Macrothink Institute™

International Research in Education
ISSN 2327-5499
2023, Vol. 11, No. 2

Construction and Analysis of Items for Assessing Learning Outcomes: Concise Models for Ensuring Validity

Samuel Ofori Bekoe

Department of Social Studies Education

University of Education, Winneba

P. O. Box 25

Winneba, Ghana

Received: June 20, 2023 Accepted: July 13, 2023 Published: July 16, 2023

doi: 10.5296/ire.v11i2.21160 URL: https://doi.org/10.5296/ire.v11i2.21160

Abstract

How can we effectively tell whether learners have acquired, and can exhibit outcomes that were initially established for them, and instructions tailored to? This question leads to assessment of learning outcomes or instructional results. It is for instance held that an outcomes-based approach to this requires assessment, in authentic ways, of what is considered to be most important of students' attainments. Unfortunately, the use of inappropriate assessment/test items/instruments is a widespread phenomenon and has become a practice/malpractice most urgently in need of improvement. To ensure such improvement is to satisfy the most important criteria in assessment/test administration; validity. The prevailing assessment culture is however still steeped in the pre-occupation with reliability. This is due to the notion that for an assessment to be reliable it must first be valid, and the subsequent assumption that the reliability of an assessment invariably ensures its validity, as there is no structured/formulaic way of determining validity. It is however known that an assessment can be reliable without necessarily being valid. This paper therefore attempts to fill this validity void, by presenting two well-structured models/flowcharts; one, for verifying the validity or usefulness/appropriateness of assessment items and the other for the construction/writing of valid/appropriate assessment items.

Keywords: Assessment item analysis model, assessment item construction model, assessment item validity model, assessment validity, assessment of learning outcomes, good assessment items

1. Introduction

According to Ebel and Frisbie (1991) "Teaching does not occur unless evaluation of learner performance occurs" (p.28). Therefore, one cannot say he/she has taught when no assessment has been done to determine whether or not his/her learners have attained the instructional/learning objectives/outcomes. This implies that teaching can be said to have taken place only when its objectives have been met, and it is only when we assess that we can determine whether these objectives have been met. The foregoing makes assessment an important and integral part of the instructional and thus the curriculum process (Pratt, 1994; Sutton, 1991; Black & Broadfoot, 1982). It is in this direction that teachers do assess their students all the time, through the use of different instruments, to find out how much they are coping with instructional and learning activities.

According to Banks (1990), the significance of assessment in the instructional and thus the curriculum process becomes more evident when the following questions are taken into consideration:

1. How well have students/pupils accomplished the goals and objectives of the course or lesson?
2. What progress have they made?
3. Have they improved upon their abilities and general performances? (p.468)

Since these questions are asked by almost every teacher and other stakeholders in the educational enterprise, assessment therefore becomes a central feature in the educational process (Sutton, 1991). Thus Pratt (1994) is of the view that "Ideally, curriculum intentions, assessment, and instruction are all part of an integrated whole" (p.104).

However, for an activity that occupies a high proportion of their professional practice, teachers, it is claimed, receive very little or no formal training in assessment during their preparatory stages (Stiggins, 2001; Stiggins & Conklin, 1992) thus, limiting their assessment literacy (Plake & Impara, 1997). This has affected the way most teachers assess in the classroom, which therefore has led to the claim that many assessments, especially external assessments, are plagued by inappropriate items (Bennett, Jenkins, Persky & Weiss, 2003) that do not do any justice to the learners, teachers and the curriculum as a whole. That is, many assessment instruments are not adequately assessing the construct or performance they are intended to assess, and also all learning outcomes to their appropriate level of demands.

The situation described above became more apparent; first, in a report of a study on the content validity of the question papers for the then Senior Secondary School Certificate Examination (SSSCE), published by the West African Examinations Council (WAEC) in 2002 and subsequently, in an analysis of WAEC's assessment items in Social Studies at the Senior High School (SHS) level in Ghana (see Bekoe, 2007 & 2006). This paper therefore examines the literature on how to ensure the appropriateness, and thus the validity, of assessment/test items. It then proposes an Assessment Item Analysis/Evaluation Model and

an Assessment Item Writing/Construction Model, in the form of flowcharts that can be employed to guide the validation analysis/selection and construction of good/appropriate assessment respectively. They are to ensure that educational assessors are able to select or write assessment items that are valid and thus useful to be used in the assessments of students' learning outcomes.

2. Understanding Educational Assessment

The conception and use of the term assessment, in education, is so varied that it connotes different things at different occasions and sometimes used inter-changeably with other terms (Ghaicha, 2016; Bachman, 2004; Mundrake, 2000). Cizek (1997) for instance argues that the term assessment "is used in so many different ways, in so many different contexts, and for so many different purposes, that it can mean almost anything" (p. 8). Assessment is for instance, sometimes used to connote evaluation (Ghaicha, 2016; Nelson & Michaelis, 1980) or measurement (Ghaicha, 2016; Kelly, 2009; Ecclestone, 1994; Gross, McPhie & Fraenkel, 1970) or testing (Ghaicha, 2016). However, many of these authors, among others, have argued against the notion of equating assessment, evaluation and measurement or testing as one and the same concept. Some have even actually tried to make clear distinctions among these obviously separate, but related, terms.

Nelson and Michaelis (1980), for instance, argue that evaluation is broader than assessment, and for that matter assessment is rather seen as part of the evaluation process (Lambert & Lines, 2000; Satterly, 1989). Measurement, on the other hand, is also seen as part of assessment (Coulby, 2000). On the contrary Ecclestone (1994) sees assessment as rather an act of measurement, by stating that "Assessment is the judgement of evidence submitted for a specific purpose; it is therefore an act of measurement. It requires two things: evidence and a standard or scale" (p. 6).

Mager (1997) however defined measurement as "the process of determining the extent of some characteristic associated with an object or person. For example, when we determine the length of a room or weight of an object, we are measuring" (p. 8). That is using a standard/universal instrument, like a ruler/measuring tape, weighing scale or a compass to determine the extent to which some characteristics of an object or person can be associated with a value/measure on such a standard instrument (criteria). In a more precise way, Ghaicha (2016), in reference to students' performance, defined measurement as, "the process by which a quantified value, usually numerical, is assigned to the attributes or dimensions related to students' performance while measuring ability or aptitude in such a way that the student's quality of performance is preserved" (p. 213). Although Ecclestone's (1994) assertion that measurement requires evidence and a standard or scale, is accepted by most writers, her view that assessment is an act of measurement is challenged by many other writers, and logically so.

Kelly (2009) for instance argues,

> The term measurement brings with it, connotations of accuracy and precision, but it is plain to anyone who will look more closely at the matter

> that there is little accuracy or precision in most forms of educational assessment. And the degree of accuracy and precision varies inversely in relation to the complexity and sophistication of what is being assessed (p.129).

Ecclestone, perhaps, made the assertion that assessment is part of measurement or even assessment is measurement, because she sees it in the like of tests and examinations. However, Rowntree (1987) disagrees with such assertions and argues, "Despite one of the assumptions commonly made in the literature, assessment is not obtained only, or even necessarily mainly, through test and examination" (p. 4). Satterly (1989) also states, "Educational assessment takes place in many ways using a variety of instrument designed for the purpose" (p. 10). Thus,

> All shades of assessment can be practiced without any kind of measurement that implies absolute standards; it may be enough simply to observe whether, for each student, some personal, even idiosyncratic, trait or ability appears discernible to a greater or lesser extent than hitherto (Rowntree, 1987: 5).

Since there is no universal standard or scale to measure the extent to which such personal characteristics as; ability, skill, attitude and value, which are all the subject matter of assessment, exist in a person, it will be inappropriate, as it is not supported by facts and logic, to accept the view of Ecclestone (1994) that assessment is an act of measurement or even is measurement. Rather, in assessment, measurement may sometimes be applied when certain characteristics, like knowledge or cognition, are seen to be amenable to a measure and thus associated with a figure or value on a standard or criterion or norm. Thus, assessment is seen as involving more than measurement (Ghaicha, 2016; Nelson & Michaelis, 1980; Gross, McPhie & Fraenkel, 1970). And thus, according to Eisner (1993, also cited by Kelly, 2009) "Assessment, like evaluation, is not one but several things" (p. 224).

Evaluation, on the other hand, involves the comparison of a measure to a standard and afterwards making judgement on the comparison (Mager, 1997). Ghaicha (2016) for instance defines evaluation as, "the process of arriving at judgments about abstract entities such as programs, curricula, organizations, institutions and individuals" (p. 213). It is therefore often considered as an appraisal of the whole curriculum or instructional process, and for which assessment is part or a tool (Kelly, 2009). In fact, assessment and evaluation, apart from the attempt by some authors, like Nelson and Michaelis (1980), to make a distinction between them and place assessment in the domain of the instructional process and evaluation at the end of the whole programme, sometimes become confusing in meaning. They look almost the same, when especially assessment is seen as being judgemental (Kelly, 2009; Cizek, 1997; Ecclestone, 1994; Wiggins, 1993) as in the case of evaluation. For instance, Wiggins (1993) defines assessment as "a comprehensive, multifaceted analysis of performance; it must be judgment-based and personal" (p. 13).

There are, however, others who are of the opinion that assessment is not judgemental (Lambert & Lines, 2000; Wiersma & Jurs, 1990; Rowntree, 1987). In this school of thought,

Wiersma and Jurs (1990) were more straightforward and perhaps daring with their opinion when they stated, categorically that "when assessment is taking place, information or data are being collected and measurement is being conducted. Assessment does not include making judgments about data, which is reserved for evaluation" (p. 8). In this case a clear distinction is being made between assessment and evaluation. Whereas assessment is indicated to connote the collection of all kinds of data about students/pupils, evaluation is seen as the act of making judgements on the data collected. Thus, assessment is seen as an important tool of evaluation.

Rowntree (1987) and Lambert and Lines (2000) were however cautious in making such a categorical assertion, as their views are implicit rather than explicit. Rowntree (1987) for instance states that assessment "can be descriptive without becoming judgemental" (p. 6). This can also imply that assessment can sometimes be judgemental. Lambert and Lines (2000) pointed out the subservience of assessment to evaluation when they wrote, as part of their explanation of the evaluative role of assessment, that the purpose is "to contribute to the information on which judgements are made concerning the effectiveness or quality of individuals and institutions in the system as a whole" (p. 4). This also places assessment, squarely, in the domain of data gathering or collection.

It is however apparent in the discussions so far that both schools, in the assessment/evaluation debate, do agree that assessment involves the collection of data about individuals or a system. That is, whether assessment is judgemental or not, there is no question about the fact that it involves obtaining some form of information about some personal or institutional characteristics or attributes. Rowntree (1987) therefore states,

> Assessment in education can be thought of as occurring whenever one person, in some kind of interaction, direct or indirect, with another person, is conscious of obtaining and interpreting information about the knowledge and understanding, or abilities and attitudes of that other person (p. 4).

It must be noted that the kinds of information being conceived in this case are not exclusively linked to those that are obtained through tests and examinations alone (measurement), but to others, including very informal or indirect ones (Lambert & Lines, 2000; Rowntree, 1987). Cizek (1997); Ferrara and McTighe (1992), Baker and Stites (1991), and Stiggins (1991) have all argued that notions about assessment need to be broadened. And according to Airasian (1994), should include "the full range of information teachers gather in their classrooms: information that helps them understand their pupils, monitor their instruction, and establish a viable classroom culture" (p. 5).

It has therefore been established that educational assessment, whether judgemental or not, is a process of obtaining all kinds of data about the characteristics of learners, in relation to set standards of attainment in the curriculum. As Satterly (1989) aptly puts it,

> Educational assessment is an omnibus term which includes all the processes and products which describe the nature and extent of children's learning, its degree of correspondence with the aims and objectives of teaching and its

> relationship with the environments which are designed to facilitate learning (p. 3).

The above definition provides a comprehensive scope or perspective of what assessment is or should be. It reflects the components of the educational process; the curriculum, instruction and assessment, and the balance/relationship, which must be maintained among these components at all cause (Pratt, 1994; Madaus, 1988).

Satterly (1989), by his definition, is calling for assessment, by whatever means, to be able to precisely describe what learners have attained in relation with curricular imperatives and how the learning was designed to take place. Thus, it will not be appropriate to assess learning attainment of a student, who has been taught how to use the computer, by asking such a student to draw and label the parts of a computer. Neither will is be appropriate to assess students on concepts that they have not been taught. Unfortunately, many assessment regimes and practices are said to be plagued by such items/questions (Mager, 1997) and especially by banal and elemental ones (Bennett, et al, 2003). The question therefore is, how do we ensure that educators, and all those involved in assessing learner outcomes, one way or the other, will always construct/select and use assessment procedures and or items that are good and useful in determining what learners know and can do, as per what they were taught and how they were taught. This question becomes particularly pertinent, when one considers the view that assessment, in this era of accountability, is a powerful instrument that can either boost or undermine students' learning (Ghaicha, 2016).

3. Criteria for Good Assessment Items: The assessment reliability versus validity debates

There are many important criteria that are traditionally employed to ensure the quality and credibility of assessment instruments/items (e.g., validity, reliability, practicability, standardization, etc.). Validity and reliability are however the two fundamental features/criteria that are mostly/commonly used (Mohajan, 2017; Darr, 2005a & 2005b), in ensuring that an assessment instrument/item gives a true reflection of the state of capability of learners (what they know and can do), and do so consistently with high level of predictability. It is further held that of the two, validity is the most important (Newton & Baird, 2016; Darr, 2005a), especially for the evaluation of assessment instrument/items (Mager, 1997; Tyler, 1949). Newton and Baird (2016), for instance, claim that "validity is the most important term in the educational and psychological measurement lexicon" (p. 173). Darr (2005a) also states that "validity can be considered as the key issue in assessment" (p. 55). This is mainly because, according to Akib and Ghafar (2015), "validity requires that an instrument is reliable, but an instrument can be reliable without being valid" (p. 67). And according to Hughes (2003), "if a test is not reliable, it cannot be valid" (p. 34). However, reliability has over the years been too often overemphasized at the expense of validity (Linn, Baker & Dunbar, 1991).

Reliability, it must be noted, is about the consistency of test scores and not the test itself (Linn, Baker & Dunbar, 1991). The Center on Standards & Assessment Implementation (CSAI) (2018) also states, "Reliability is a measure of consistency. It is the degree to which

student results are the same when they take the same test on different occasions, when different scorers score the same item or task, and when different but equivalent tests are taken at the same time or at different times" (p. 1). Darr (2005b) expatiates on the issue of consistency, by indication,

> "This may mean:
>
> - Consistency across time—would the results have been the same if the test or assessment had taken place on another day, or at another time?
> - Consistency across tasks—would the result have been the same if other tasks had been chosen to assess the learning?
> - Consistency across markers—would the results have been similar if another marker had scored the assessment?" (p. 59)

And according to Bachman (1990) "if it was possible for a test candidate to take the same test in an unaffected environment several times, it is conceived that the eventual mean score would provide a total that would closely equate to the participants true score" (p. 167).

In spite of the position that validity is the most important criteria, it has rather been variously argued that much emphasis, in assessment, is often placed on reliability at the expense of the validity (Linn, Baker & Dunbar, 1991). This is perhaps because the traditional notions of reliability and validity holds that for something to be valid it must first be reliable (Akib and Ghafar 2015; Hughes, 2003). It is therefore often assumed that once an assessment is computed/determined to be of high reliability, it automatically becomes valid. This assumption is apparently problematic and flawed, in the sense that there is evidence, both from research and experience, that this is not always the case, as according to Akib and Ghafar (2015), "an instrument can be reliable without being valid (p. 67).

For instance, if a teacher teaches his/her students the four basic operations in Mathematics, and ends up assessing them on only additions, the reliability of the assessment could be quite high as students' scores would possibly show consistency. However, such a test will definitely not be valid in respect of the content lacking breadth of coverage and thus being unrepresentative of the intended curriculum objective/outcome. It would also not provide a complete picture of what the students have actually acquired and can do. The scenario above clearly indicates that the validity of an assessment instrument cannot be guaranteed just because it has been determined to have a high degree of reliability. Haydn, Arthur & Hunt (2001) also argued that "A test loses validity if the pupils are being assessed on content, skills or concepts which they have not been taught" (pp. 237-238). That is, while such a test may score high on the reliability scale, the fact that it is assessing outcomes that pupils have not been instructed in makes it invalid. Thus, although reliability is a necessary criterion, it is however not sufficient for ensuring validity (Thompson, 2013).

It is also argued that the issue of test reliability itself is inconclusive and various questions have been raised against it. The first of such questions is about the computation of the reliability coefficient of a test. According to Darr (2005b) "determining reliability has traditionally been seen as a statistical exercise. It usually involves calculating a reliability

coefficient to indicate how well assessment results agree over repeated uses of the assessment tool" (p. 59). Examples of methods usually employed in estimating reliability include test-retest reliability, internal consistency reliability, parallel-test reliability and analysis of Classical True Score (CTS) (Nadasdy, 2011; Bachman, 1990).

In the case of test-retest reliability, for instance, the argument is that it is practically absurd to give the same test to pupils on two separate occasions (McMillan, 2002). The second issue has to do with the appropriate duration in-between the two tests, as it is argued that whether the second test is taken immediately or sometime after the first one, many things could happen between the time spans to impact on the subsequent performance (McMillan, 2002). Bachman (1990) also argues that "other factors concerning what we are measuring will affect test reliability. Factors including test participants' personal characteristics i.e., age, gender, and factors regarding the test environment and condition of the participants can contribute to whether or not a test is effectively reliable" (p.164). It is also wondered, against the split-half reliability test, whether two different items can really measure the same thing or construct (McMillan, 2002). It must also be noted that reported error in reliability of traditional test scores is often underestimated (McMillan, 2002; Rogosa, 1999). Nadasdy (2011) also argues that the results of CTS are still in the theoretical realms and therefore may not take into account variables that could be established through empirical investigations.

The question, which perhaps arises from the foregoing discourse is that should we not rather be emphasising and ensuring the validity of an assessment instrument/item, instead of solely focusing on the reliability of test scores which cannot be absolutely relied upon. This view is supported by Tyler (1949) who argued to the effect that the most important criterion for an evaluation/assessment instrument is validity, which needs to rather consciously ensured all the time (Mager, 1997). Thus, inasmuch as measurement experts and test constructors go to great lengths to ensure the reliability of assessment instruments/items, they must also verify, and even more rigorously, the validity of such instruments/items. Certainly, a test cannot qualify to be a good test if it has high reliability, but then lacks validity. It is further argued that assessment decisions are reliable when they are based on evidence that is generated by valid assessments (SQA, 2001). It therefore holds that in constructing or evaluating assessment instruments or test items, reliability should be considered as just one of the criteria and not the only criterion, and validity as rather the most important of the criteria.

Validity itself has been embroiled in contentious debates, therefore becoming a thorny issue among assessment and measurement professionals, as to what it is supposed to mean (Newton & Baird, 2016). According to Darr (2005a)

> In the past, validity has often been treated as the degree to which a test or assessment tool measures what it claims to measure, as if this was something inherent in the assessment instrument itself. More recently, however, assessment specialists have argued that validity should not be considered as a fixed property of an assessment instrument. Instead, they propose that validity is better understood as an evaluation of the quality of the interpretations and decisions that are made on the basis of an assessment result—that is, how well

> the inferences we make or actions we take on the basis of an assessment result can be justified (p. 55).

Brualdi, (2002), for instance, claims that "Test or assessment validity refers to the degree with which the inferences based on test scores are meaningful, useful, and appropriate" (p. 12). Moskal and Leydens (2002) also refer to the American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) definition of validity as, "The degree to which the evidence supports that the interpretations are correct and that the manner in which the interpretations are used is appropriate" (p. 77). AERA, et al (2014) have provided a recent definition of validity, which states that "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). This thus suggests that validity is about whether interpretations or inferences made of a student's test scores are true reflections of his/her ability to really perform the task that the test indicates.

However, Sireci (2015) argues that "to support the use of a test for a particular purpose will require evidence that the test is measuring its intended construct, the scores are interpreted as appropriate manifestations of that construct, and the actions based on those interpretations are defensible" (p. 8). He (Sireci, 2015) also argues that "providing evidence that a test is measuring what it purports to measure is a necessary component of a validation effort and so it is certainly germane to validity. Such evidence helps us evaluate, and validate, the interpretations made on the basis of test scores" (p. 5). Thus, according to Linn, Baker and Dunbar (1991) "questions of validity focus their attention on long-range objectives, criterion situation… and the extent to which they are reflected in the tasks presented to learners on a test" (p.1). CSAI (2018), also goes back to the earlier notion of validity to state that "validity can be summarized as how well a test measures what it is supposed to measure" (p. 2).

It has however been respect argued that the traditional notion of validity has viewed the concept too narrowly (Brualdi, 2002; Messick, 1996 & 1989; Linn, Baker & Dunbar, 1991). For instance, the traditional means of accumulating validity evidence have been grouped under several categories (Brualdi, 2002; Bachman, 1990; Hughes, 2003). Thus, evidence for validity has been sought in terms of the following:

1. Construct Validity: The correlation between tests measuring the same construct or between a test and the criterion behaviour of interest (Taylor & Nolen, 1996; Nitko, 1996; Linn & Gronlund, 1995; Hanna, 1993). According to Darr (2005a) "constructs are specific psychological characteristics or traits, such as a type of reasoning or thinking, that we are interested in assessing" (p. 56).
2. Content Validity: Darr (2005a) states that "the content of our assessments as part of a validity argument involves evaluating how well our assessment tasks represent or sample the learning domain in question" (p. 55). In most cases this involves the use of tables of specifications to determine whether the content of a test measures the breadth of content targeted (Taylor & Nolen, 1996; Oosterhof, 1996; Linn & Gronlund, 1995).

3. Criterion-related Validity: Using a range of strategies to build a logical case for the relationship between scores from the assessment and the construct the assessment is intended to measure (Taylor & Nolen, 1996). According to Cohen, Manion and Morrison (2000), this type of validity is said to contain two primary forms; predictive and concurrent validity. An assessment is said to have strong predictive validity, if a separate but related assessment produces similar results, as it did. Concurrent validity is similar but it is not necessary to have been measured over a span of time, but can be determined simultaneously with another instrument (Cohen, et. al, 2000). Darr (2005a) also claims that "sometimes, developing a validity argument involves looking at how well our assessment results compare with or predict other measures recorded on a separate assessment or criterion" (p. 56).

4. Face Validity: This term relates to what degree a test is perceived or appears to be measuring what it is supposed to measure (Bhandari, 2022; Middleton, 2022; Nadasdy, 2011)

5. Consequential Validity: This type of validity involves evaluating the consequences of using assessment results, since the weight given to the results of an assessment and the use thereof will have an impact on teaching and learning (Darr, 2005a). According to Darr (2005a) "we should question the validity of our assessment when there is evidence that the consequences of using the assessment results to make decisions or inform students of progress are detrimental to our overall educational goals" (p. 56).

Brualdi (2002), however, claims that "there are no rigorous distinctions between them; they are not distinct types of validity" (p. 12). The modern concept of validity, as advanced by Cronbach (1988) and Messick (1989), instead views construct validity as the unifying concept underlying all validity (see Gipps & Murphy, 1994; Brualdi, 2002). The argument is that the traditional notion of validity is fragmented and incomplete, as it fails to take into account evidence of the value implications of inferences made from scores as a basis for action and also the social consequences of the way inferences are made from the scores (Brualdi, 2002; Messick, 1989; Cronbach, 1988). The proponents of the modern conception of validity therefore suggest that validity should be seen as a unitary concept and its categorisations rather as its components. They also called for an expanded view of validity to include other important concerns (Linn, Baker & Dunbar, 1991).

Linn, Baker & Dunbar (1991) are of the view that the idea of an expanded notion of validity becomes more imperative and central to the evaluation of the adequacy of new forms of educational assessment. They claim that such criteria "provide a framework that is consistent with both current theoretical understandings of validity and the nature and potential uses of new forms of assessment" (p. 4). It must be clarified that the components of the expanded conceptualisation of validity is currently inexhaustive (Linn, Baker & Dunbar, 1991), with different authors producing different lists of what should constitute the components of validity. However, a close examination of some of the lists shows that they have similarities

among them, and that two or more components of validity, by one author could as well fit into a component of the other.

Linn, Baker & Dunbar (1991) and Brualdi's (2002) lists of components of validity, for instance, could all be merged to produce the following:

1. CONTENT: This will be the merging of Brualdi's (2002) components of 'Content' and 'Substansive' with Linn, Baker and Dunbar's (1991) 'Content Quality' and 'Content Coverage'. Content, in this case refers to the extent to which the content of assessment is consistent with best current understanding of the field and at the same time reflects what are judged to be aspects of quality. It also refers to the comprehensiveness of content coverage and the extent to which the assessment is relevant and representative of the construct domain (Brualdi, 2002).

2. STRUCTURE: According to Brualdi (2002), this "is about how the internal structure of the assessment is consistent with what is known about the internal structure of the construct domain" (p. 13)

3. TRANSFER AND GENERALISABILITY: The evidence that the performance in a specific task can be transferred to other tasks to allow for consistency and thus generalisation. (Brualdi, 2002; Linn, Baker & Dunbar, 1991).

4. CONSEQUENCES: This involves the collection of evidence about both the intended and unintended effects of assessment on the way teachers and students spend their time and think about the goals of education. In other words, this is about whether the interpretation of assessment results leads to either positive or negative consequences (Herman, 1992). Gipps and Murphy (1994), in reference to a TGAT Report (Department of Education and Science, 1988), reiterated that external assessment should, among other things, "not have undesired effects on the curriculum" (p. 187).

5. FAIRNESS: This is whether assessments and the interpretations of their results take into consideration the cultural and socio-economic background of students and whether there is evidence of offensive items to some students and/or sources of irrelevant difficulty for students (Linn, Baker & Dunbar, 1991)

6. COGNITIVE COMPLEXITY: This is about the evidence that no matter the difficulty of the subject matter, assessment/test items really require students to exercise higher order thinking and reasoning processes (Linn, Baker & Dunbar, 1991).

7. MEANINGFULNESS: This answers the question as to whether an assessment task is meaningful to students and whether it does provide worthwhile educational experiences (Brualdi, 2002; Linn, Baker & Dunbar, 1991).

8. EXTERNAL FACTORS: This is about the extent to which the relationship of assessment scores with other measures and non-assessment behaviours reflect the expected relations implicit in the intended construct. That is, is the score

interpretation externally substantiated, by appraising the degree to which empirical relationships are consistent with the meaning of the construct or subject matter of the assessment (Brualdi, 2002).

9. COST AND EFFICIENCY: This aspect of validity is about the cost effectiveness of the assessment instrument, especially for large-scale assessment (Linn, Baker & Dunbar, 1991).

Whether we are to view validity in its traditional or modern expanded form, it is argued that it should be seen as minimizing invalidity and maximizing validity (Cohen, et. Al, 2000). And to Nadasdy (2011) validity should be a matter of degree rather than a pursuit of perfection. The argument, therefore, is that in constructing or selecting assessment instruments/items, every effort must be made to ensure that they are valid to a greater degree and thus potentially useful in assessing what they are supposed to assess (Mager, 1997).

4. Guidelines for Ensuring the Validity of Assessment Instruments/Items

Having established the meaning of assessment/test validity and having identified the components therein, as per the modern expanded unitary notion of validity, the next issue for consideration is how one can ensure that an assessment instrument or item encompasses all these components and thus is potentially useful for assessing learning outcomes. One practical guideline, as provided by the SQA (2001), is that an assessment is valid when it is appropriate to or fit for purpose (e.g., using practical assessment to assess practical skills), and allows the production of the evidence of students' performance which can be measured against defined standards. This implies that in the construction of assessment items every effort must be made to ensure that their quality is mostly assured by their validity. In this direction Herman (1992) and Linn, Baker & Dunbar (1991) identified the characteristics of a good assessment to guide such an endeavour. It must be noted the characteristics, so identified, are similar to most of the criteria already identified as components of validity. This perhaps goes to support the assertion made by Mager (1997) and Tyler (1949) that the only good assessment is the assessment that is valid. In order words, test validity should capture the characteristics of good assessment (Dietel, Herman & Knuth, 1991).

Many models/checklists have been proffered to aid assessors to ensure that their assessment instruments/items are valid or good, and thus potentially useful in assessing students' learning outcomes. These include, but not limited to, Quelmalz's model (Quelmalz & Hoskyn, 1997), Darr's validity checklist (Darr, 2005a), McMillan's characteristics of good assessment (McMillan, 2002) and Mager's checklist and flowchart (Mager, 1997). The Quelmalz model (Quelmalz & Hoskyn, 1997), for instance, provides the following criteria:

1. Problems or tasks should represent important recurring issues or activities.
2. Emphasise purposeful, sustained, reasoning that requires integration of reasoning strategies rather than demonstration of discrete isolated skills.

3. Assessment tasks should permit multiple interpretations or solutions, rather than one right answer. That is the encouragement of alternative points of views and conclusions.
4. Assessment formats should elicit explanations of inquiry processes, not just the answer.
5. Assessment tasks and problems should represent a range of generalisation and transfer.
6. Assess reasoning strategies directly, not as undifferentiated components of a more complex solution.
7. Assess meta-cognitive strategies for planning revision and self-evaluation (p.105).

Darr (2005a) also provided the following, as a validity checklist:

- Do the tasks match the learning intentions we are interested in?
- Does the test cover a wide enough range of content?
- Are there enough items or tasks to cover the scope of what is being assessed?
- Do the tasks require use of the desired skills and reasoning processes?
- Is there an emphasis on deep, rather than surface knowledge?
- Are the directions for the assessment task clear?
- Are the questions unambiguous?
- Are the time limits sufficient?
- Do the tasks avoid favouring groups of students more likely to have useful background knowledge- for instance, boys or girls?
- Is the language used suitable?
- Are the reading demands fair? (p. 55)

McMillan (2002), on the other hand, identified the following as characteristics of good assessment:

1. Good assessment must enhance instruction, by its integration with instruction in the classroom.
2. Good assessment should be valid, in its modern and expanded form.
3. Good assessment should be fair and ethical, in that it must ensure students' knowledge of learning targets and the nature of the assessments prior to instruction and avoid stereotypes.
4. Good assessment must use multiple methods to ensure that a complete picture of what students understand and can do is put together in pieces comprised by different approaches to assessment.
5. Good assessment is efficient and feasible in the sense where benefits outweigh cost.

Apparently, the components of validity and characteristics of good assessment, even though

similar and are to be taken as one and the same thing, seem unwieldy if one is to consider the extent of the issues or concepts they embrace. These criteria therefore seem difficult for one, without the appropriate guidance, to meet when constructing an assessment item or tool. Mager (1997) therefore provided a simplified solution, by insisting that one simply needs to match the performance and condition of the item to that of the curriculum objectives. His mantra is, "write or select items that will ask students to do what the objectives say they are able to do" (p. 15). This is supported by Farris (2015) when she, in reference to Cangelosi (1990) and Popham (1995) stated, "if the assessment items do not match the content and the behavioral construct of the objective, then the assessment is of the little value" (p. 68). SQA (2001) also provides that in devising assessments, one should ensure that all outcomes are covered to the appropriate level of demand, as described by the performance criteria or objective.

In furtherance to the above position, Mager (1997) presented two models (The Objective/Item Checklist and The Objective/Item Flowchart) that could be used for the construction or selection of assessment items, and also serve as useful tools for the evaluation of test items. These models can be seen as classic guidelines, which can help teachers and test constructors to come out with appropriate tools for assessing learning outcomes. They can also be used to ensure that the use of inappropriate test items, which is a widespread phenomenon, is done away with and the whole assessment culture improved.

The checklist is a six (6) steps instruction that presents a systematic approach to item evaluation. It starts with the identification of the performance as stated in the curriculum objective, and verification as to whether it is an overt/covert main intent or indicator of the main intent, followed by the establishment of its clarity. The next step is to identify the performance being demanded of the students, by the item, after which a match/congruence between the two performances is determined. The last step on the checklist calls for the establishment of a match between the objective and item conditions (i.e., does the condition in which the assessment is to be done match the condition, in which the teaching and learning took place).

The flowchart on the other hand is an eighteen (18) steps model, which is an expanded form of the checklist, described above. However, anyone intending to use Mager's models for the evaluation of assessment/test items could as well expand it to include other important criteria like; fairness, cognitive complexity, meaningfulness (Herman, 1992), and the contextualisation of task in real-world applications (Dietel, Herman & Knuth, 1991). This will ensure that the criteria for evaluating test items will be more comprehensive and rigorous, and should result in very good test items that can really assess, in an objective and valid manner, whatever curriculum objective they intend to assess.

5. Models for Analysing/Selecting and Constructing/Writing Valid Assessment Items

Bekoe (2006), in an attempt to employ a more comprehensive/concise validity model for the evaluation of some selected Social Studies examination questions, by the West African Examinations Council (WAEC), developed a new model, which was an adaptation of Mager's (1997) Flowchart. This new model went beyond Mager's (1997) Flowchart to

include steps in ensuring that an assessment item meets some other criteria, not included in Mager's model. This model has been modified over the years, to include more details and for ease of use. Thus, the current model (Figure 1), is a build-up of Bekoe's (2006) model and Mager's (1997) flowchart. A second model (Figure 2), which is closely related to the first model, is also being proposed to aid the construction/writing of more appropriate, and thus valid assessment items.

The Assessment Item Analysis/Selection Model (Figure 1) contains twelve (12) major steps in the whole process. It starts with the selection of an item, followed by determining whether the curriculum content area or topic/unit, in relation to the item, can be identified in the curriculum. The next step is the identification of the performance in the item whether it is the main intent or an indicator of the main intent. This is to be followed up by noting the precise/explicit performance in the item, and identifying same in the curriculum objective/learning outcome that relates to the item being analysed.

Another major step is to determine whether the performance, identified in both the item and the curriculum objective/learning outcome, do match. This step is followed by matching the conditions for both the attainment of the objective/learning outcome and performing the task in the item. Subsequently, if the item is so determined to be contextualised in real-world application, to a great extent, then it can be accepted as a potentially useful and valid item to be used to assess what students know and can do. There are five (5) main decision stages in the process, where an assessor may reject an item completely, as invalid/inappropriate, or continue through with the process to determine whether the item should be accepted, as valid and thus potentially useful.

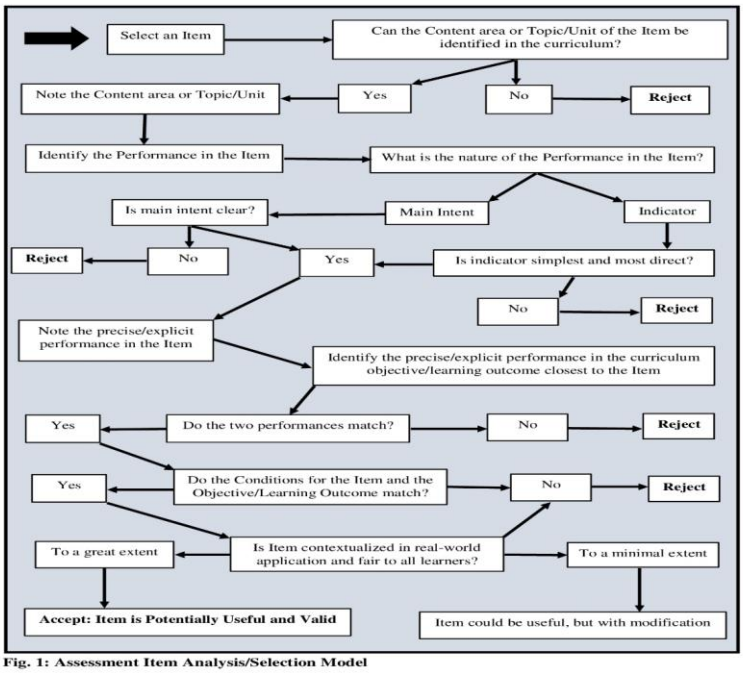

Fig. 1: Assessment Item Analysis/Selection Model

The Assessment Item Writing/Construction Model (Figure 2), on the other hand, contains eleven (11) major steps. This begins with selecting the curriculum content area or the topic/unit one wants to assess students' learning outcomes on. The next step is to select a specific curriculum/learning objective/outcome pertaining to the topic/unit. This is to be followed by identifying the performance in that objective/outcome. The item writer will then have to determine the nature of the performance, whether it is an indicator or the main intent.

Furthermore, after the precise/explicit performance in the objective/outcome has been established, the writer then moves to write an item with the established explicit performance. This is to be followed up by ensuring that the performance in the objective/outcome really matches the one in the written item. The next step is to ensure that the condition for performing the task in the item matches the one for the attainment of the objective/outcome. The item can be deemed to be potentially useful and valid to be used to assess students' learning attainment, if it is finally determined to be well contextualised in real-world application.

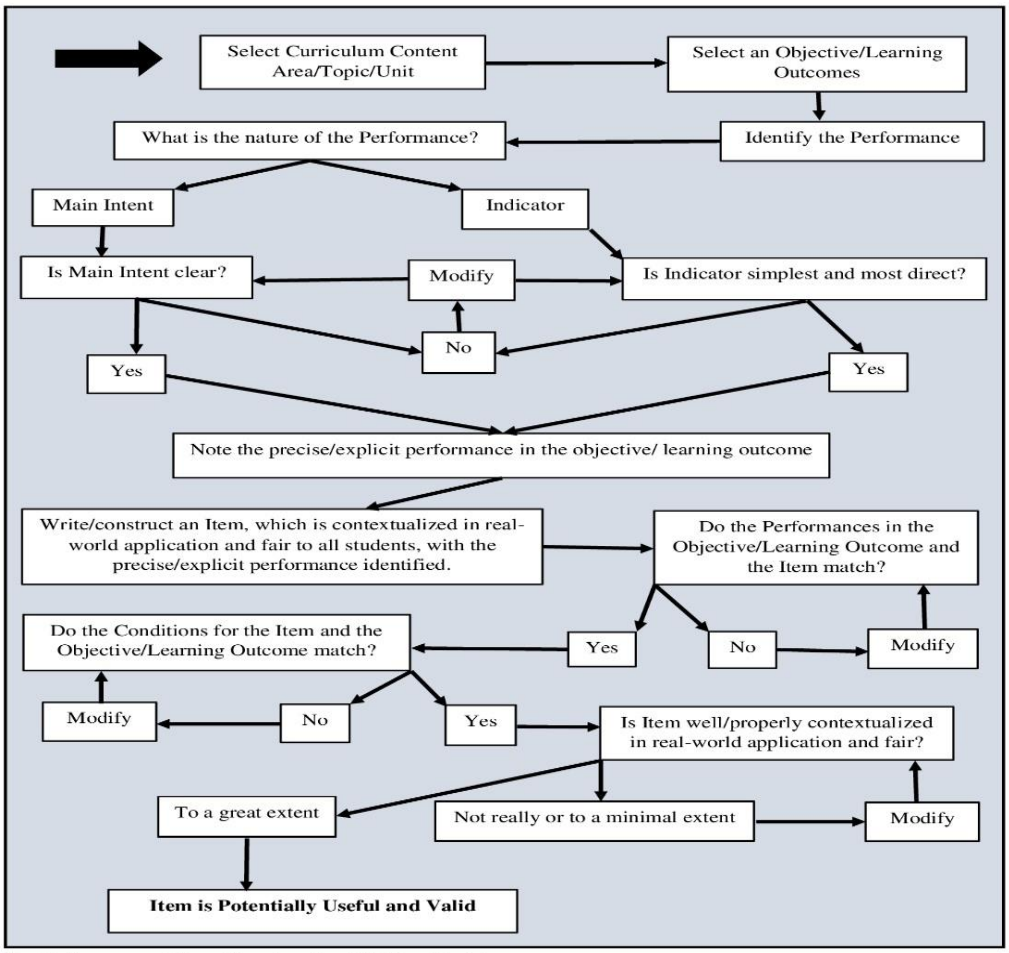

Fig. 2: Assessment Item Construction/Writing Model

Both models (Figure 1 & Figure 2) throw up some very important and yet challenging activities, for assessors that need further and detailed explanations. These include the identification of the performance, matching the conditions, the contextualisation of the item in real-world application, and making the item fair. The challenge may be how to identify the performances and match them; identifying the conditions and matching, contextualising an item in real-world application, so as to make it have complexity, be educationally worthwhile and thus meaningful, and making it fair to all students taking it.

5.1 Identifying the Performance

To be able to successfully identify the performance, one has to decode (Mager, 1997) the objective or the item. The procedure for decoding the objective or item is as follows:

1. Identify the performance stated in the objective or item. Just note the word or phrase that specifies what students should be doing when demonstrating their achievement of the objective or performing the task in the item. For example, in "students will be able to identify some of the socio-cultural practices in Ghana", the underlined word, identify, is the performance.

2. Note whether the performance is the main intent or an indicator. If the performance describes the main thing to be done, then it is the main intent. However if the performance is an act that will tell us whether a student can really do what the objective is about, then it is an indicator. For example, "be able to kick a football" is a main intent, whereas in "be able to underline all natural resources in a list of resources", the underline is an indicator. The main intent is differentiating among the various types of resources.

3. The precise/explicit performance therefore is the ability to differentiate among the various types of resources, by selecting only natural resources from a list that includes all the types.

5.2 Matching the Conditions

The specific environment in which a particular learning is designed to take place (Satterly, 1989) or the performance of an assessment is supposed to take place is what is referred to as condition (Mager, 1997). For instance, if students learn how to bake cake, by going through the actual process of baking in a bakery of any place with all the necessary equipment made available, then that will be the condition for the learning. The assessment item must equally ask them to bake a cake in a similar setting and not rather to describe, by writing, how to bake a cake. Such an item will be inappropriate, and thus invalid. There are however instances where the learning objective does not indicate any particular condition for its attainment by students. In that respect, there wouldn't be the need to match any conditions.

5.3 Contextualising the Item in Real-world Application

Mager (1997) argues that assessment items should require students to performance tasks that are as close to the real things as possible. That is, if they were made to learn a skill, then the assessment should ask them to perform that skill in its actual/real setting, to demonstrate their

attainment of the skill. Mager (1997) further argued that it is only when it becomes impossible or too risky for students to perform the task in the real setting that one can resort to approximation in finding out as to whether they have attained the objective. Thus, contextualising assessments in real-world application (Dietel, Herman & Knuth, 1991), either fully or approximately, will be the best way of assessing students. The contextualisation is more likely to make the assessment meaningful to the students and therefore educationally worthwhile, as they will come to know the real-life situations where they can apply what they have learnt and thus appreciate what the learning.

5.4 Making the Item Fair

Ensure that the assessment and the interpretations of the result, thereof, takes into consideration the cultural and socio-economic background of students. Also, make sure that there is no evidence of offensive words or stereotypes in the item, to some students. The item should not also contain evidence and/or sources of irrelevant difficulty for students (Linn, Baker & Dunbar, 1991).

6. Conclusion

This paper does not intend or attempt to completely settle the debates about the pre-eminence or otherwise of validity and reliability or the great validity debates itself. The propositions herein are that since validity is held to be a very important criterion in establishing the quality of an assessment item/instrument, similar efforts must be put into ensuring it, just like reliability, instead of the use of reliability as a pseudo justification for validity. Secondly, no matter how one would want to define validity, the quality of the items and their ability to assess the exact construct/outcome they intend to assess cannot ne grossed over. We cannot be deemed to be making valid interpretations or uses of the results of an assessment, when the items in that assessment are useless, as they might have no relation and thus relevance to what students were taught in the first place.

The foregoing was, therefore, what motivated the development of the models (Fig. 1 & Fig. 2) to aid in ensuring that assessment items, used in assessing what students know and can do, are appropriate, useful and thus valid. There is also no attempt to present these models as the ultimate validity models, but rather as a contribution to the discourse out there, and the efforts being made to ensure that there is a more concise, objective and robust way of validating an assessment item/instrument, instead of the subjective processes proffered in the literature. The debates may therefore go on, to bring the best out of validity.

Author Declaration

This paper resulted from authors' own efforts and work. There was no sponsorship or award, as well as any conflict-of-interest issue involved in the production of this paper.

References

Airasian, P. W. (1994) *Classroom assessment* (2nd ed.). New York: McGraw-Hill

Akib, E. & Ghafar, M. N. A. (2015). The validity and reliability of assessment for learning (Afl). *Education Journal 4*(2), 64-68.

American Educational Research Association, American Psychological Association, & National Council, on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council, on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, E. L. & Stites, R. (1991). Trends in testing in the USA. In S. H. Fuhrman & B. Malen (Eds.), *The Politics of curriculum and testing* (pp. 131-157). London: Falmer.

Banks, J. A. (1990). *Teaching Strategies for the Social Studies: Inquiry, valuing and decision-making*. (4th ed.). New York: Longman

Bekoe, S. O. (2007). Assessment of Social Studies learning outcomes: An evaluation of the appropriateness and validity of the Senior Secondary School Certificate Examination's items in Ghana. *The Social Educator, 3*(1), 119-135.

Bekoe, S. O. (2006). Assessment and curriculum goals and objectives: Evaluation of the systemic impact of the SSSCE on the senior secondary school Social Studies curriculum in Ghana. An unpublished PhD Thesis presented to the University of Strathclyde, Scotland, UK.

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem-solving performances. *Assessment in Education, 10*(3), 347-359.

Bhandari, P. (2022). *What is face validity? Guide, definition & examples*. Scribbr. Retrieved June 5, 2023 from https://www.scribbr.com/methodology/face-validity/

Black, H. & Broadfoot, P. (1982). *Keeping tracks of teaching: Assessment in the modern classroom*. London: Routledge & Kegan Paul.

Brualdi, A. C. (2002). Classroom questions. In L. M. Rudner & W. D. Schafer (Eds.) *What teachers need to know about assessment*. Washington, DC: National Education Association.

Cangelosi, J. S. (1990). *Designing tests for evaluating student achievement*. London: Longman.

Center on Standards & Assessment Implementation (2018). Valid and reliable assessments.

CSAI Updates. CSAI. Retrieved May 20, 2023 from CSAI-Update_Valid_Reliable_Assessments.pdf (wested.org)

Cizek, G. J. (1997). Learning, achievement and assessment: Constructs at a crossroads. In G. D. Phye (Ed), *Handbook of classroom assessment: Learning, adjustment and achievement*. 2-32. San Diego: Academic Press.

Cohen, L., Manion, L. & Morrison, K. (2000). *Research methods in education*. London: Routledge/ Falmer.

Coulby, D. (2000). *Beyond the National Curriculum: Curriculum Centralism and Cultural Diversity in Europe and the U.S.A*. London: Routledge Falmer.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*, 3-17. New Jersey: Lawrence Erlbaum Associates, Inc.

Darr, C (2005a). A hitchhiker's guide to validity. NZCER Set 2, 55-56. https://doi.org/10.18296/set.0639

Darr, C (2005b). A hitchhiker's guide to reliability. NZCER Set 3, 59-60. https://doi.org/10.18296/set.0623

Department of Education and Science (1988). *Task Group on Assessment and Testing: A report*. London: HMSO.

Dietel, R. J., Herman, J. L., & Knuth, R. A. (1991). *What does research say about assessment?* Oak Brook: NCREL.

Ebel, R. J. & Frisbie, D. A. (1991). *Essentials of educational measurement*. (5th ed.). New Jersey: Prentice-Hall.

Ecclestone, K. (1994). *Understanding assessment: A guide for teachers and managers in post-compulsory education*. Leicester: National Institute of Adult Learning- NIACE.

Eisner, E. W. (1993). Reshaping assessment in education: Some criteria in search of practice. *Journal of Curriculum Studies, 25* (30), 219-234.

Farris, P. J. (2015). *Elementary and middle school Social Studies: An interdisciplinary approach*. (7th ed.). Long Grove, IL: Waveland Press, Inc.

Ferrara, S. & McTighe, I. (1992). A process for planning more thoughtful classroom assessments. In A. Costa, J. Bellanca & R. Fogarty (Eds.), *If minds matter: A forward to the future* (pp. 337-341). Washington, DC: Skylight Professional Development.

Ghaicha, A. (2015). Theoretical framework for educational assessment: A synoptic review. *Journal of Education and Practice 7*(24), 212-231.

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment achievement and equity*. Buckingham: Open University Press.

Gross, R. E., McPhie, W. E., & Fraenkel, J. R. (1970). *Teaching the Social Studies: What,*

why and how. Pennsylvania: International Textbook Company.

Hanna, G. S. (1993). *Better teaching through better measurement*. Texas Harcourt Brace Javanovich College Publishers.

Haydn, T., Arthur, J., & Hunt, M. (2001). *Learning to teach history in the secondary school: A companion to school experience*. (2nd Ed). London: Routledge Falmer.

Herman, J. (1992). Accountability and alternative assessment. *CSE Technical Report 348*. University of California, Los Angeles.

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

Kelly, A. V. (2009). *The curriculum theory and practice*. (6th ed.). Thousand Oaks, CA: Sage Publications Ltd.

Lambert, D. & Lines, D. (2000). *Understanding assessment: purposes, perceptions, practice*. London: Routledge Falmer.

Linn, R. L., Baker, E. L., & Dunba, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *CSE Technical Report 331*, UCLA.

Linn, R. L. & Gronlund, N. E. (1995). *Measurement and assessment in teaching*. (7th ed.). New Jersey: Merrill

Maduas, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum, 87th year book of the National Society of the Study of Education*. Chicago: NSSE.

Mager, R. F. (1997). *Measuring instructional results: or got a match?* (3rd ed.). Atlanta: Center for Effective Performance.

McMillan, J. H. (2002). Fundamental concepts common to all assessment. In L. M. Rudner & W. D. Schafer (Eds.), *What teachers need to know about assessment*. Washington, DC: National Education Association.

Messick, S. (1996). Validity of performance assessment. In G. Philips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Educational Statistics

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. (3rd ed.), (pp. 13-104). New York: Macmillan.

Middleton, F. (2022). *The 4 types of validity in research: Definitions & examples*, Scribbr. Retrieved June 10, 2023 from https://www.scribbr.com/methodology/types-of-validity/

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University, 17*(3), 58-82.

Moskal, B. M., & Leydens, J. A. (2002). Scoring rubric development: Validity and reliability. In L. M. Rudner & W. D. Schafer (Eds), *What Teachers Need to Know about Assessment*. Washington, DC: National Education Association.

Mundrake, G. A. (2000). The evolution of assessment, testing, and evaluation. In J. Rucker (Ed.) *Assessment in Business Education, 38, NBEA Yearbook*. Reston: NBEA.

Nadasdy, P. B. (2011). Reliability and validity of a test and its procedure conducted at a Japanese high school. *Niigata University of International and Information Studies Repository*, 23-40. Retrieved May 13, 2023 from oai:nuis.repo.nii.ac.jp:00002601

Nelson, J. L. & Michaelis, J. U. (1980). *Secondary Social Studies: Instruction, Curriculum, Evaluation*. New Jersey: Prentice-Hall Inc.

Newton, P. E., & Baird. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice, 23*(2), 173-177.

Nitko, A. J. (1996). *Educational assessment of students*. New Jersey: Merrill.

Oosterhof, A. (1996). *Developing and using classroom assessment*. New Jersey: Merrill.

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, adjustment, and achievement* (pp. 55-68). New York: Academic Press.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know*. Needham Heights: Allyn and Bacon.

Pratt, D. (1994). *Curriculum planning: A handbook for professionals*. Fort Worth: Harcourt Brace College Publishers.

Quelmalz, E. & Hoskyn, J. (1997). Classroom assessment of reasoning strategies. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, adjustment, and achievement* (pp. 103-130). New York: Academic Press.

Rogosa, D. (1999). *How accurate are the STAR National Percentile Rank Scores for individual students? – An interpretative guide*. Palo Alto: Stanford University.

Rowntree, D. (1987). *Assessing students: How shall we know them?* London: Kogan Page.

Satterly, D. (1989). *Assessment in Schools* (2nd ed.). Oxford: Basil Blackwell Ltd.

Scottish Qualifications Authority (2001). *Guide to internal moderation for SQA centres*. Glasgow: SQA.

Sireci, S. G. (2015). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, DOI: 10.1080/0969594X.2015.1072084.

Stiggins, R. J. (2001). *The unfulfilled promise of classroom assessment*. Oregon: Assessment Training Institute.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(10). 534-539.

Stiggins, R. J. & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom*. New York: State University of New York Press.

Sutton, R. (1991). *Assessment: A Framework for Teachers*. London: Routledge.

Taylor, C. S., & Nolen, S. B. (1996). What does the psychometrician's classroom look like? Reframing assessment concepts in the context of learning, *Educational Policy Analysis Archives, 4*(17), 1-35.

Thompson, N. A. (2013). Reliability & validity. *Whitepaper- September, 2013*. Retrieved may 25, 2023 from *Test-reliability-and-validity.pdf (assess.com)*.

Tyler, R. W. (1949). *Basic Principle of Curriculum and Instruction*. Chicago: The University of Chicago Press.

West Africa Examinations Council (2002). *The Content Validity of WAEC Question Papers for the SSSCE in Ghana: 1997-2000*. Accra: Department of Research Division, WAEC.

Wiersma, W. & Jurs, S. G. (1990). *Educational measurement and testing*. (2nd ed.). Boston: Allyn & Bacon.

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass Publishing.

Copyright Disclaimer

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).