Testing for professional qualification in conference interpreting

Andrew C. Dawrant and Chao Han

Introduction

Conference interpreting has been described as a “testing-intensive” profession (Setton & Dawrant 2016: 373) in which formal tests accompany practitioners “at least into the intermediate stages of [their] career” (Sawyer 2000: 32).

Different kinds of tests can be categorised according to their purpose: in training programmes, entrance exams aim to be predictive of readiness and suitability (and possibly even of ‘aptitude’) for conference interpreter training; in-course assessments are formative, aiming to generate feedback for learning and teaching; and end-of-course professional exams are summative, aiming to assess if graduation requirements have been met. Freelance accreditation tests and staff interpreter recruitment tests organised by conference interpretation services and certification tests offered by public credentialling bodies are professional qualification tests that aim to assess competence for professional practice.

In principle, if conference interpreter training programmes target job-readiness as their graduation standard—as recommended inter alia by AIIC (1999), the European Masters in Conference Interpreting consortium (EMCI 2012), and the chief interpreters of conference interpretation services (HINTS 2012)—then their end-of-course graduation exams are also de facto professional qualification tests. Indeed, in many markets, training programme-issued professional diplomas are the only test-based credential of competence available.

In this chapter, we will focus on professional qualification tests for conference interpreters, in all these categories, first, because of their high stakes—for candidates, schools/testing bodies, interpretation services, users of conference interpretation, and indeed the profession—and, second, because these consequential tests have been chronically under-researched and there is reason to believe that many may not meet basic standards of test quality.

The Standards for Educational and Psychological Testing (AERA, APA & NCME 2014) are widely recognised as the most authoritative statement of professional consensus on best practices and technical standards in behavioural measurement and deserve to be carefully consulted by all testers—including in conference interpreting. These will be referenced extensively throughout this chapter and referred to below as the Standards. Our objective in this
chapter is to examine existing conference interpreter qualification tests based on available published descriptions, to benchmark them against the Standards, attempt to identify potential gaps in their validity, and propose recommendations for improvement to bring testing in line with best practices.

For more general background on interpreting testing and assessment (ITA), readers are referred to Han (2021) for a state-of-the-art review; Pöchhacker and Liu (2014) and Russo, Chapter 23 in this volume for research on aptitude testing; Sawyer (2004) and Setton and Dawrant (2016) for discussion of testing and assessment in conference interpreter training programmes; Han (2016b) for a review of measurement issues in interpreting research; and for broader discussion of testing and assessment in different domains of T&I, Chen and Han (2021), Huertas-Barros et al. (2018), Tsagari and van Deemter (2013), and Angelelli and Jacobson (2009).

Finally, survey research on user perception or expectation of quality in conference interpreting is covered in Pradas Macías and Zwischenberger, Chapter 19 in this volume. This line of research provides upstream input for testing by helping testers better understand the multifaceted nature of ‘interpreting quality’, a construct that is related to, but not equivalent to, the crucial construct of ‘interpreting competence’ that underlies the design and development of professional qualification tests, which are the subject of the present chapter.

### Criterion-referenced testing: an overview

In the parlance of testing, conference interpreter qualification tests are performance tests designed to elicit samples of behaviour on representative tasks in the target domain for the purpose of distinguishing test-takers who are qualified for professional practice (‘masters’) from those who are not (‘non-masters’). Such professional qualification (also known as credentialling, certification, or licensure) tests should be criterion-referenced, meaning that test-takers’ performance is assessed against clearly defined performance standards that are explicitly linked to requirements for successful practice in the real-life performance domain. In criterion-referenced testing, these performance standards are “usually defined in a process that is external to the examination’s development” (Johnson et al. 2009: 257) and are operationalised by test developers in a cut score (or ‘pass mark’), which represents achievement of the criterion, or acceptable level of performance, on the test.

Criterion-referenced tests are documented in detailed test specifications which indicate critical dimensions to be measured (e.g., skills and knowledge, cognitive processes, context for performing the tasks) so that tasks selected for testing will systematically represent the critical dimensions, leading to a comprehensive coverage of the domain as well as consistent coverage across test forms.

>(AERA, APA & NCME 2014: 78)

Because performance tests rely on subjective scoring by human raters, a paramount concern is how to minimise arbitrariness and maximise consistency and quality in scoring. This is achieved through quality control processes that include clear, explicit, and detailed scoring guides; careful selection, training, and qualification of raters; requiring independent scoring by two or more raters, and applying statistical checks to establish inter-rater consistency (AERA, APA & NCME 2014: 33–47, 118). In credentialling tests, verifying “the consistency of decisions on whether to certify is of primary importance” (AERA, APA & NCME 2014: 182).
Finally, as perfect measurement accuracy is unattainable, even in a well-designed test there will always be a risk of classification errors resulting in false positives and false negatives. Different rules for mastery decisions can shift the balance of risk one way or the other. In professional credentialing tests, where the relative harm of false positive misclassification outweighs that of false negative misclassification, measures may be taken to minimise the risk of false positives (Cizek & Bunch 2007: 25–29).

As these professional qualification tests are high-stakes tests of conference interpreting competence that regulate candidates’ access to graduation, accreditation, certification, and employment, they should meet higher technical standards than low-stakes ones such as formative assessments (AERA, APA & NCME 2014: 33, 42). Sufficient evidence of their quality and usefulness should therefore be established through test validation.

Validity and test validation

The most fundamental consideration in developing and evaluating tests is validity, defined as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA & NCME 2014: 11).

In recent years, the concept of validity has evolved in two important ways. Traditional views of test quality were characterised by a dichotomy between reliability and validity, with the former referring to whether a test is able to measure a given construct consistently, and the latter to whether a test really measures what it purports to measure. Such dichotomisation can be counterproductive, when testing agencies sacrifice one in the pursuit of the other (see Weir 2005: 6–7). Thus, reliability has come to be subsumed within validity and seen as one category of validity evidence. At the same time, the traditional conceptualisation of distinct types of validity (e.g. content validity, construct validity, predictive validity) has given way to a unified validity theory that treats them as different kinds of validity evidence (Han & Slatyer 2016). This approach originated in a seminal chapter by the validity theorist Samuel Messick (1989), and is reflected in the latest edition of the Standards (AERA, APA & NCME 2014: 14).

Thus, validity is a unitary, yet multifaceted, concept, meaning that multiple lines of validity evidence need to be marshalled to support validity claims. Programmatic and consistent efforts are required to identify, collate, and produce relevant (counter)evidence to substantiate or refute validity claims and assumptions. This process is known as test validation. To the best of our knowledge, no professional qualification test for conference interpreters has ever been scrutinised in a systematic and rigorous validation study.

While contemporary frameworks for test validation are all generally compatible, we deem Weir’s (2005) socio-cognitive model particularly relevant for conference interpreter tests, for three reasons. First, its socio-cognitive focus is highly relatable to conference interpreting, a cognitively taxing activity situated in specific communication settings. Second, it embraces all aspects of the evaluative (scoring) dimension of testing (e.g. assessment criteria, scoring guides, rater characteristics, score reliability), which is critical in rater-mediated, criterion-based performance assessment. Third, Weir’s framework is accessible for non-specialist readers.

Weir’s (2005) framework focuses on validity evidence in five categories:

1. **cognitive validity**, referring to the degree to which the cognitive processes underlying real-life performance are elicited by test tasks;
2. **context validity**, which resides in the degree of closeness or authenticity between test conditions and the real-life performance domain;
3. *scoring validity*, which concerns the degree to which test scores are stable, consistent, and free from sources of measurement error;
4. *criterion-related validity*, which consists in the relationship between test scores and an external criterion of performance with known properties; and
5. *consequential validity*, meaning the extent to which the potential and actual social consequences of test score-based inferences and actions are supportive of intended testing purposes and consistent with social values (including *fairness*).

In the following review of existing qualification tests and subsequent recommendations for improvement, we will employ Weir’s framework and rely on the *Standards* (AERA, APA & NCME 2014).

**Review of available descriptions of existing conference interpreter qualification tests**

Although detailed literature on conference interpreter credentialing tests is scant, it is nonetheless possible to try to identify and analyse potential problems in tests organised by the United Nations, the European Union, certification bodies, and training programmes based on available publications: a recent PhD dissertation by a UN chief interpreter on the United Nations Language Competitive Examination (UNLCE) (Diur 2015); current public information on the EU’s freelance accreditation test (European Union 2018, 2020a, 2020b, n.d.); published descriptions of large-scale certification tests like the China Accreditation Test for Translators and Interpreters (CATTI) (Han & Slatyer 2016); and surveys and reports—albeit dated and non-comprehensive—on final testing practices in at least a subset of conference interpreting training programmes (Liu, Chang & Wu 2008; Sawyer 2004; Setton & Dawrant 2016: 379–383, 391–399).²

As shown below, our review of this available literature failed to identify any test that demonstrably meets basic standards for test quality in a professional qualification test and revealed potentially major problems across the board in terms of their *context validity* and *scoring validity*.

In terms of context validity, none of the reviewed tests appeared to have proper test specifications that ensure systematic and representative coverage of all critical aspects of the target domain (AERA, APA & NCME 2014: 76, 175, 178–179). Apparent gaps in coverage and authenticity include:

- insufficient and/or inconsistent efforts to test *routinely encountered discourse parameters* and ensure a *representative level of difficulty* in the test speeches:
  - Many surveyed training programmes had no standardised criteria for the selection of test speeches (Liu et al. 2008).
  - The EU’s test uses artificial speeches delivered semi-extemporaneously by interpreters in their native language, at comfortable speeds, containing “no deliberate traps or peculiar phrases” (EU 2018) despite evidence that fast speech rates, reading from written text, and non-native accents are common in EU meetings (Seeber 2017). For consecutive in particular, test speeches do “not seek to reflect real life speeches … [which] would be too difficult …, [and] too variable in difficulty, accessibility and quality” (EU examiner, pers. comm.).
- Conversely, the UNLCE fails to test communicative spoken discourse as routinely interpreted in discussion-based meetings at the UN, instead testing only the ‘oral
translation’ (Shermet 2012) of formal text-based statements read out for the record in the General Assembly and Security Council (Diur 2015).3 • Failure to test relevant modes: for example, simultaneous with text, checking against delivery, is not tested in the UNLCE, in the EU’s test, or in many of the surveyed training programmes, despite evidence that this mode is common in the United Nations (Diur 2015) and probably in general.

In terms of scoring validity (AERA, APA & NCME 2014: 33–47, 95–109, 118), potentially serious problems appeared to be endemic across the reviewed tests:

• In the UNLCE, a survey of raters found that “there are no clear criteria for the marking of the LCE and jury members (in the same booth and in different booths) have clearly different ways of assessing candidates” (Diur 2015: 226). No measures to check inter-rater consistency were reported.

• In many training programmes and in the EU test, checklists of assessment criteria may be provided to raters (e.g. EU n.d.), but without an explicit definition of the passing standard on each of them with differentiated descriptors and benchmarks.

• In many training programmes and in the EU test (EU 2018, 2020a; EU examiner, pers. comm.), raters assess performances in real time, including for simultaneous interpreting—even though live assessment generally allows for checking content fidelity on a “sampling basis” only (Gile 1999b: 54);

• In many training programmes and in the EU test, scoring is conducted through group deliberations among raters to reach consensus on pass/fail decisions (European Union 2020b; EU examiner, pers. comm.; Liu et al. 2008), instead of recording and comparing independent rater scores to establish inter-rater consistency, which is an essential practice whenever scoring relies on human judgment (AERA, APA & NCME 2014: 43–44, 118). Indeed, in the EU test, “individual jury members have an evaluation sheet but this is filled in after the deliberations” (EU examiner, pers. comm.).

Such departures from basic testing standards, especially in scoring, are arguably prima facie evidence of insufficient validity in a high-stakes qualification test. As such, the reviewed UN, EU, and training programme-based tests do not appear to meet relevant standards of technical quality. Consequently, these tests may not be functioning as intended.

A further word about collective ‘deliberations’ among raters, given that they appear to be both commonplace and taken for granted: these are evidently motivated by a desire to iron out inter-rater differences by seeking compromise and reaching consensus on a final pass/fail outcome for each candidate in the belief that this will result in better reliability and fairness in scoring. However, this belief is not borne out by social psychology research, which has found that this kind of group decision-making may be prone to undesired phenomena such as group-think, motivation losses, group polarisation, and conformity pressure (Olbrecht & Bornmann 2010). Moreover, empirical research has found that although raters within the same panel do tend to agree more after group discussion, different panels scoring the same inputs agree less (Pier et al. 2017). In other words, ‘deliberations’ may result in worse, not better, quality of scoring, despite the surface appearance of consensus.

In the case of large certification tests like CATTI, the lack of detailed information on the test’s design and of any substantive validity evidence is even more conspicuous (Han & Slatyer 2016). As Weir (2005: 219) puts it, there is “little excuse for examining bodies delivering high stakes tests that fail to meet … standards for the provision of validity evidence”.

262
In sum, based on our review of the available literature, it appears that not much has changed in conference interpreter testing since Hatim and Mason (1997: 198) wrote of the widespread “unsystematic, hit-and-miss methods of performance evaluation” and Campbell and Hale (2003: 211) drew attention to “the intuitive nature of test design and assessment criteria” that is prevalent in interpreter assessment.

Over the years, many educators and researchers have called for learning from the broader disciplines of educational measurement, psychological testing, and language assessment to better the practice of interpreter testing (Arjona 1984; Campbell & Hale 2003; Sawyer 2004; Clifford 2005; Angelelli 2009; Setton & Dawrant 2016; Han 2016a; Wu 2010). In the next section, we will suggest how this can be accomplished with practical recommendations for improving conference interpreter testing.

Recommendations for improvement in conference interpreter qualification tests

Our overarching recommendation is that conference interpreter qualification tests should be firmly grounded in established best practices in criterion-referenced testing as set out in the authoritative Standards and on research in language testing. Here we will offer concrete recommendations to accomplish this in each of the five dimensions of Weir’s model.

1 Cognitive validity

First, the three prevalent modes of conference interpreting—consecutive, simultaneous, and simultaneous with text, checking against delivery—appear to engage differently configured cognitive processes operating under different constraints (Gile 2009: 167–182). Each mode should therefore be tested separately (Recommendation 1.1).

Test developers may also consider testing SI with slides separately. In its basic form, this mode is cognitively similar to SI with text; but more demanding applications may impose different cognitive processing requirements, for example, where a speaker discusses numeric data in complex spreadsheets, as is common in business and budgeting meetings.

For certain job settings like the UN General Assembly and Security Council (and possibly the European Parliament), it may be necessary to test separately the kind of simultaneous interpreting that Shermet (2012) calls ‘oral translation’, which is performed on carefully crafted text-based speeches and requires that virtually every source-language word be accounted for in the target-language output with a high degree of formal correspondence. In cognitive terms, this can be seen as a rerouting, or reweighting, of the main processing route away from conceptual and intentional representations towards retrieval from the bilingual phrasebook of fixed equivalents, frames, and formulas (Setton 1999: 65; Setton & Dawrant 2016: 166–167), and deliberate suppression of the ‘condensation norm’ in interpreting that “not only condones but often encourages strategic macroprocessing” (Shlesinger 1999: 69).

Second, to ensure that the relevant cognitive processes are fully engaged, input speeches with overly familiar, practised, or predictable content should be avoided (Recommendation 1.2). Similarly, speeches should not be delivered by speakers whom candidates may be overly used to interpreting, such as their own instructors.

Third, speeches should be sufficiently contemporary, such that the matters discussed and any persons/events/timelines referenced are not grossly inconsistent with the candidates’ current frame of reference (Recommendation 1.3).
2 Context validity

The validation of professional qualification tests “depends mainly on content-related evidence” demonstrating that the test adequately covers the content domain of the relevant occupation (AERA, APA & NCME 2014: 175; 89, 178–179). For conference interpreter testing, to ensure adequate representation of the universe of conditions and challenges that practitioners will typically encounter, we recommend that testers develop detailed test specifications, based on empirical job analysis, that systematically sample all of the following (Recommendation 2.1):

(a) A sufficient range of authentic topics and speech genres, for example, international political discourse, specialised economic discourse, (semi-)technical topics requiring advance preparation, and possibly unexpected topics which would test range and adaptiveness.

(b) A sufficient range of authentic discourse parameters at real-world levels of difficulty as confirmed in empirical studies (e.g. Diur 2015; Seeber 2017; Varela García et al. 2019), including faster input speeds (>160 wpm), reading from written text not supplied to the interpreter, and non-native accents, as well as different registers, from very formal through colloquial and idiomatic.

(c) A sufficient number of ‘problem triggers’ (Gile 2009: 171) or ‘local hazards’ (Setton & Dawrant 2016: 422–423). These are SL speech elements that increase local processing capacity requirements and may be frequent sources of errors. These can be used by raters as ‘look-fors’ (Johnson et al. 2009: 185) in candidates’ interpretation performances that provide evidence about their level of interpreting ability. For conference interpreting, the most critical include (i) hard information like (lists of) numbers, unfamiliar proper nouns (including names of institutions and document titles), and technical terms; and (ii) culture-bound language, such as realia and idioms. In language pairs with asymmetrical word order, such as Chinese/Japanese-English, another important category is long asymmetrical structures which have been empirically demonstrated to be associated with a higher number of errors in SI (Wang 2014).

Test specifications must be sufficiently detailed to ensure a consistent level of difficulty across test forms, i.e. parallel forms reliability, which is a sine qua non for valid performance testing (AERA, APA & NCME 2014: 35, 176; Weir 2005: 250).

If input speeches are specially recorded, speakers should not be professional interpreters or others who may tend to “speak quite differently from how most people generally speak” (Weir 2005: 81) in terms of diction, speech rate, coherence, etc.; the ideal speakers are those who are “typical of the target language use situation” (Weir 2005: 81).

Another important parameter in the test specifications is test length, i.e. the total amount of interpreting performance to be sampled, in each language pair-direction, and in each mode. Longer tests with more tasks are more reliable (AERA, APA & NCME 2014: 38; Weir 2005: 68–69) and allow for better domain coverage. As a rule of thumb, we recommend eliciting at least 45 minutes of total performance in each pair-direction—broken up over different test sessions, with breaks—to allow for systematic sampling of different topics, genres, and input parameters, in different modes (Recommendation 2.2).

For context (and cognitive) validity, we strongly recommend that speeches be adequately contextualised (Recommendation 2.3): candidates should be given a brief as to who is speaking about what, on what occasion, and given any unfamiliar specialised terms, but without giving away the substantive content of the speech.
Finally, test tasks should be performed under authentic environmental conditions (Weir 2005: 56). Simultaneous should be tested in a proper booth in a conference-room setting, not a language lab, with light on-ear headphones and a clear view of the speaker on a large screen (Recommendation 2.4).

### 3 Scoring validity

Scoring validity appears to be the weakest link in current conference interpreter testing practice. To ensure scoring validity, in line with the Standards (AERA, APA & NCME 2014: 33–47, 79–80, 95–119), we generally recommend the adoption of validated scoring guides and procedures; precise definition of criterion performance; a robust process for selecting, training, and qualifying raters; independent scoring by each rater, without discussion; and statistical measures to check and document inter-rater consistency (Recommendation 3).

Given the paramount importance of scoring validity, the apparent poor practice across many tests, and the difficulty of getting it right, we will describe in detail key considerations in each of these areas based on the current state of the art.

#### Assessment criteria and scoring guides

Scoring guides should be based on explicit assessment criteria and sufficiently detailed to ensure their consistent application. For conference interpreting, there is widespread agreement on three main assessment criteria (Han 2018; Setton & Dawrant 2016: 433):

(a) **fidelity**, as realised in the completeness and accuracy of the interpretation and its faithfulness to the speaker’s communicative intent;

(b) **language quality**, including the subcomponents of natural and appropriate usage, correct grammar, command of register and terminology, and—especially important in conference interpreting—good pronunciation and clear diction;

(c) **delivery**, including fluency and communicative prosody\(^4\) (and in consecutive, time control).

Crucially, however, there exist no defined standards specifying exactly how good a performance must be on each of these criteria to merit a pass. No formal standard-setting process has ever been undertaken in the conference interpreting profession, or even, to our knowledge, by any testing body. For criterion-referenced conference interpreter testing, it is imperative that criterion performance be defined as precisely as possible on each of these major criteria (with rubrics and benchmarks, see below).

For all subcomponents of language quality (and fluency), validated scoring guides from language testing can be thoughtfully adapted, differently for A and B languages (Recommendation 3.1). In contrast, fidelity is a construct unique to interpreter testing, and is the primary determinant of interpreting quality. It is therefore of paramount importance to operationalise it in a scoring guide that is consistent with obtaining professional norms and sufficiently detailed for consistent scoring application (Recommendation 3.2). For any kind of interpreting test, the fidelity scale most fundamentally represents the theoretical basis upon which the test is founded.

#### Rubrics and benchmarks

Rubrics, also known as descriptor-based rating scales (i.e. a graduated set of descriptors that set out in sufficient detail critical features of performance at different levels), have emerged as a reliable, valid, and practical alternative to traditional scoring methods such as error...
analysis (see Angelelli 2009; Han 2018; Setton & Dawrant 2016). We strongly recommend that the assessment criteria of fidelity, language quality, and delivery be operationalised in explicit scoring rubrics (Recommendation 3.6) that clearly distinguish different performance levels with detailed descriptors. Furthermore, rubrics should be accompanied by exemplar recordings (‘benchmarks’) that amply illustrate performance at each level (Recommendation 3.4). Benchmarks help anchor raters to the rating scale during rater training and operational scoring (Johnson et al. 2009: 180–181).

**Holistic vs. analytic scoring**

Holistic rubrics are appropriate “when an overall judgment is desired and when the skills being assessed are complex and highly interrelated” (AERA, APA & NCME 2014: 79). Both these conditions apply in a conference interpreter credentialing test, whose purpose is to arrive at an overall judgment of job-readiness, and in which the main criteria of fidelity, language quality, and delivery are only partially distinct, interact in various ways, and may become inseparable in certain common error patterns (Setton & Dawrant 2016: 433–435). Indeed, the limited empirical research conducted to date indicates a fairly strong level of correlation: see Clifford (2005: 121) \( r = 0.690–0.746 \) and Liu (2013: 171) \( r = 0.668–0.743 \). In terms of scoring reliability and practicality, holistic rubrics are therefore probably more suitable for scoring interpreting performances in a credentialing test.

Analytic rubrics, which call for separate rating of different performance features, are preferable when assessment aims to provide diagnostic feedback to stakeholders, particularly test-takers, and different facets of a construct (such as the individual subcomponents of language quality) can be differentiated to a large extent. Analytic rubrics may therefore be generally more suitable for formative assessments.

For the common scenario where testers decide to adopt holistic scoring but also wish to generate useful feedback, we recommend holistic scoring for a summary score supplemented by analytic information for diagnostic purposes (cf. Johnson et al. 2009: 171) (Recommendation 3.5).

**Sourcing or creating and validating rubrics**

Many descriptor-based rating scales, both holistic and analytic, have been developed to score interpreting performance or ability for different purposes. (For a review of rating scales published before 2015, see Han 2018; and for an account of an ambitious recent project in China to develop descriptors and rating scales of interpreter competence, see Wang 2021). For our purposes, basic selection criteria are: (a) **fitness for purpose**, i.e. the rubric has been specifically developed for professional qualification testing of conference interpreters (not formative assessments) and reflects expert-judged performance requirements; (b) **appropriate coverage of all main assessment criteria** for conference interpreting; (c) **theoretically valid operationalisation of fidelity**, discriminating between different kinds of errors and omissions; and (d) **detailed, concrete descriptors at each level**, ideally refined through field-piloting, that correspond to typical discriminating features that raters actually observe in performances at different levels. Additionally, if the rating scale contains an embedded pass/fail standard between two band levels (which is not necessary, or even superior), the difference between the passing level and the immediately lower one should be as detailed and explicit as possible.

One published rubric that fits the above criteria is Setton and Dawrant’s (2016: 436–437) six-level holistic scale. This rubric has been locally piloted in small-scale, single-administration tests, but remains to be more widely field-piloted and validated on the basis of empirical testing data. We may therefore suggest that testers critically evaluate this rubric as a starting point. Of
course, this or any rubric should be accompanied by benchmark recordings in the relevant language combination, without which “even the best scoring rubrics are abstractions” (Johnson et al. 2009: 180).

We strongly recommend that current and future rubrics be subjected to extensive field testing and validation studies, including evaluation of their psychometric properties (Recommendation 3.6).

Score augmentation
In professional qualification tests (especially with pre-screening), a large number of scores may tend to be concentrated in the band levels immediately around the pass/fail borderline. In order to achieve higher discriminability and scoring precision, the scale’s granularity may be increased through score augmentation (Johnson et al. 2009: 173, 202–203), i.e. allowing raters to add a plus (alternatively, a plus or a minus) to a rating level (Recommendation 3.7). This allows for more fine-grained resolution of scoring levels, without creating new ones. Doing so is appropriate in cases where a given performance is best characterised by a combination of descriptors from two adjacent band levels. (But see below on ‘jagged profiles’.)

Setting the passing standard (cut score)
After a scoring system has been developed and fine-tuned in field piloting, the next key step is to define the passing standard with reference to the adopted rating scales. In criterion-referenced testing, this is operationalised in a cut score that represents criterion performance, separating masters from non-masters.

Setting the cut score essentially consists in deciding on what rubric-referenced level is minimally acceptable for professional qualification. This determination can be made by expert judgment alone with reference to the detailed descriptors in the rubric (see Setton & Dawrant 2016: 444–448, for an example), but ideally should be informed by empirical data such as the scores achieved during test piloting by test-takers with known levels of competence (AERA, APA & NCME 2014: 107–109). The cut score does not have to correspond to one band level, but could be, for example, 4.75 or 4.5 on a six-level rubric. We recommend that the cut score be defined through a formal standard-setting process (Cizek & Bunch 2007) with experts and stakeholders upstream of operational test implementation (Recommendation 3.8).

Recruiting, training, and qualifying raters
With a scoring system in place, the next step is recruiting, training, and qualifying raters. The ideal rater is one who “demonstrates the ability to attend to relevant criteria, work in a structured environment, follow directions, and recognize and monitor errors in his or her own judgment” (Johnson et al. 2009: 191). For conference interpreting, we recommend that raters be expert interpreters with these characteristics and all with (a) the appropriate working languages (including at least two native speakers of the target language) and (b) relevant domain expertise (Recommendation 3.9). Such raters’ in-depth understanding of the relevant constructs—especially of fidelity—can presumably contribute to the cognitive validity of score-based inferences.

Rater training is necessary because it is well known in testing that “similarity of background is not sufficient to predict who will be willing and able to apply a scoring system” (Johnson et al. 2009: 190). Unsurprisingly, empirical research has shown that professional interpreters assessing performances without training in a well-defined scoring system are inconsistent in their judgments (Gile 1999b; Wu 2010). To minimise unwarranted variability, we strongly recommend systematic rater training and qualification procedures, requiring raters to demonstrate
sufficient scoring consistency in practice sets prior to operational scoring (AERA, APA & NCME 2014: 25, 92; see also Han 2015, 2018; Liu 2013; Setton & Dawrant 2016: 438–440) (Recommendation 3.10).

**Number of raters**
The use of two raters is always a minimal practice (Angelelli 2009; Campbell & Hale 2003; Han 2018), with third-rater adjudication in cases of discrepancy. But in a small-scale, single-administration test, the reliability and quality of scoring will be much more convincingly demonstrated if all performances are independently scored by at least four raters (including two native speakers of the target language) and the necessary level of inter-rater consistency is achieved (see below) across this panel (Recommendation 3.11). After all, high agreement between just two raters may simply reflect that both were equally unfairly harsh or lenient. Also, empirical research has shown that score reliability estimates tended to be lower when raters assessed interpretation into their less dominant language (Han 2016a).

In a large test with many hundreds of candidates, like the UNLCE, it is usually infeasible to implement a fully-crossed rating design in which each rater scores each performance. We would therefore recommend a partially-connected rating design in which each rater scores a different set of performances, but is also linked or connected flexibly by a common set of performances (Recommendation 3.12). This data linkage is of critical importance, as it is a precondition for statistically analysing rater reliability and severity and for calibrating test candidates’ ability estimates (i.e. after controlling for rater severity). Rasch analysis, conducted via the Facets programme (Linacre 2021), is the most popular means to examine rating data from a partially-connected design (for an introduction, see Linacre 1989).

**Cognitive issues in scoring**
If simultaneous interpreting is one of the most cognitively demanding tasks, frequently saturating available processing capacity (Gile 1999a: 153), scoring SI performances may be even more cognitively demanding (Han 2015; Wu 2010), generally making it impossible during live scoring for raters to compare the SL and TL speeches on anything more than a “sampling basis” (Gile 1999b: 54). It is therefore necessary to take measures to reduce cognitive load for raters.

First, before raters start scoring test performances, we might suggest that they try out the tasks themselves under test conditions (Setton & Dawrant 2016: 462; Wu 2010: 110). This serves the dual purpose of familiarising raters with the input speeches generally and helping them better appreciate their global difficulty and local hazards.

Second, live scoring should be avoided wherever possible. Raters need to attend to both the source- and target-language speeches while assessing the output carefully against descriptors of different criteria at different performance levels, and noting salient characteristics of the performance (Gile 1999b; Han 2015; Wu 2010). To prevent cognitive overload and ensure scoring accuracy, we strongly recommend individual *post hoc* scoring from recordings, pausing after a passage and re-listening as necessary (Recommendation 3.13). Scoring from anonymous recordings is also a best practice to counter potential bias associated with irrelevant test-taker characteristics such as age, ethnicity, and attractiveness.\(^5\)

Third, we strongly recommend that verbatim transcripts of all source-language speeches be provided to raters (Recommendation 3.14). Wu’s empirical study of rater behaviour when scoring simultaneous interpretations found that having the input speech script at hand helped raters avoid falling back on impressionistic judgement (Wu 2010: 246–248). Similarly, Han (2016a) attributed higher score generalisability on the information completeness criterion to the availability of the source language script.
Finally, another potential source of cognitive interference is *order effect* (or serial position effect) (Han 2018; Wu 2010). To mitigate this, we recommend that interpreting performances be presented to raters in a different, or randomised, sequence, so that each rater assessing the same batch of recordings scores them in a different order (*Recommendation 3.15*).

**Jagged profiles**

A major potential challenge to valid scoring—whether on holistic or analytic rubrics—is presented by uneven, or ‘jagged’, performance profiles, where candidates present at markedly different levels (i.e. non-adjacent band levels) on different assessment criteria. This situation needs to be covered in rater training, with an explicit policy. Weir (2005: 181) points out that in high-stakes tests, the “simple solution is that [such candidates] get the level of their lowest performance” on any of the criteria, and this is indeed our recommendation (*Recommendation 3.16*). Thus, a candidate with poor diction, grammar, and/or fluency would receive an overall score corresponding to that band level, regardless of better performance even on fidelity.

**Granularity of scoring units**

A final important consideration in scoring practice is the granularity of scoring units, especially in relation to fidelity. Scoring fidelity at the whole-task level, i.e. assigning a single score to a candidate’s performance on an entire speech, may pose a serious risk to scoring validity, for the simple reason that performance levels can fluctuate, sometimes widely, throughout the task.

In recognition of this problem, several authors have suggested that input speeches be divided into cohesive passages, roughly corresponding to paragraphs, for passage-by-passage scoring (Liu 2013; Setton & Dawrant 2016: 432; Tiselius 2009).

While no empirical studies have yet been conducted in conference interpreting to compare the reliability of whole-task vs. part-task (passage-by-passage) scoring, recent empirical research (Khabbazbashi & Galaczi 2020) in the assessment of speaking performance by experienced raters found that choice of scoring model had a very significant impact on candidates’ CEFR® classifications: 30.1 per cent of performances were assigned a different CEFR classification (at an upper or lower adjacent level) on the basis of whole-task vs. part-task scoring by the same raters. Whole-task scoring resulted in higher overall scores and CEFR levels, entailing an elevated risk of false positives. Part-task scoring demonstrated higher reliability, superior measurement qualities in separating candidates into more ability strata, and was also confirmed in rater feedback to be “feasible in operational conditions” (Khabbazbashi & Galaczi 2020: 12).

We therefore recommend that test developers consider subdividing input speech transcripts into cohesive segments for passage-by-passage scoring. (*Recommendation 3.17*).

If passage-level scoring has been adopted, then these subscores will need to be aggregated (e.g. summed, averaged, or weighted) into an overall score from each rater for that task. In language testing, score aggregation is commonplace and considered “not a very complex process”, since “the use of the same scale levels throughout the process gives a conceptual basis for combining the scores” (Luoma 2004: 173). As rubric-referenced rating scales are not interval but ordinal measures, on strict interpretations of theory, subscores should not be combined by summing or averaging, as has been noted by one interpreting studies author (Sawyer 2004: 105), but instead should be analysed non-parametrically. However, empirical and simulation research has demonstrated that ordinal data can be reasonably described and analysed by parametric statistics, enabling more powerful and useful analysis in applied settings than non-parametric methods (Carifio & Perla 2008), and this approach has been practised in psychological and educational testing for decades (see DeVellis 2012). Which approach to take in aggregating scores can be decided during test piloting with the assistance of a testing specialist.
Essential statistics
Statistical analysis of scoring consistency is an indispensable step in criterion-referenced performance testing (AERA, APA & NCME 2014: 44), because “we need to be able to demonstrate that scores … are reliable, and that the ways in which we interpret and use … [these] scores are valid” (Bachmann & Kunnan 2004: 3). Familiarity with essential statistics is therefore an integral part of testing literacy.

At the minimum, two key statistics should be calculated and reported for any criterion-referenced performance test: the Standard Error of Measurement (SEM), and the correlation coefficient phi (φ) for the mastery judgments of different raters (Recommendation 3.18).

First, since perfect precision of measurement is unattainable, any test score should “always be viewed as a point within a range—not a single, absolute point” (Shrock & Coscarelli 2007: 283). SEM indicates what that range is, and how accurate any test score is likely to be, by defining a confidence interval around the score (AERA, APA & NCME 2014: 37, 39–40). SEM is a crucial measure in CRT for two reasons: first, as a “cautionary flag” if the confidence band around a score is too broad (Luoma 2004: 183), indicating low reliability/precision (AERA, APA & NCME 2014: 34); second, for deriving an operational cut score for each test administration by adjusting the theoretical cut score by a predetermined multiple of SEM.

The true score of a candidate whose observed score is \( x \) is contained within the interval \( x \pm 2 \) SEM at a 95 per cent confidence level. Thus, in a test with a theoretical pass score of \( y \), only by applying an operational cut score of \( y + 2 \) SEM (i.e. the judged pass score plus 2 SEM) can false positives be ruled out with high certainty.

A decision not to adjust the theoretical cut score by some multiple of SEM is a decision to be equally accepting of false positive and false negative classification errors (Cizek & Bunch 2007: 29). In a professional qualification test where the consequences of false positives are deemed more serious than those of false negatives, we recommend that test developers carefully consider, in consultation with stakeholders and testing specialists, adopting an explicit policy of adding a pre-determined multiple of SEM on top of the judged cut score to derive the operational cut score for each test administration (cf. Cizek & Bunch 2007: 299–306) (Recommendation 3.19). In practice, for a school-based diploma exam, a lower multiple of SEM might be added than in a UN or EU staff interpreter recruitment test, given the different stakes.

For reasons of cognitive validity discussed above, we recommend that this operational cut score be applied conjunctively to candidates’ scores in each separate mode (i.e. requiring a separate pass in consecutive, simultaneous, and simultaneous with text) (Recommendation 3.20).

Second, in a professional qualification test, “the consistency of decisions on whether to certify is of primary importance” (AERA, APA & NCME 2014: 182). To demonstrate sufficient consistency of their mastery decisions, the independent judgments of each rater must be compared statistically. In practical terms, this means first converting the quantitative score returned by each rater into a master/non-master decision from that rater by comparing it to the operational cut score. The consistency of these mastery decisions across the panel of raters can then be checked using the phi (φ) coefficient (Shrock & Coscarelli 2007: 335–340). In any test, a phi coefficient of less than 0.75 is unacceptable. A high-stakes test in an applied setting should target a phi coefficient of 0.90 or higher (Johnson et al. 2009: 292), and ideally 0.95 (Shrock & Coscarelli 2007: 340) (Recommendation 3.21).

For large-scale and potentially well-resourced testing programmes, like the UNLCE, a more sophisticated approach to examining rater behaviour and, indeed, an entire rater-mediated performance assessment system would be many-facet Rasch measurement (MFRM) analysis (Linacre 1989), implemented through the Facets programme. In conference interpreter testing, the test design consists of four assessment facets: test-takers, tasks, raters, and assessment...
criteria. MFRM analysis is capable of generating measures for each of these four different assessment facets on a single interval-based scale (in logits), including ability estimates for each test-taker, severity estimates for each rater, and difficulty estimates for each task and each assessment criterion, thus creating a common frame of reference for interpreting and comparing assessment results. (For examples, see Han 2015, 2017.)

4 Criterion-related validity

Obtaining criterion-related validity evidence is extremely challenging for conference interpreter testing as there are very few criterion measures available, and their own validity is almost certainly not well established. In some limited cases, it may be possible to correlate test scores with other criterion-referenced test measures. For example, scores on EMCI programme’s graduation exams, which target conference interpreting in EU institutions (and have EU examiners on the jury), could potentially be correlated with those of the same graduates on EU accreditation tests. Similarly, investigation of predictive validity may be possible only in a large interpretation service such as those of the UN and the EU, where scores on accreditation/recruitment tests could be correlated against future in-house performance assessments.

5 Consequential validity

Adoption of criterion-referenced professional qualification testing can be expected to bring about positive impacts for all stakeholders in conference interpreting by ensuring appropriate and justified performance standards for professional practice. Additional considerations include:

First, given that ethics and professionalism are bedrocks of professional interpreting, we recommend that they also be assessed (for example, in multiple-choice test, open-ended questions, interview, and/or simulation formats) (Recommendation 5.1).

Second, feedback should be provided to test-takers, especially those who narrowly failed, so they can improve and re-test (Recommendation 5.2). Aggregated feedback could also be provided to training programmes for benchmarking purposes.

Third, it may be worthwhile considering issuing different accreditations (diplomas, certificates) signifying different levels of competence. For example, two different cut scores could be set: candidates who achieve the lower cut score would be passed at the ‘journeyman’ level (Sawyer 2004) (with the opportunity to retest and upgrade), and those who achieve the higher cut score would be passed as fully-qualified ‘expert’ conference interpreters. Again, SEM would have to be considered in setting these cut scores, to avoid (significant) overlap between the two bands.

Conclusion

Our review of the available literature on conference interpreter qualification tests leads us to conclude that criterion-referenced testing has not been widely applied in our profession. We call for test developers and testing bodies to design, develop, administer, and validate tests in accordance with basic testing standards, and to publish detailed information on their tests in order to generate knowledge for testers and for the profession. Virtually every aspect of testing practice discussed above can benefit considerably from empirical research and we sincerely hope that this chapter will prompt more research on high-stakes testing of conference interpreting competence.
Notes


2 At the time of writing, Australia’s National Accreditation Authority for Translators and Interpreters (NAATI) was in the process of developing a new test for the certification of conference interpreters. Detailed information was not yet available, and this test was therefore not included in our review.

3 And, in another departure from authenticity, candidates in the UNLCE are not given the text, even though UN interpreters typically would have a copy of the speech in the booth (Diur 2015: 178).

4 The importance of prosody (fluency, pauses, intonation) in user perception of quality has been seen in empirical research: Collados Aís (2002) reported that monotonous intonation resulted in poorer evaluation by monolingual listeners, including of (perceived) fidelity. See also Collados Aís et al. (2007); Rennert (2010).

5 In certain markets with small pools of candidates who may be personally known to the raters, it may be advisable also to consider employing voice-altering software to disguise the identities of test-takers.

6 The Common European Framework of Reference for Languages.

References


Han, Chao 2016a. Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. Language Assessment Quarterly 13 (3), 186–201.


Liu, Minhua, Chang, Chia-Chien & Wu, Shao-Chuan 2008. Interpretation evaluation practices: Comparison of eleven schools in Taiwan, China, Britain, and the USA. Compilation and Translation Review 1 (1), 1–42.


