Nikola Dobrić
University of Klagenfurt, Austria

2018, Vol. 15 (2), 9–24(120)
revije.ff.uni-lj.si/elope
doi: 10.4312/elope.15.2.9-24
UDC: 37.091.26:808.1(091)

Reliability, Validity, and Writing Assessment: A Timeline

ABSTRACT

Looking at the issue of validity and test validation, the historical and the theoretical progression has been well described both when it comes to educational assessment in general and language assessment in particular. A clear progression can be seen starting in the 1920s and culminating in the late 1980s/early 1990s (with minor notable developments since), and it is an advancement motivated and driven almost solely by new theoretical and practical considerations. Securing validity and validation with regard to writing assessment in particular, however, took a more winding route and was primarily shaped by a power struggle between externally administered standardized testing (and the supporting administrative bodies) on one side, and the practicing teachers of writing at higher education institutions on the other. The paper at hand outlines this evolution and gives a timeline of the events and major developments that have fueled it and explores the cutting edge of today.

Keywords: language testing; writing assessment; historical development; overview; validation; validity; standardization

Zanesljivost, veljavnost in ocenjevanje pisanja: časovni pregled

POVZETEK

Vprašanje veljavnosti in vrednotenja testov je bilo z vidika zgodovinskega in teoretičnega razvoja že dobro opisano tako na področju splošnega pedagoškega ocenjevanja kot tudi na specifičnem področju ocenjevanja jezikov. Začetek razvoja sega v dvajseta leta preteklega stoletja in doseže vrhunec v poznih osemdesetih oziroma v začetku devetdesetih let 20. stoletja (z manjšim opaznim napredkom tudi po tem času). Gibalo razvoja so predvsem nova teoretična in praktična dognanja. Zagotavljanje veljavnosti in validacije ocenjevanja pisanja pa je potekalo po bolj ovinkasti poti in predvsem pod vplivom boja za premoč med zunanjimi standardiziranimi testi (ob podpori administrativnih teles) na eni strani ter visokošolskimi učitelji pisanja na drugi. Prispevek prikazuje ta razvoj in podaja časovni pregled dogodkov in večjih dosežkov, ki so ga omogočili, hkrati pa tudi raziskuje današnje najsodobnejše smernice.

Ključne besede: jezikovno testiranje; ocenjevanje pisanja; zgodovinski razvoj; pregled; veljavnost; validacija; standardizacija

CC BY SA

Reliability, Validity, and Writing Assessment: A Timeline

1 Introduction – Indirect Assessment of Writing

If we leave aside the writing examinations one reads of starting in Ancient China almost two millennia ago (as a part of the rigorous assessment of civil servants underway during the Han Dynasty in the 2nd century AD) as too far removed contextually and in time, and somewhat historically inaccessible, we can say that the earliest more contemporary records dealing with the assessment of writing in the Western countries emerged in the mid-19th century. At the time, the turn was being made from the Western tradition of open debate (stemming originally from the Ancient Greek philosophical tradition) towards a more standardized model of uniform examinations (actually conceptually influenced by the previously mentioned Chinese tradition). In essence, across the Western world universities usually favored giving oral-based examinations, a tradition stemming from the Socratic approach to higher education. However, because of the displacement created by the Industrial Revolution, school-age children were suddenly being forcibly put behind school desks (a result of new compulsory education laws), a development which demanded a faster and more resource-friendly way to test larger numbers of students. One of the first actual records of written examinations coming up within a national dialogue is one by Horace Mann calling for written tests to replace oral examinations in Boston schools in 1840 (Huot, O'Neill, and Moore 2010, 496).[1] In the 1840s oral tests were the standard in American schools, and the move was to introduce written tests in order to facilitate objectivity and reduce bias. Additionally, he also wished to build a transparent (standardized) system according to which one could compare the quality of different schools and teachers. As a result of this and similar calls, written tests were relatively quickly introduced across the United States, such as for example the *Harvard's Exams in Writing* in 1874. With this a new writing component was added to the Harvard University entrance exam, which included a short, speeded essay rated by the resident teachers, and was based on 'such works of standard authors as shall be announced from time to time' (Hobbs and Berlin 2001, 252). However, not long after the launch of the first essay-based writing exams at American universities came critical voices. On the one hand, the criticism focused on unreliability and subjectivity in terms of rating the essays (ironically enough, the same criticism that motivated the switch from oral examinations in the first place), and their lack of practicability and usability in a situation of increased numbers of students to process, on the other hand. Already in 1880 we have first publications highlighting the problems of unreliability that were inherent in the early essay-based writing examinations (Huddleston 1952). For example, the study published by Hopkins in 1921 demonstrated that the score a student received as a part of the College Board exam statistically depended more on the rater and on the administration conditions (when and where the exam took place) than on the actual writing performance, a situation also found in many European countries administering writing tasks as part of high-stakes examinations (Nikolov 2009). The Certificate of Proficiency in English (CPE) was instituted in London in 1913 by Cambridge Assessment (an organization that went on to become the leading international English language examining institution in the

[1] The history of the development of standardized testing and its influence on writing assessment is best documented within the American framework, and hence the brief historical overview will mostly follow the story in this context, with a note that the development in other Western countries took a similar form with the exception of the UK, where the emphasis on direct methods of writing assessment and a focus on (context) validity was always more present.

world), and with its emphasis on essay (composition) writing was to face similar criticism. These very real issues were further emphasized not just by a continuous academic scrutiny (in a number of studies, such as Starch and Elliott (1912), Hopkins (1921), Sheppard (1929), and others), but also by both demographic and political developments impacting higher education in the USA at the time. Demographically, the numbers of pupils and students in the US rose exponentially, because starting from 1852 the different states began adopting the recently approved universal (and obligatory) public education laws. Politically, this fueled the foundation of the College Entrance Examination Board (CEEB) in 1899 as a not-for-profit organization aiming to expand access to higher education. And while this did mark a very important development, certain actions regarding organization of assessment within higher education (such as the centralization and outsourcing of language testing) created a massive power struggle and by extent had a great impact in terms of how validation of writing assessment developed in the subsequent century.

Backed by academic publications pointing out the severe problems of the unreliability of essay-based assessment of writing, weighed on by the increased numbers of students in need of processing, under pressure from high school representatives for a more standardized arrangement of entrance examinations into universities, and somewhat swayed by internal power issues, the decision was made by the CEEB to replace the then-perceived unreliable direct writing assessment (essays) with a much more reliable (though, of course, much less valid) indirect method. Unlike the US, the socio-economic context in the UK did not demand such an increase in processing power, and was able to maintain a much larger focus on context validity and maintain (for CPE unwaveringly for more than a century now) direct assessment of rating as a predominant method (Weir, Vidakovic, and Galaczi 2013). Nevertheless, assessment of writing in higher education institutions in the USA (and many other countries) became as of that point almost solely conducted in the form of multiple-choice tests and, to a large extent, mostly outsourced to private institutions as well. This was to remain so up until the 1960s, as the issue of unreliability came up periodically and repeatedly in academic publications as the major factor in favor of indirect methods (such as Traxler and Anderson (1935), Stalnaker (1936), Coward (1952), and others). Fortunately, this emphasis on continuously pointing out unreliability of direct writing assessment was to backfire in 1961 with a notable publication by Diederich, Frech, and Carlton. Their study dealt with the problems of readers (raters) not being able to agree in terms of rating writing performances, identifying that as the major source of error when it comes to score stability (reliability) and, ultimately, fairness of direct writing examination. In brief, they asked 53 raters to score 300 writing performances on a 9-point scale, with the results of 94% of papers receiving at least seven different scores (and the inter-rater agreement peaking at only 0.31). Looking from today's perspective, we can easily see the flaws in the study as readers were bound to disagree seeing that they worked without any guidance, common training, or shared point of reference (as was pointed out even at the time in Braddock, Lloyd-Jones and Schoer (1963)). However, the results of the study regarding the inter-rater (dis)agreement are not the reason why this earlier work is to be considered so relevant. This importance actually stems from the fact that it got the academic community thinking about what criteria raters actually focus on when rating written work. This is because the study, apart from looking at the agreement, also collected some eleven thousand introspective reports from the raters, revolving around the features they were looking at when rating the performances (coming up, via factor analysis, with *ideas*, *form*, *flavor*, *mechanics*, and *wording* as the most salient features). This particular aspect of looking beyond just inter-rater agreement as an expression of reliability into issues such as the features of writing performance that influence rating, rater scores, and rater training and experience, was to establish a new base-line for researching writing assessment that was to mark

the next stage of the development in the field (Huot, O'Neill, and Moore 2010, 502). Studies moving towards establishing proof that direct writing assessment can produce reasonably reliable scores (such as Godshalk, Swineford, and Coffman (1966) and their outline of holistic scoring, for example) further opened the door to a generally increased focus on direct writing assessment (but not, as we shall see, to a much needed shift of focus from reliability to validity).

2 Direct Assessment of Writing – Focus on Reliability

Despite these developments and motions, the CEEB fought back all the way up to the late 1980s to preserve multiple-choice tests as the foundation of standardized (and usually externally managed) assessment of writing at universities and colleges in the USA (such as the well-known TOEFL examinations). The struggle was, however, futile, as the push in the other direction (towards direct examination of writing proficiency) was once more coming both from political and academic quarters (Breland 1983, 2). Politically, one of the issues was that the overwhelming focus on indirect assessment of writing was extremely worrying, because it technically led to hardly any writing being taught within the American educational system by the mid-1970s (Applebee 1981). Another political issue was the pressure coming from the English Departments at higher education institutions across the country. Rankled by both the feeling of disempowerment caused by the prescribed external testing services and the feeling that multiple-choice testing of writing did not really conform to their experiences as teachers of English and writing of what writing ability represented, a firm stand was made to change things. Within the US, we can trace the first major battle being fought in 1971 and the English Departments at California State University, the staff successfully rejecting the institutionalization of an externally administered multiple-choice test for their first year English equivalency exam (White 2001, 308), following this up with a creation of their own locally administered examination. Academically, with authors such as Paul Diedrich changed their stance towards direct writing assessment, and similar voices coming collectively from other publications (such as Brown (1978), Coffman (1971), Cooper (1977), Cooper and Odell (1977), and others), the move was a clear one towards the abandonment of the indirect method of assessment of writing ability. Having a wealth of information on where the source of reliability problems is to be found when it came to rating written performance, the two main focuses being on rater inconsistency and sampling bias, the majority of studies in this period revolved around the attempts to address these. With a better understanding of the rating process and what it entailed, there was a rise in research on scoring (such as that related relating to analytic scoring with Diedrich (1974), holistic scoring with Godshalk, Swineford, and Coffman (1966) or Cooper (1977); primary-trait scoring with Mullis (1980); or syntactic scoring with Hunt (1977); writing scales in Jacobs et al. (1981); Huot (1990) or Ericsson and Simon (1993); rater training and agreement with McIntyre (1993) or Weigle (1994); and other facets of score stability).

The fact that the socio-economical and academic conditions were favorable for direct assessment of writing to be promoted meant that by the mid-1980s most universities (and their language departments) were already implementing locally administered direct tests of writing ability in much the same way. This was a very positive step, as it allowed for tests to be developed by the faculty that worked together within the same program for their own purposes, with their own goals and within the shared system and shared curriculum, and yet also able to maintain reliability at an acceptable level. However, as we mentioned, while the switch to performance-based testing of writing was a very important development, what did not change (but should have) was the focus, which needed to shift from practicability of testing and reliability issues to much more important questions of aspects of validity unrelated to scoring (mostly present within the UK tradition, though).

3 Direct Assessment of Writing – Focus on Validity

Validity in terms of assessment indicates that the results obtained from the given measurement procedure objectively reflect the phenomenon the said procedure is intended to measure, and that the measurement at hand has not been obtained due to any measurement-irrelevant factors or chance). In this respect validity depends on the degree to which quantified measurements of presumptive behavior or ability are blurred by other factors (Sigott 2004, 44). Different from reliability, which can be seen as the quality of the data collected, validity is the quality of the inferences (and decisions) we can make following the measurement (Chan 2014, 9). In this sense validity also depends on the degree as to which aspects of the said behavior or ability which the test is supposed to measure are covered by the given test (Sigott 2004, 44). Validation is the process in which we gather and evaluate evidence to support the said appropriateness, meaningfulness, and usefulness of the inferences and decisions we make based on measurement scores (Zumbo 2007; 2009). Unfortunately, one of the products of the described power struggle between the (external) testing providers (who were seen as one of the actors advocating the indirect writing examinations) and the public bodies (such as the CEEB) on one side and the researchers and practicing teachers of writing (who felt unheard and disempowered) on the other, was that writing assessment did not catch up with the theoretical developments revolving around validity that had actually been under way in educational and language testing since the 1950s. Although there were voices calling for more insight into the content of the tests (the likes of Wiseman (1949) or Wiseman (1956)), most of the research on the assessment of writing taking place between 1960 and 1990 was into the reliability of the direct methodology of writing assessment (the same as during the previous stage of development regarding indirect methodology). The focus did not appropriately really shift towards validity until the early 1990s.

The earliest accounts of validity considered within the framework of language testing can be traced to authors such as Lado (1961) and Davies (1968)[2]. Their account of validity focused on the face values of the test (the so-called validity by assumption), on the content of the test, on control of extraneous factors, on conditions required to answer test items, and on empirical insight (D'Este 2012, 63). At the same time, Campbell and Fiske (1959) described the validity of language tests as having a *convergent* dimension (measures that should be related are perceived as related) and a *discriminate* dimension (measures that should not be related are found to be not related). Campbell and Stanley (1966), as one more early account, featured the concepts of *internal* and *external* validity. It is only decades later that we can see perhaps the most influential account of validity within language testing, that of Bachman (1990). It was written as a reaction to the unified theory of validation famously put forward by Messick in 1989, which Bachmann embraced, but went on to identify three factors that support this overall validity (D'Este 2012, 67): *content relevance* and *content coverage* (what is known as content validity); *criterion relatedness* (i.e. criterion validity); and meaningfulness of construct (construct validity). *Content validity* here refers to the domain specifications which underlie the test, *criterion relatedness* refers to a meaningful relationship between test scores and other indicative criteria, while *construct validity* relates to the extent to which performance of the test is consistent with the predictions we make on the basis of the theory of abilities (Bachman 1990, 246–69). In addition, Bachman also imported the *consequential* (or *ethical*) *basis* of validity from Messick and other authors, which

[2] While the earliest considerations of validity within psychological and educational settings trace back earlier, with beginnings in the 1920s (and authors such as Scott (1917), Thorndike (1918), Ruch (1929), or Tyler (1934)), with the major movement here starting in of the 1950s (and the work related to the first Standards of Educational and Psychological Testing of 1955).

refers to the fact that tests have not been designed to be used in an academic vacuum but rather have real-life applications and are influenced by society as a whole. The final important notion, as argued by Bachman, was of the inclusion of the concept of reliability within validity itself. His argument was that when it comes to language testing of any kind (and it so especially in the case of writing assessment), it is not easy to distinguish between effects of different test methods or between traits and test methods (1990, 239).

Stemming from Bachman's account, the 1990s and early 2000s saw several related discussions and models of validity in language assessment, such as Kane (1992; 2001), Sigott (1994), Alderson, Chapman and Wall (1995), Miller and Linn (2000), and Wier (2005). If we were to take as an example, as the most recent and most language-assessment-related account, Weir's socio-cognitive model of evidence-based test validation (2005, 17–37), this presents overall test validity as an interplay between five different types (discussed in more detail further on): influenced by test-taker characteristics, there are cognitive validity, context validity, scoring validity, consequential validity, and criterion validity. Cognitive validity has as its emphasis the determination of the cognitive processes which are to be used as a model for designing test items. Context validity follows the idea of moving away from the sole focus on linguistic representation and including the social and cultural dimension within which the writing performance has been produced (Weir 2005). Scoring validity, in line with Bachman's already discussed elaborations on reliability belonging within validity, focuses on all aspects of the assessment that could have an impact on the scores. Finally, criterion-related validity in this account keeps the focus as traditionally established, and revolves around a comparison to any reliable external measurements.

4 Direct Assessment of Writing – Discussion

What we can see if we dissect all of the accounts presented above, from Bachmann onwards, is the contemporary view of validity and validation in language testing understands that validity is a unified concept (which includes reliability), and is an aspect of a test's use to be regarded as a whole. However, we can also see a practical need for a focus on particular individual aspects (types) of validity that should be covered. These individual aspects reflect sources of evidence for validity that we can tap and steps to be taken within any validation effort (regardless of the theoretical preference of the terminology marking the given different aspects. This is clearly evident in terminological disagreements such as construct validity vs. validity, division on content, criterion, and consequential validity, and so on). Applying this kind of a practice-driven (authenticity-driven) model to direct writing assessment, we can start with understanding it as being *strong* or *weak* (McNamara 1996). Direct writing assessment in a strong sense incorporates tasks which represent real-world performance and are judged according to real-world criteria, the focus being on the successful fulfilment of the task and not (only) on language proficiency. Direct writing assessment in a weak sense puts the focus on language use, where the task does indeed resemble a real-world purpose, but the goal is only to extract a writing performance for the purpose of ascertaining language competence. It is a distinction that influences everything from the construct to the scoring, and it is safe to say that in most cases writing assessment in higher educational setting takes the form of the latter (Tsai 2002, 1). Once a choice is made in this direction, further conceptualization (and retroactively understanding) of the construct (including all other elements of validity as well) comes essentially at one of the three stages (Bachman and Palmer 1996; Weigle 2002): *design* stage, *operationalization* stage, and *test administration*.

The design stage is extremely important for ensuring validity and it should, according to McNamara (1996, 43), revolve around sampling of the task from the communicative situation the test is to be a proxy of – this would include consulting expert informants, examining available literature, analyzing and categorizing communicative tasks, collecting and examining relevant texts, and deciding on a broad test method (Tsai 2002, 2). In the operationalization stage information gathered in the design stage is transformed into concrete test specifications and detailed procedures the test takers and readers (raters) are to follow. These specifications should generally include the description of the test content, criteria for correctness, and a sample of tasks (Douglas 2000), and comprise of scoring rubrics, writing prompts, rating scales, and similar (Weigle 2002). Finally, the test administration phase focuses on the actual tests being administered, where the validation revolves around the data that stem from the elicited writing performances (such as the scores and the washback information), while the stage is firmly footed in different statistical methods. This division into practical stages of test development and administration is useful, as it points to the different steps which should be taken at different stages in order to ensure the validity of a particular interpretation and use of a particular test, and to the fact that a useful way to observe the process of validation is to imagine it as *a priori* (before the test administration) and *a posteriori* (after it has been administered). Additionally, it is useful to observe validity in terms of the types of evidence that support it, along the lines of Bachman's division into evidential and consequential bases of validity (1990, 248)[3]. Finally, the interplay between validity and reliability is also worth discussing, together with how Bachman (following other similar voices such as Loevinger (1957), Messick (1989) or Cronbach (1990)) argues that the two are in fact complementary aspects of identifying, estimating, and interpreting different sources of variance in test scores (1990, 239).

4.1 *A Priori* Validation

Unlike the subsequent stage of *a posteriori* validation that is largely grounded in quantitative data and statistical methodology (and which is usually feared and avoided by practicing teachers), *a priori* validation efforts are usually conducted on a regular basis by most practitioners. The reason is that *a priori* validation takes places during the test compilation phase, and most commonly includes the very logical, common-sense, steps anyone seriously compiling an assessment tool would think of (though it can also incorporate some statistics-based operations). *A priori* efforts basically focus on the commitment of the test designer to make sure that he or she understands the behavior or performance which is the target of the assessment, that the content of the test is relevant to the real-life behavior or performance at hand, and that the criteria of success (correctness) are clear both to the test-taker and the test-rater. This stage of validation is hence much less opaque to non-experts and, as indicated, covers the stages of test design and test operationalization.

Starting always with the understanding of the behavior or performance of interest, the first step in test design (and *a priori* validation) is the consideration of construct validity in the narrow sense. As we noted previously, a construct can be seen as a definition of people's attributes that are assumed to be reflected in their performance (Cronbach and Meehl 1955, 283). This means that to properly measure the 'extent' of someone's writing ability we need to understand the nature of that ability as it is, in our minds, and we need to be able to describe it in a sufficiently

[3] Evidential basis of validity, according to Bachman, is grounded in evidence that supports the relationship between test scores and their interpretations and subsequent use (1990, 248). The consequential (or ethical) basis of validity refers to the fact that tests have not been designed for use in an academic vacuum, but rather have real-life applications (Bachman 1990, 280).

clear manner. Hence, construct validation is related to the basic question of what is the nature of that something that an individual possesses or displays that is the object of our measurement (Messick 1975, 957). The part that takes place *a priori* (sometimes referred to as the *logical* part of construct validation) involves first defining the construct (the ability) theoretically, or rather scouring the existing literature on descriptive accounts of writing ability. Accounts such as these try to capture the inner-workings of the cognitive processes at the basis of what we could call 'an ability to write' (and often perhaps the 'ability to write in a foreign/second language' as well). Sometimes this is quite challenging, as writing assessment as a discipline has not really excelled at writing about writing (Huot 1990), but rather at the more practical work on assessment methodology (such as writing scales and rater training procedures). That is why there is a feeling that theoretical accounts of writing competence as a phenomenon (unlike accounts of language competence in general) are somewhat lacking, and that more work still needs to be done in this direction (Yancey 1999). Nonetheless, finding (or setting up) a comprehensive and meaningful theoretical account of the construct in the design stage allows us to perform many actions which can assure subsequent high levels of validity. Understanding the construct well allows us to identify the sources for the most meaningful sampling of test content, and to identify which test method would be best to elicit behavior (performance) which would most resemble the construct at hand and the relevant real-life performance. These developments we would carry over to the operationalization stage, where the logical (theoretical) part of construct validation provides us with means of linking the actual ability we are measuring with the content of the test and with the scores that end up quantitatively representing our measurement. Adjacent to this understanding of the ability itself is the insight into the metacognitive strategies (or strategic competences, as termed in Bachman and Palmer (1996)) seen as mediating between the trait and context, and comprising of faucets such as goal setting, assessment of needs to achieve the purpose, planning, monitoring, and organization of language and topic – all contributing, in fact, to what Wier refers to cognitive validity (2005, 86). In terms of writing assessment, Shaw and Weir (2007, 34) recognize six different aspects of cognition behind writing ability:

- macro-planning: gathering of ideas and identification of major constrains such as genre, readership, and goals;
- organization: ordering the ideas and identifying relationships between them;
- micro-planning: focusing on individual parts of the text and considering issues such as the goal of the paragraph in question, including both its alignment with the rest of the text and the ideas and the sentence and content structure within the paragraph itself;
- translation: the content previously held in a propositional form is transferred into text;
- monitoring: focusing on the surface linguistic representation of the text, on the content and the argumentation presented in it, and its alignment with the planned intentions and ideas; and
- revising: results from the monitoring activity and involves fixing the issues found to be unsatisfactory.

Alongside the cognitive aspects behind writing as a performance, Weir also emphasized the need to consider the test-taker actively in terms of personal individualities. He distinguished between three classes of test-taker characteristics (Shaw and Weir 2007, 5):

- physical/physiological characteristics, which include any special needs on the side of the test-taker, such as those stemming from dyslexia or eyesight impairment;

- psychological characteristics, including test-taker motivation, personality type, learning styles, and more; and
- experiential characteristics, incorporating factors such as the degree of test-taker familiarity with the test format or content.

These individual characteristics can then be viewed as *systematic*, if they affect a test-taker's performance consistently (such as dyslexia or personality traits) or *unsystematic*, when they have a random, perhaps one-off effect (for example, motivation or test format familiarity).

Once we have a reasonable grasp of the nature of writing ability (and a complete grasp in theoretical terms is pretty hard to achieve when it comes to social sciences and humanities), the test-takers' characteristics at play, and the cognitive processes taking place in our minds while writing, then while still in the design stage we need to tackle the context surrounding the test. Related to the social and situational background (Weir 2005, 57), this revolves around the task setting and conditions necessary to ensure appropriateness of the test content, clarity and conformity to the construct, the intended use, and the intended stakeholders (Shaw and Wier 2008, 63):

- setting = task: includes the *expected task format* (consideration the genre), the *purpose* (expected form and communicative function or the nature of information in the text), the *clear knowledge of the criteria* (task instructions), the *weighing* (focusing on the relative contribution of the different parts/aspects of the test), the *text length*, the *constraints* (i.e. speededness, additional resources, and more), and the *expected writer-reader relationship* (addressee information);
- setting = administration: incorporates *physical conditions* (venue, background noise, lighting, and more), *uniformity of administration* (same specifications for all test takers), and the *security* involved (controlled access); and
- linguistic demands = text input and output: the focus is on the lexical resources (vocabulary), structural resources (morpho-syntax), the discourse mode (considerations of genre, rhetorical task, and pattern of exposition), the functional resources (referring to the fulfillment of the intended communicative purpose of the writing), and the content knowledge (background and the subject matter knowledge).

Also important to mention here is that, unlike most of the other *a priori* validation efforts, ascertaining validity of contextual (content) features such as the appropriateness of the lexical resources, structural resources, or the discourse mode can be undertaken empirically (quantitatively) as well.

The last *a priori* puzzle-piece in the jigsaw that is validation (and ultimately the degree of validity) of a particular issue of writing assessment is in the operationalization stage, where we define instructions on how to, essentially, obtain reliable scores. Called scoring validity by Weir (2005, 117), Shaw and Weir list the *a priori* considerations as follows (2007, 146):

- criteria and type of the rating scale: focuses in essence on the marking scheme and the different approaches to assigning a number to the measurement (i.e. *primary trait* scoring vs. *holistic* scoring vs. *analytic* scoring);
- rater characteristics: identified as one of the biggest causes of score variability when it comes to writing assessment; the rater is observed both in terms of the rater-candidate and the rater-item interactions (again dividable into *physical/physiological*, *psychological*, and *experiential* characteristics);

- rating process: different prescribed rating procedures (such as for example the *principled and pragmatic two-scan approaches*, the *read through approach*, or the *provisional mark approach*);
- rating conditions: includes the setting (on site or at home, for example), the medium of the writing performance (handwritten vs. electronic), time constraints (deadlines), and scaffolding (ways in which examiners are advised); and
- rater training: perhaps the most crucial aspect in achieving higher levels of inter-rater agreement (lowering the influence of severity and other effects that stem from the rater).

After having completed all of listed actions, and ultimately having arrived at a version of the test that in its content and specification reflects entirely the theoretical description of the behavior (ability) we are interested in measuring, it is time to administer the test and then see how much the results (scores) obtained do in fact conform to the expectation of validity we have built up in the design and operationalization stages.

4.2 *A Posteriori* Validation

One of the first aspects of validity that most research into test interpretation employs after the test has been administered is generally termed *criterion validity*. As indicated, it revolves around correlations with other comparable measures of the same ability. This is one of the go-to procedures in smaller-scale validation efforts, as it is simpler to set up in comparison to other procedures – all one needs is an established comparable existing (or future) measurement which is already considered as valid, to which you can make comparisons in terms of obtained measurements (scores). Traditionally, we find two kinds of criterion-based validity: concurrent and predictive (Bachman 1990, 248). Concurrent criterion relatedness involves one of the several commonly employed procedures: examining differences in test performance among individuals at different levels of proficiency, examining correlations among different measurements of the given ability, correlations with existing standardized tests, comparisons with teacher ratings, with informant data, with self-assessment, and with different versions of the same test (Weir 2005, 209). Predictive criterion relatedness focuses on demonstrating a link between test scores and some future performance, where the test scores predict the criterion behavior of interest – for example, having a writing test serve as a predictor of place allocations in writing courses of different levels, or linking predictive criterion validity to comparison with external performance benchmarks such as the CEFR, for example (Weir 2005, 209). In particular, for the purposes of writing assessment, Shaw and Wier list the following three procedures (2008, 230):

- cross-test comparability: the procedure encompasses the effort of comparing different and yet related language proficiency measures (such as comparing the ratings of two different tests and ascertaining the levels of their relatedness);
- comparison with different versions of the test: relies on the existence of a relationship between two or more different versions of the test involving the same test takers and other comparable conditions; and
- comparison with external standards: using standardized and reputable existing measurements (such as national graduation (Matura) exams or well-known British Council or ESOL examinations) as benchmarks for making comparisons to.

The statistical methodology of pursuing further aspects of *a posteriori* validity then becomes increasingly more complex than just looking at the correlation coefficients one would normally find in criterion-related validity. This is the reason why we mostly see them in more serious

validation efforts, starting with the focus on scoring validity in an empirical sense. *A posteriori* analysis of the obtained scores looks into rating scales, inter- and intra-rater agreement, moderation of scores, and then beyond into use and consequences (Weir 2005), and as such involves complex dealing with quantitative data. *A posteriori* analysis of scoring validity involves focusing on two major aspects (Shaw and Weir 2008, 190) that are also strongly grounded in statistical methodology:

- post-exam adjustments: they include statistical methods (such as scaling or Rasch measurements) aimed at artificially correcting scores based on the attested interplay of different features surrounding the test (severity and leniency, central tendency, and more); and
- grading and awarding: the final process is the one of assigning cut-off scores and issuing certificates (if applicable).

However, one can argue that this part of empirical validation deals much more with consistency of measurement (and hence reliability) than what we would normally understand as validity, even if we look at them as merged. Hence, most *a posteriori* efforts at validation in the strictest sense revolve around construct validation in its narrow meaning. Empirical construct validation focuses on finding empirical evidence (in terms of correlations and experimentation) which goes towards confirming or disproving a particular interpretation of the obtained test scores (Bachman 1990, 260). It functions in terms of two very broad approaches at gathering validity evidence – correlational (and other psychometric methods) and experimental approaches. The correlational approaches can be seen as either exploratory or a confirmatory. The exploratory approach concentrates on identifying the traits that influence test performance by examining the correlations[4] among a large set of measures. The confirmatory approach focuses on a particular hypothesis about the traits and attempts to confirm or reject it. On the other hand, the classical experimental design for testing hypotheses (which has up to recently received less attention) involves attempts at manipulating the variables involved in a testing situation, such as example pre-test/post-test experimentation, and similar (Bachman 1990, 258–66).

Finally, looking beyond the test and the scores, we have the social effect a test use may have, usually observed within the framework of consequential validity. This involves us considering the ethical basis of validity as incorporating deliberations which are neither scientific nor technical, and which focus on the influence of a particular (educational) system on the interpretation of a test, as well as on the washback effect that test use has on that particular system in reverse (1990, 279). In practice, this means observing (see Bachman (1990, 280–84) and Weir (2005, 210)):

- the rights of the test-takers (secrecy, confidentiality, privacy);
- the values inherent in test developers and raters;
- the values inherent in the particular social system;
- background knowledge; and
- influence on teaching and learning.

[4] Related psychometric approaches also include factor analysis, multi-trait-multi-method investigation, and more.

5 Direct Assessment of Writing – The Cutting Edge and Outlook

If we review the entire discussion presented in the previous two sections, looking at both the general push and pull within development of writing assessment and at the issues of reliability and validity within language testing in general, and assessing writing in particular, we can with some degree of confidence suggest a global outline of what validity in terms of direct assessment of writing entails, and how it can be supported:

- validity entails the application of scientific rigor to the process of interpretation of test scores by focusing on both rational argument and empirical evidence;
- even though it can be observed as a conceptual whole, validity is in essence comprised of different aspects and this division is extremely useful when it comes to the practical (empirical) conduct of validation efforts;
- reliability is to be considered as an integral part of overall validity, and especially so when it comes to writing assessment;
- writing as a language skill deserves special consideration when it comes to its theoretical description within the larger concept of language competence; and
- equal attention should be given to the *a priori* and *a posteriori* efforts of ensuring validity.

If we were to translate this general overview, emerging from the overall discussion, into practical steps needed to assure the highest possible degree of validity, we can dilute them into a validation checklist, as follows in Figure 1 below (largely revolving around and adapted from the Shaw and Wier (2007)). Following a list such as this one can lead any practitioner to easily pinpoint most of the steps they need to take when setting up the assessment, and can/should take after they have administered the test, in order to ensure the highest level of validity in the context of the notoriously problematic phenomenon that is the assessment of writing.

In the end, when it comes to further development of securing validity within assessing writing ability, it is hard to foresee any new theoretical movements in the future. The core deliberations regarding the nature of validity in educational (and psychological) testing took place between 1950s and 1990s, culminating in the extensive account presented by Mesick in 1989. Similarly, Bachman's (1990) translation of Messick's (1989) influential account into the framework of language testing basically set the ultimate theoretical underpinnings of validity for this particular field, and even added the somewhat novel consideration of incorporating reliability within validity. However, Bachman's account cannot be seen as bringing anything conceptually new with regards to Messick (perhaps only adding the said notion of reliability being merged with validity), nor can any other account seen as coming after his. The real contribution he and other of the authors discussed in this paper made was in furthering the practice of validation and fine-tuning the outline of the sources of evidence and procedures one should focus on to address different aspects of unified validity. This is the direction the development of validity and validation research in terms of language assessment, and in particular that of writing, will most likely follow in the future, with conceptual novelty highly unlikely in this content (apart from any possible changes related to our understanding of the construct itself).

One of the areas that will see particular development is the validation of scoring, both in an *a priori* and *a posteriori* sense. *A priori*, the development will most likely focus on rating scales and

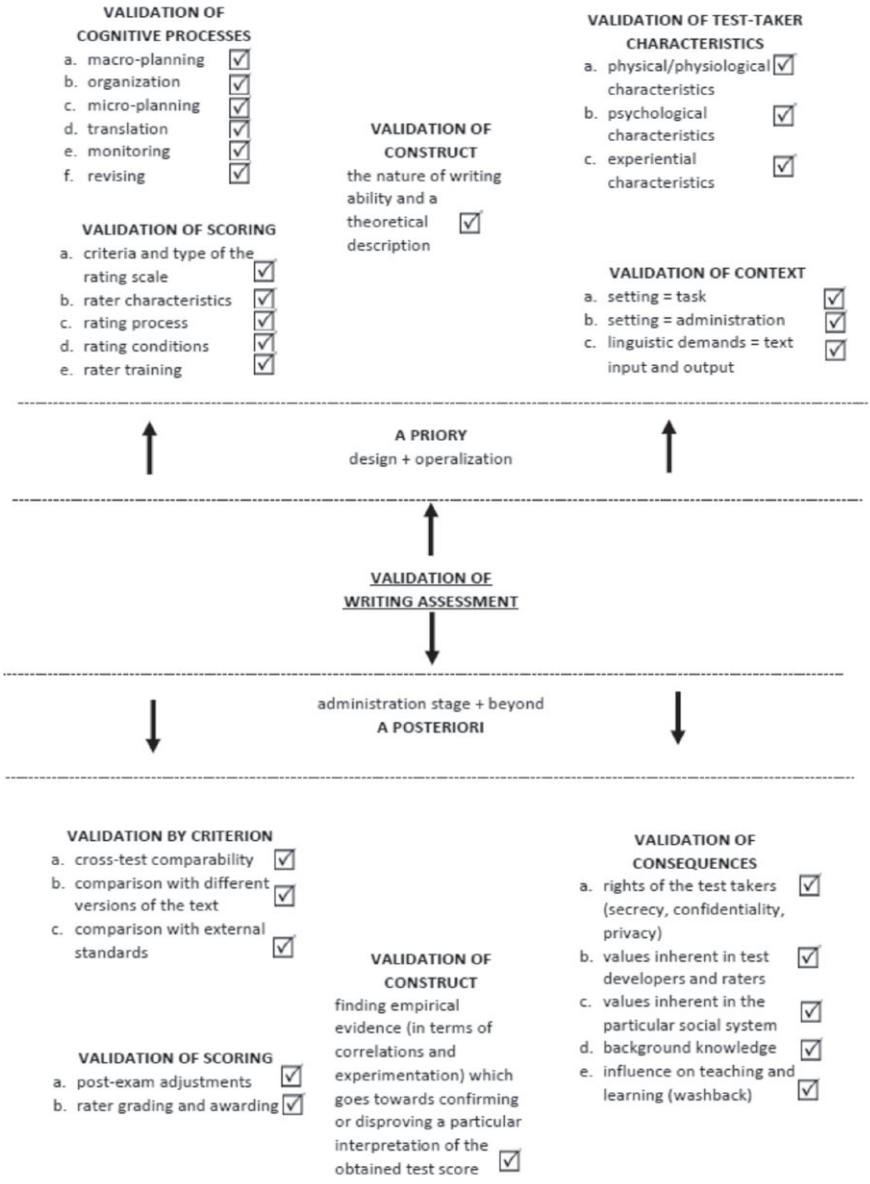

Figure 1. A summary of the prototypical validity effort of any writing assessment.

their application. For instance, one of the more interesting directions the advance of rating scales is taking is in moving away from the vague 'can do' and 'has got' descriptors often found in rating scales to the actual (linguistic) features being weighted and ticked off as present (or not) in relation to ratings. This represents the first step of actually turning away from a purely 'judging' perspective predominant in rating writing performances to a more 'counting' oriented one. Likewise, *a posteriori*, the shift to a more 'counting' oriented approach would mean the introduction and reliance on new methods of score adjustment, quality assurance (via agreement studies), computer-aided processes being introduced on a wider scale, eye-tracking investigations, and more.

References

Alderson, J. Charles, Caroline Clapham, and Dianne Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

American Psychological Association. 1955. *Standards for Educational and Psychological Testing and Manuals*. Washington, DC: Author.

Applebee, Arthur N. 1981. *Writing in the Secondary School: English and the Content Areas*. Urbana, IL: National Council of Teachers of English.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, Lyle F., and Adrian S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Breland, Hunter M. 1983. *The Direct Assessment of Writing Skill: A Measurement Review. College Board Report No. 83–6*. New York: Educational Testing Service.

Brown, Rexford. 1978. "What We Know Now and How We Could Know. More about Writing Ability in America." *Journal of Basic Writing* 1 (4): 1–6.

Campbell, David, and D. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (2): 81–105. https://doi.org/10.1037/h0046016.

Campbell, David, and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Boston: Cengage Learning.

Chan, Eric K. H. 2014. "Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments." In *Validity and Validation in Social, Behavioral, and Health Sciences,* edited by Bruno D. Zumbo and Eric K. H. Chan, 9–24. New York: Springer.

Coffman, William E. 1971. "On the Reliability of Ratings of Essay Examinations in English." *Research in the Teaching of English* 5: 24–36.

Cooper, Charles R. 1977. "Holistic Evaluation of Writing." In *Evaluating Writing: Describing, Measuring, Judging,* edited by Charles R. Cooper and Lee Odell, 3–31. Urbana: National Council of Teachers of English.

Cooper, Charles R., and Lee Odell, eds. 1977. *Evaluating Writing: Describing, Measuring, Judging*. Urbana: National Council of Teachers of English.

Coward, Ann F. 1952. "A Comparison of Two Methods of Grading English Compositions." *Journal of Educational Research* 46 (2): 81–93. https://doi.org/10.1080/00220671.1952.10882003.

Cronbach, Lee J. 1990. *Essentials of Psychological Testing.* 5th Ed. New York: Harper & Row.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281–302. https://doi.org/10.1037/h0040957.

Davies, Alan, ed. 1968. *Language Testing Symposium. A Psycholinguistic Perspective*. London: Oxford University Press.

Davies, Alan, and Catherine Elder. 2005. "Validity and Validation in Language Testing." In *Handbook of Research in Second Language Teaching and Learning,* edited by Eli Hinkel, 795–813. Mahwah, New Jersey: Lawrence Erlbaum Associates.

D'Este, Claudia. 2012. "New Views of Validity in Language Testing." *EL.LE* 1 (1): 61–76. https://doi.org/10.14277/2280-6792/5p.

Diederich, Paul B., John Winslow French, and Sydell T. Carlton. 1961. "Factors in Judgments of Writing Ability." *ETS Research Bulletin* 1961 (15): i–93. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x.

Diederich, Paul B. 1974. *Measuring Growth in English*. Urbana: National Council of Teachers of English.

Douglas, Dan. 2000. *Assessing Language for Specific Purposes*. Cambridge: Cambridge University Press.

Ericsson, K. Anders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Godshalk, Fred I., Frances Swineford, and William Eugene Coffman. 1966. *The Measurement of Writing Ability*. New York: College Entrance Examination Board.

Hobbs, Catherine L., and James A. Berlin. 2001. "A Century of Writing Instruction in School and College English." In *A Short History of Writing Instruction: From Ancient Greece to Modern America,* edited by James J. Murphy, 247–89. Mahwah: Lawrence Erlbaum Associates.

Hopkins, Levi Thomas. 1921. *The Marking System of the College Entrance Examination Board*. Cambridge: The Graduate School of Education, Harvard University.

Huddleston, Edith M. 1952. "Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Testing Techniques." *ETS Research Bulletin Series* 1952 (2): 165–213. https://doi.org/10.1002/j.2333-8504.1952.tb00925.x.

Hunt, Kellogg W. 1977. "Early Blooming and Late Blooming Syntactic Structures." In *Evaluating Writing: Describing, Measuring, Judging*, edited by Charles Cooper and Lee Odell, 91–106. Urbana: National Council of Teachers of English.

Huot, Brian. 1990. "Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know." *College Composition and Communication* 41 (2): 201–13.

Huot, Brian, Peggy O'Neill, and Cindy Moore. 2010. "A Usable Past for Writing Assessment." *College English* 72 (5): 495–517.

Jacobs, Holly L., Stephen A. Zinkgraf, Deanna R. Wormuth, V. Faye Hartfiel, and Jane B. Hughey. 1981. *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Kane, Michael T. 1992. "An Argument-Based Approach to Validity." *Psychological Bulletin* 112 (3): 527–35. https://doi.org/10.1037/0033-2909.112.3.527.

—. 2001. "Current Concerns in Validity Theory." *Journal of Educational Measurement* 38 (4): 319–42. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x.

Lado, Robert. 1961. *Language Testing: The Construction and Use of Foreign Language Tests: A Teacher's Book*. New York: McGraw-Hill.

Loevinger, Jane. 1957. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports* 1957 (3); 635–94. https://doi.org/10.2466%2Fpr0.1957.3.3.635.

McIntyre, Philip N. 1993. "The Importance and Effectiveness of Moderation Training on the Reliability of Teachers' Assessments of ESL Writing Samples." MA Thesis, Department of Applied Linguistics, University of Melbourne.

McNamara, Tim F. 1996. *Measuring Second Language Performance.* Harlow: Addison Wesley Longman.

Messick, Samuel. 1975. "The Standard Problem. Meaning and Values in Measurement and Evaluation." *American Psychologist* 30 (10): 955–66.

—. 1989. "Validity." In *Educational Measurement*, edited by Robert Linn, 13–103. Washington, DC: American Council on Education and National Council on Measurement in Education.

Miller, David M., and Robert L. Linn. 2000. "Validation of Performance-Based Assessments." *Language Testing* 24 (4): 367–78. https://doi.org/10.1177%2F01466210022031813

Mullis, Ina V. S. 1980. *Using the Primary Trait System for Evaluating Writing (Report No. 10-W-51)*. Denver: National Assessment of Educational Progress, Education Commission of the States.

Nikolov, Marianne, ed. 2009. *Early Learning of Modern Foreign Languages: Processes and Outcomes.* Clevedon: Multilingual Matters.

Ruch, Giles Murrel. 1929. *The Objective or New-Type Examination: An Introduction to Educational Measurement*. Chicago: Scott, Foresman and Co.

Scott, W. 1917. "A Fourth Method of Checking Results in Vocational Selection." *Journal of Applied Psychology* 1: 61–66. https://doi.org/10.1037/h0073494.

Shaw, Stuart D., and Cyril J. Weir. 2007. *Examining Writing: Research and Practice in Assessing Second Language Writing.* Cambridge: Cambridge University Press.

Starch, Daniel, and Edward C. Elliott. 1912. "Reliability of the Grading of High-School Work in English." *School Review* 20 (7): 442–57.

Sheppard, Everett M. 1929. "The Effect of Quality of Penmanship on Grades." *Journal of Educational* Research 19 (2): 102–5.

Sigott, Günter. 1994. "Language Test Validity: An Overview and Appraisal." *AAA: Arbeiten aus Anglistik und Amerikanistik* 19 (2): 287–94.

—. 2004. *Towards Identifying the C-Test Construct.* Frankfurt am Main: Peter Lang.

Stalnaker, John M. 1936. "The Problem of the English Examination." *Educational Record* 17: 140–43.

Thorndike, Edward L. 1918. "The Nature, Purposes and General Methods of Measurements of Educational Products." In *The Measurement of Educational Products – National Society for the Study of Education Yearbook*, edited by Guy Monteose Whipple, 16–24. Chicago: National Society for the Study of Education.

Traxler, Arthur E., and Harold A. Anderson. 1935. "Reliability of an Essay Test in English." *School Review* 43 (7): 534–40.

Tsai, Constance Hui Ling. 2002. "Issues of Validity in the Assessment of Writing Performance." *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* 4 (2): 1–3.

Tyler, Ralph W. 1934. *Constructing Achievement Tests*. Columbus: Bureau of Educational Research, Ohio State University.

Weigle, Sara C. 1994. "Effects of Training on Raters of ESL Compositions." *Language Testing* 11 (2): 197–223. https://doi.org/10.1177%2F026553229401100206.

—. 2002: *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Weir, Cyril J. 2005. *Language Testing and Validation: An Evidence-Based Approach*. Houndgrave, UK: Palgrave.

Weir, Cyril J., Ivana Vidakovic, and Evelina D. Galaczi. 2013. *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012.* Cambridge: UCLES/Cambridge University Press.

White, Edward M. 2001. "The Opening of the Modern Era of Writing Assessment: A Narrative." *College English* 63 (3): 306–20. https://doi.org/10.2307/378995.

Wiseman, Stephen. 1949. "The Marking of English Compositions in Grammar School Selection." *British Journal of Educational Psychology* 19: 200–209. https://doi.org/10.1111/j.2044-8279.1949.tb01622.x.

—. 1956. "Symposium: The Use of Essays in Selection at 11+." *British Journal of Educational Psychology* 26: 172–79. https://doi.org/10.1111/j.2044-8279.1957.tb01390.x.

Yancey, Kathleen Blake. 1999. "Looking Back as We Look Forward: Historicizing Writing Assessment." *College Composition and Communication* 50 (3): 483–503. https://doi.org/10.2307/358862.

Zumbo, Bruno D. 2007. "Validity: Foundational Issues and Statistical Methodology." In *Handbook of Statistics 26: Psychometrics*, edited by C. R. Rao and S S. Sinharay, 45–79. Amsterdam: Elsevier Science.

—. 2009. "Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice." In *The Concept of Validity: Revisions, New Directions and Applications*, edited by Robert W. Lissitz, 65–82. Charlotte: Information Age Publishing.