GERGELY A. DÁVID

School of English and American Studies, ELTE
david.gergely@btk.elte.hu
0000-0003-2280-5451

Gergely A. Dávid: Three Mediums of Test-taking and Performance in the Measurement of Foreign Language Competence at Level B2
Alkalmazott Nyelvtudomány, XXIII. évfolyam, 2023/1. szám, 18–35.
doi:http://dx.doi.org/10.18460/ANY.2023.1.002

Three Mediums of Test-taking and Performance in the Measurement of Foreign Language Competence at Level B2

The construct of foreign language competence has become broader in the past decades, progressively incorporating several performance factors. A more comprehensive, multi-faceted construct must also be matched by equivalent multi-faceted measurement, which few programmes currently can do. However, technological developments allow the logging, among other potential factors, of what test-taking medium the test was taken in, computer-based at a testing centre or "home-based" online. According to Many-facet Rasch analysis, taking the test online was significantly easier than the same test at a testing centre, giving rise to the medium of test-taking as a performance factor. The emerging factor poses a problem for many test providers in that they should not administer tests both online and offline, as it would be unfair unless they are prepared to upgrade their analytical capabilities to Many-facet Rasch or equivalent.

Keywords: Communicative competence, facets of performance, Many-facet Rasch Measurement

Introduction

The influence of the pandemic is just becoming visible, as is shown by frequent conference calls and presentations. Partly as a result of the pandemic, completely new tests and examinations have been launched, and „old" examinations that had been in operation for years have added internet-based, online capabilities. In Hungary, at least, an immediate outcome of the pandemic seems to have been a boost in the development of computer-based (CBT) and internet-based, "online" testing (IBT). This paper will discuss the possible effect of introducing online or internet-based capability in the foreign language testing field. The introduction of online capabilities gave rise to the research focus for this article, permitting an investigation of three different test-taking mediums in a single examination suite.

iTOLC as the context of research

The International Test of Language Competence (iTOLC) is a fully computerized foreign language proficiency examination. It was the first to appear on the Hungarian scene in 2019, although it should be conceded that another foreign language examination introduced a mixed CBT-PBT design earlier. Their exam comprised CBT reading and listening components, while the writing and speaking "papers" were still the traditional, paper-based face-to-face medium (PBT).

Thus, iTOLC is a relatively new exam administered in English and German. The test-takers can take this examination at four CEFR-based levels (A2, B1, B2, and C1) (Council of Europe, 2001). It is also a communicative, skills-based exam with four "papers" in reading, writing, listening, and speaking. iTOLC recognizes its roots in Communicative Language Testing (chiefly Morrow, 1979). It is not a system-based test of grammar, lexis, etc. (Baker, 1989). In this paper, only one level, the more popular B2, will be discussed.

iTOLC was initially conceived as a CBT exam to be taken at local testing centres, with the data collected on a central server for analysis, calibration, and score reporting. Following the COVID lockdown in March 2020, iTOLC obtained a temporary licence (accreditation) for online examining from the ministry agency, *Accreditation Centre for Foreign Language Examinations* [Nyelvvizsgáztatási Akkreditációs Központ]. This paper will refer to the modified exam version for online use as IBT.

The temporary licence expired at the end of August 2020. Following the submission of appropriate documentation, iTOLC was granted an extended licence for online examining at the beginning of November 2020, which meant that CBT and IBT operations were officially acknowledged from that point onwards. The two operations ran largely parallel, increasing the comparability of these two mediums. Between May and November 2020, iTOLC still had its original licence for the CBT operation. Except for September-October 2020, iTOLC simultaneously offered both the CBT and IBT mediums (modes of operation) for the test. At any particular exam date, when both mediums are offered, the test version is the same for both mediums.

By the end of 2021, there were cumulatively more test-takers for the IBT medium of the test in the item bank than those taking the CBT. It should be added that, during the test development phase, paper-based versions of the test material also participated in the pretests, which will be referred to as PBT below. In this way, this research comprises three mediums of test taking, with a focus on the CBT-IBT comparison as the chief interest. By the end of 2021, data from CBT and IBT accounted for nearly all the data, while PBT data constituted a small portion by comparison.

Thus, while iTOLC provided the context and background to this research, this paper focuses on three different test-taking mediums: CBT, IBT, and PBT. Since the mediums affect test-takers' foreign language performance, this paper deals with aspects of foreign language performance rather than the competence that performance entails. Like all other foreign language tests, iTOLC, as an accredited examination, seeks to measure test-takers' foreign language competence, to be inferred from their test performance.

The research questions

In this paper, the following questions will be answered.

- Is the medium of test-taking a method-related facet of performance, or is it related to the test-taker?
- Is there statistical evidence that the medium of test-taking is a performance factor?
- How do the three test-taking mediums in this study compare in terms of difficulty?

A review of the literature

The distinction between competence and performance is customarily attributed to Chomsky (1965), but the distinction, or a similar distinction, goes back to earlier times (de Saussure, 1997). The concept of competence in this time-honoured tradition was limited to knowledge of the language systems, leaving everything else to the performance domain. However, in the writings of several experts (Hymes, 1972; Chomsky, 1980; Canale & Swain, 1980; Canale, 1983a,b; Bachman, 1990; Bachman & Palmer, 1996; 2010) later on, the notion of competence has gone through a marked broadening, to include several non-language specific aspects, or factors, of performance. Endorsing a broad conception of language competence enriched by performance factors still leaves actual manifestations or observed instances of competence, alternatively actual use, as performance components. The issue of test-taking mediums will be discussed below in light of the development of competence and performance.

The development of foreign language competence theories

As has been indicated above, the question of whether and how different mediums to the same examination affect test results is, in the first place, a question of performance and not of competence. However, in apparent contradiction, it is test-takers' foreign language competence, i.e., communicative competence that stakeholders are always interested in, irrespective of the context or situation. (Bachman & Palmer, 1996: 18-19). In addition, it is competence that measurement specialists should be interested in when they use sophisticated measurement technology to tease out the test-taker's competence behind all that is observable in test performance.

A further twist to the logic is that foreign language competence itself is not observable without using measurement instruments and conditions. The mainstream thinking in language testing, referred to as the epistemological approach, posits that foreign language competence is an internal trait, not directly observable. It is to be inferred on the basis of performance in observed responses obtained with the measurement instruments. Thus, to get a proper focus on foreign language competence, researchers and practitioners must deal with the complexities of performance. Therefore, it is not surprising that complexities of

competence and performance resulted in considerable confusion and inconsistency in the use of relevant terminology, as McNamara (1996: 51, 55, 57, 68, 90) observed and pointed to the need for clarification.

Following successive theories of competence, McNamara (1996) identifies Hymes (1972) as the originator of a broadened sense of foreign language competence. This broader conceptualisation includes the language user's *ability for use* and which leads the reader to Hymes' communicative (foreign language) competence (McNamara, 1996).

McNamara defines Hymes' (1972) model of performance ability for use as a "broadly psychological model of performance" (McNamara, 1996: 55). Hymes' model includes a range of "cognitive and non-cognitive factors, none of them exclusive to language performance" and motivation (1996: 56). Performance is also formulated inclusively as "language-relevant but not language-exclusive factors," elsewhere (1996: 59). In addition to crucially pointing to the non-exclusive nature of these factors, McNamara also lists examples, such as Goffman, who names courage, composure, presence of mind, emotional states, personality factors, among others, as belonging to performance (1967: 56).

By contrast, Canale and Swain thought performance could not be modelled because they felt it was too complex (1980: 6); therefore, their model of grammatical, sociolinguistic, and strategic competences is the narrow version of the concept of communicative competence. Later, adding discourse competence to their earlier theory of competences, Canale (1983a,b) returned to the broader version of communicative competence, in which strategic and discourse competences were the clearest performance elements.

Bachman (1990) and Bachman and Palmer (1996) also introduced performance elements into their concept of communicative language ability. A central construct to Bachman (1990) is a broadly conceived strategic competence that activates the language user's language competence and knowledge of the world and engages the context of the situation through psychophysiological mechanisms. In Bachman and Palmer (1996), we find a further elaborated version of communicative language ability with affective factors and personal characteristics being the main additions. Modifications of lesser importance, in comparison with Bachman (1990), are hardly more than the relabelling of an earlier concept, such as *knowledge of the world* in Bachman (1990), which became *topical knowledge* in Bachman and Palmer (1996). McNamara (1996) criticised Bachman and Palmer (1996) for their limited exploration of performance and summed up his disappointment using the metaphor "having confidently lifted the lid on Pandora's Box, they shut it again" (1996: 74), meaning that Bachman and Palmer (1996) failed to go far enough.

It is Bachman and Palmer (2010) who present the most well-developed and complete version of communicative competence, i.e., their (communicative) language ability (CLA) construct (33-58): In addition to language knowledge,

topical knowledge and personal attributes contribute towards the speaker's language ability and feed into their broadly conceived strategic competence. Bachman and Palmer's strategic competence interacts with other cognitive strategies where the language user's strategic competence ultimately prompts the question of the particular cognitive strategy or strategies implemented. Both strategic competence and cognitive strategies operate through affective schemata, including, most notably, motivation.

The strategic competence of Bachman and Palmer (2010) is much broader than Canale and Swain's strategic competence, which comprised only of compensatory mechanisms (1980: 30) or coping strategies (1980: 31), following Stern (1978). Bachman and Palmer's strategic competence includes goal-setting, appraisal, and planning (2010: 49).

From performance factors towards a theory of performance

Theories of foreign language competence were repeatedly proposed, most of them years ago. As a result, some (Canale & Swain, 1980; Bachman, 1990) appear to be dated now. It is perhaps fair to say that while the debate centred on elements of competence and the breadth of the communicative foreign language construct, theories of *actual* performance have not evolved with the same speed, or to the same extent, and as coherently and comprehensively. As performance factors have been progressively incorporated into broader theories of foreign language competence, it might be asked what factors are left to complement a theory of actual competence.

Writers have generally been very cautious about putting forward their ideas about actual performance, most notably Canale and Swain, who did not think it could be modelled (1980: 6). Since then, the successive constructs of language competence have not only become "enhanced" constructs of communicative language competence (ability) but have also become increasingly complex. In this way, teasing out communicative language competence from many other variables becomes a genuinely daunting endeavour. No wonder that McNamara (1996: 48-90) and Widdowson (2001: 13) both use the same metaphor *Pandora's Box*, to describe the outcome of the activities of the very influential communicative movement at the turn of the 1970s and 1980s (Brumfit and Johnson, 1979; Morrow and Johnson, 1981; Littlewood, 1981). McNamara (1996) and Widdowson (2001) recognized that the theorists of Communicative Language Testing (CLT), having brought a multitude of performance factors into the test, embarked on what McNamara and Widdowson see as a challenging, almost impossible mission. Thus, a theory of actual performance, whatever components it might comprise, will probably also be highly complex, not in the least because of the complex expectations of communication by the practitioners and theorists who espoused Communicative Language Teaching and Testing.

Categories of performance factors

A theory of actual performance can probably only be formulated regarding contingencies because testing contexts are incredibly varied. The mediums of test-taking, CBT or IBT, and the use or non-use of dictionaries during the test may be examples of such contingencies. It is unlikely ever to be a definitive list, I believe, not in the least, because they are dependent on options. In one assessment context, the test-taker might choose online assessment, while in another context, they will choose to be examined at a testing centre. Similarly, in one particular testing context, the test provider may allow the use of dictionaries while taking the test, while in another context, dictionaries may not be used.

The medium of test-taking might belong to one of two categories of performance. First, it may belong to the category of test methods. Test method performance factors are also the most readily identifiable and observable factors. Method factors are crucially important because, without methods or instruments, no response data may be collected from the test-takers. In addition, test materials, or items and tasks, belong to a method as well, and it is plain to see that without them, no evidence from the test-taker would be forthcoming. Alternatively, the medium of test-taking may also belong to the category of test-taker factors because it is up to the test-taker's decision between the CBT and IBT modes of operation.

Test method-related factors of performance

The testing methods are the results of the test provider's informed decisions, which the test provider is ultimately not interested in. However, they constitute performance factors because they affect the scores even if they do not belong to language competence.

In addition to test items or tasks being an explicit component of method, the rating scales used by the examiners should also be seen as part of method. Item formats, or, for example, the categories of source of text (originally audio, originally audio-visual, written, etc.) in a listening test should also be seen as part of method. In addition, examiners, or raters, may also be categorised as *methods or facets of method,* even if the term is intuitively less appealing than *characteristics* in Bachman and Palmer (1996; 2010).

In this discussion of method facets, or factors, rating scales and raters may be selected as good examples to show a growing acceptance of CLT as represented in Bachman and Palmer's writing. Bachman (1990) presented an elaborate framework for test method facets (1990: 111-159), further elaborated in Bachman and Palmer (1996) as a framework for language task characteristics (47-57), which labelling was maintained in Bachman and Palmer (2010). In very general terms, it may be stated that rating scales and raters, with each reworking of the method characteristics CLT, are represented with more detail and explicit formulations in Bachman and Palmer's writing.

Test-taker-related factors of performance

The performance factors related to test-takers are more of a "grey zone" because our understanding of them seems less well-developed. O'Loughlin (2002, p. 189), for example, reported contradictory results from the research literature (of speaking tests) and found no gender effect in their own study. In addition to gender, there is evidence in the literature that a range of test-taker-related variables is investigated, such as test-taker attitudes, motivation, or test anxiety. Such research shows that test-takers' constituency is very far from being monolithic and varies from place and time.

An early monographic treatment of test-taker characteristics is Kunnan (1995), who investigated the fit of four proposed models to the data collected, reflective of a typical American starting point, the Indo-European or Non-Indo-European background of the test-takers. None of the models fit perfectly. Oliveri et al. (2015) is another interesting example of a project in which they try to take account of the different racial, cultural, and social backgrounds of the citizens of the countries their tests, developed in the US, are exported to where such tests may not function as expected, may not produce equally valid results due to differing test-taker characteristics. A study similar to this author is Kenyon and Malabonga (2001), in which the authors compared attitudes to two technology-mediated oral tests: a simulated oral proficiency interview (SOPI) and the then new computerized oral proficiency interview (COPI), the latter being an adaptive test. Primarily lower level test-takers favoured the COPI.

In Hungary, according to a recent survey (Kiszely, 2022), it appears that no study investigated the differential action of the mediums of test-taking, but the last reference entry dates from 2020, the year of the pandemic. However, even more recently, Babos et al. (2022) compared the CBT and IBT mediums with the traditional PBT medium in a nationwide survey of accredited exam results. The authors wanted to know (1) whether the new CBT/IBT mediums, labelled together as "computerised," or the PBT medium was more difficult in the study period (2018-2021). They also wanted to know (2) whether the overall increase in the success rate was attributable to the appearance of the computerised (CBT/IBT) mediums alone. From the point of view of this study, it is unfortunate that the CBT/IBT frequencies are not reported separately in Babos et al. (2022). It should be added that *Akkreditációs Kézikönyv 2023* (2023) does not make a difference between IBT and CBT, as is understood in this article. However, the data Babos et al. present have at least shown that the results from computerised and PBT exams are not consistently higher or lower in the years between 2019 and 2021. (Results for computerised examinations could not be reported from 2018 since no such examinations were functioning yet.) Answering question 2, Babos et al. (2022) concluded that the appearance of computerised exams could not have been the single cause of the observed higher success rate.

It is also the case that some possible performance factors related to the test-taker result from options present in some assessment contexts but not in others. The variability of the assessment context makes research difficult because a performance variable may be relevant in one assessment context while it may not be so in another. If test-takers can use a dictionary, for example, there is a choice involved, and the dictionary factor ought to be acknowledged, but if dictionaries and choice are not allowed, the factor will not play a role. Ultimately, it should be important that the test-taker's choice and decision are needed for the mediums of test-taking.

Multi-faceted competence, performance, and Many-facet Rasch Measurement

As has been agreed upon within the language testing community since the falsification of Oller's Unitary Competence Hypothesis (UCH; 1976), foreign language competence is understood to be multi-faceted (or multi-factorial). That there is agreement may be inferred from the fact that no unitary concept, or construct, gained prominence in the past decades after Oller (1976). Multi-faceted models have become prevalent instead. From the point of view of our research, it does not seem important what internal parts, components or dimensions, or facets different models stipulate. Canale and Swain's model (1980), Canale's revision (1983a,b), or the successive versions of Bachman's model (1990), later in Bachman and Palmer (1996, 2010), all identify partly different components of foreign language competence.

Similarly, once foreign language competence is multi-faceted, foreign language performance must also be multi-faceted, as is shown in the discussion above. However, no comprehensive model of performance has emerged to date. As is discussed above, the very elements of performance that have become part of the communicative foreign language competence construct are but an indication of the multi-faceted nature of performance.

Many-facet Rasch modelling

From the above, it logically follows that the measurement of foreign language competence should also be multi-faceted to match the nature of language competence, with the addition of performance elements. Two (or three) software programmes can statistically operationalize the multi-faceted view of language proficiency. Both are probabilistic, based on modern test theory. One of these is appropriately called *Facets* (Linacre, 2014a) because language competence and the various performance factors are like facets of the complete performance generated in the measurement process. *Facets* is an extension of Rasch methodology, itself a branch of Item Response Theory. According to https://www.rasch.org/software.htm, the other software to operationalize a multi-faceted view of foreign language proficiency is *Conques*t (Adams et al., 2020).

However, the information therein is at variance with what the author of this paper knows from practice because *Conquest* is labelled as "multidimensional," whereas *Facets* is not. Other authors would also be wrong if Facets were not multidimensional (Bond and Fox, 2001; Eckes, 2015). A more recent addition to Many-facet Rasch Methodology may be the programme package *R*, which is labelled as having "some Rasch functionality" at rasch.org. In contrast, Wind and Hua (2022) explicitly address the Many-facet Rasch capabilities of the R package.

Apart from these two (or three), a host of other software may be used to estimate test-takers' foreign language competence. Still, they generally operationalize only one factor, that of items (tasks), in addition to the test-takers: they are dichotomous models. This aspect is clearly a limitation to most measurement contexts, especially those that follow communicative ideals. Thus, it appears that, although most software operationalises a dichotomous model, a growing number can operationalise more than a single factor (facet) in addition to items/tasks, catering to the demands of communicative language teaching and testing.

The factors that *Facets* can typically operationalise, in addition to items (tasks), are raters and rating scales, thus making *Facets* a suitable analytical tool to process data from productive skills tests in the first place. In addition to the above, *Facets* can process data from additional facets also, such as the formats of different item/task types, dictionary use (or non-use), and last but not least, the facet of the medium of test-taking, the focus of this paper. All the above make this branch of Rasch measurement into what is called Many-facet Rasch Measurement (MFRM). (This extension of Rasch methodology is not to be confused with classifying models into one, two, and three-parameter models in IRT.)

As with basic Rasch methodology, most facets in MFRM constitute "hurdles" to the test-taker, for example, the items/ tasks facet or that of the raters, where higher scores indicate lower difficulty and lower severity (more leniency). As a result, compensation in the test-taker ability will occur. In compensation for performance factors that constitute "hurdles," the measure of the test-takers' ability is modified, raised, or lowered. Some performance factors might act otherwise, contributing to or "boosting" ability, again raising or lowering abilities.

As has been said, the original Rasch model (Rasch, 1960/1980) comprised only two facets: items and persons, which was adequate at the time since it was the structuralist era, strongly associated with behaviourism in psychology and discrete-point techniques and dichotomous data in the field of language testing. With the advent of communicative language teaching, there was a growing need to recognize additional factors (raters, scales, etc.) and allow them to shape the scores (Morrow, 1979; Weir, 1990).

In the age of communicative language teaching and testing, almost thirty years after Rasch, Linacre (1989) extended the basic model to include, in addition to person abilities (Bn) and the difficulty of items (Di), adding the difficulty of tasks, "challenge" in Linacre's wording (2014b: 13), and the severity of judges (Cj) and

Fk, which is “the barrier to being observed in category k relative to category k-1” (2014b: 13).

Thus, MFRM may be summarised as follows:

Fig. 1 A basic model for MFRM

$$\log (P_{nmijk} / P_{nmij}(k-1)) = B_n - D_i - A_m - C_j - F_k$$

While Linacre (2014b) included four facets in this equation, there could be more, theoretically, without an upper limit. Indeed, an iTOLC productive test component, apart from items(tasks) and raters, additionally includes the facet of rating scales, such as *accuracy*, *vocabulary*, etc., which strongly resemble items in language testing.

Note that all the facets to the right of the equal symbol are mathematically made out as subtractions (Fig. 1). This is to reflect what happens to the total amount of variance in test-taker scores, revealing that MFRM distributes the observed variance between the assumed facets, creating measures of test-taker ability not contaminated by “freak” items and idiosyncratic raters, etc. To the left of the equal sign, note “log,” which reminds us that we are dealing with the logarithmic conversion of probability, the ratio between Pnmijk (the probability of category k being observed) and Pnmij(k-1) (the probability of category k-1 being observed). Linacre (2014b: 280) also explains the lowercase letters. They are represented below with minor changes:

B_n is the ability of person n, e.g., examinee Nora,

A_m is the challenge of task m, e.g., an essay "My day at the zoo,"

D_i is the difficulty of rating sale item i, e.g., punctuation,

C_j is the severity of judge or examiner j, e.g., Dr. Smith,

F_k is the barrier to being observed in category k relative to category k-1.

P_{nmijk} is the probability of category k being observed.

As has been alluded to above, in the design of iTOLC, a few more facets were assumed, all as modifications to Linacre’s Many-facet Rasch model (1989). The medium of test-taking was one of them.

The research questions

In this section, as a reminder, I present the research questions. The related rationale and operationalisations will follow below.

- Where should the medium of test-taking be located in a theory of performance? Is it a facet related to test method or the test-takers?
- Is there statistical evidence for a medium of test-taking as a facet of test performance?
- How do the three test-taking mediums compare in terms of difficulty? How do CBT and IBT, as mediums most relevant for use in the future, compare?

The rationale

The operation of iTOLC lends itself to comparing large numbers of test-takers taking the online (IBT) and the CBT versions of the exam in each of the four "papers" or test components. The basis of this research is comparing the item banks, one for each component. The item banks are constructed from the response data generated in the "live" administrations, the test versions administered to test-takers, allowing direct comparisons between the three mediums in this research. It should be added that each test version is anchored through at least two tasks, or "testlets" of individual items, to the same common item bank. This meant that in the combined-collated dataset (item-bank), from the skills components of reading, listening, writing, or speaking, from the period between 2019-2021, a web of common (anchor) items connected the individual test versions so that connectivity (comparability) of data could be achieved *within* the item banks. Comparability was achieved to believably interpret the difference in the difficulty of the three test-taking mediums concerning each other and compare item difficulties, test-taker abilities, etc., on the same scale.

One needs to address why the difference in mediums is particularly important at this juncture. As pointed out above, the original Rasch model (1960/1980) stipulated items that each had a difficulty level; thus, the notion of difficulty has always been central to Rasch methodology. The additional factors brought in by Linacre's extension (1989) of the model were also difficulties of a particular kind. Raters constitute "hurdles" or "challenges," -- to use another metaphor -- by being either easy to convince of the merit of the essay (lenient rater) or more difficult to convince (strict rater) of the same, all this resulting in a "rater difficulty" continuum between lenient and strict (severe) raters. All other facets in the Rasch model also constitute difficulties, including test-taking mediums.

The research reported here is mainly statistical (quantitative) since it is challenging to expect stakeholders (test-takers, examiners, or teachers) to discuss a proportion of the test score variance and comment on the relative differences observed between the mediums. The one question, however, about where in the competence-performance framework the medium of test-taking could be located should be decided based on logic and judgement, informed by the relevant literature, since there is no way to construct a hierarchy from construct elements and answer the question on the basis of that.

Operationalisations

In the case of the reading and listening papers (Fig. 2), the percentages were calculated as follows:

Fig. 2. The calculation of reading and listening percentage means

$$\frac{\text{the sum of all point scores for the medium}}{\text{the count of all responses for the medium}}$$

The calculation was somewhat more complex in the productive test components (Fig. 3).

Fig. 3. The calculation of speaking and writing percentage means

$$\frac{\text{the sum of all point scores for the medium}}{\text{the count of all scoring events for the medium}} = \frac{\text{mean score}}{\text{length of score scale}}$$

This way, a mean is obtained, 2.25, for example, 56% of 4, since the maximum obtainable score is 4 (0-4) on any of the iTOLC rating scales.

The mediums of test-taking are first illustrated with their raw score percentage means (Tables 1-2). Given that the mediums are like test items in that low mean scores indicate difficulty and high mean scores indicate easiness, some inference can already be made at this point about how the three mediums can be compared. However, it should be added immediately that raw score means might be deceptive because a host of performance factors influences them, but whose effect they do not show.

Table 1. English B2 raw score means in percentages

English		Reading	Writing	Listening	Speaking
N		5664	5499	4501	4346
Raw means	CBT%	52	57	59	56
	PBT%	62	76	63	49
	IBT%	56	60	62	58
SD %		4	8,3	2	4

Table 2. German B2 raw score means in percentages

German		Leseverstehen	Schriftlicher Ausdruck	Hörverstehen	Mündlicher Ausdruck
N		1365	1377	1088	1071
Raw means	CBT%	45	56	50	50
	PBT%	57	55	43	41
	IBT%	49	57	55	55
SD %		5	1	5	6

To estimate the difficulty of the mediums on their own without possible interaction from other factors, a second round of analysis was necessary, where the medium of test-taking was taken into account in isolation with respect to the rest of the facets (items, raters, scales, formats, etc.). These other facets were made not to participate, i.e., make no contribution to the calibration of abilities. In light of the multi-faceted nature of language proficiency, the calibration of the difficulty of test-taking mediums was made using MFRM, with logit values converted into "fair" difficulty values as described in Linacre (2014b: 120), which in turn were represented here as percentages (converted back to the percentage metric) for the sake of comparability.

The data

The analyses were made on the basis of collated/pooled response data in both languages, which included all the examinations since accreditation over a period of over two and a half years. The data also included pretests in the project period that preceded the accreditation of iTOLC. The data size in English is more extensive, with the collated data ranging between 4000 to 5000 test-takers per "paper." At the same time, the same in German is considerably smaller, with response data from 1000 to 1400 test-takers per "paper." (The breakdown to more precise counts (N) is in Tables 1-2.) The breadth of the data collected could also be appreciated by knowing that 90-95 pieces of information (Dávid, 2014) have been collected from each test-taker of the complex exam. This number includes the responses to the discrete-point items in the two receptive skills tests as well as the ratings in the productive skills test components. Thus, it may be stated that this research is not based on samples but on what might be called the iTOLC population since all the response data from the years 2019-2021 were collected, making the data as comprehensive as possible.

Results and discussion

It appears the answer to research question 1 may be formulated based on the choice test-takers make, a helpful "crutch," whether they take the exam in a testing centre or take the online version from outside a testing centre (e.g., from home). Once test-takers are allowed to decide the medium of test-taking for themselves, the medium can no longer be part of the testing method or belong to method-related performance factors. Still, it will become characteristic of the test-taker (a test-taker-related factor of performance). The test-taker choice, of course, is not entirely free, as we all know, since questions of available equipment, experience, access, etc., modify how the test-taker might want to take the test. Dictionary use, to take another example, should also be a test-taker-related factor because it depends on the decision of the test-taker to use or not use a dictionary, provided the test provider allows their use.

To answer research questions 2 and 3, the investigation of the raw score means provided an initial idea of the comparative difficulty of the mediums. In the English examinations (Table 1), the CBT version of iTOLC consistently appears to be more difficult than the IBT version. The PBT version seems to be the easiest except for the speaking component, where the IBT version appears to be the easiest. In German (Table 2), the tendency is less straightforward. The IBT version is the easiest except for reading (Leseverstehen), where the PBT version is the easiest. Most importantly, however, it is also true for the German examination that the IBT version is consistently easier than the CBT version in all four test components.

Due to some of the means in Tables 1-2 being very close and that raw scores do not consider performance variables, a second round of analysis was necessary. This round focused on the "fair" scores, which are logit values scaled back onto the percentage metric (Linacre 2014b: 279). For the sake of the experiment, the effect of the mediums without interactions from other factors, e.g., dictionary use, item format, etc., was investigated. This was necessary because when test-taker scores are calibrated in the standard mode of the operation of *Facets*, all other facets of performance are active and modify the calibrations of the mediums the researcher is interested in. As seen in Tables 3 and 4, the IBT test-taking medium consistently constitutes a lower "hurdle," i.e., presenting a lower mean difficulty than the CBT.

Table 3. English B2 "fair" means in percentage metric

English		Reading	Err.	Writing	Err.	Listening	Err.	Speaking	Err.
"Fair" means	CBT%	51	0.12	56	0.04	57	0.15	55	0.08
	PBT%	56	0.71	76	1.6	66	1.85	58	1.20
	IBT%	64	0.01	62	0.04	67	0.15	62	0.08
SD and mean err. %		1.31	0.32	2.00	0.56	3.08	0.72	3.14	0.32
Chi-square and d.f.		1407,8	2	2492.9	2	1199.4	2	3635.1	2
Sig:		0.00		0.00		0.00		0.00	

Table 4. German B2 "fair" means in percentage metric

German		Lesevers.	Err.	Schriftl.	Err.	Hörvers.	Err.	Mündl.	Err.
"Fair" means	CBT %	42	0.32	56	0.17	48	0.34	48	0.2
	PBT %	40	1.61	58	0.51	41	2.07	56	0.4
	IBT %	50	0.32	60	0.17	57	0.34	60	0.2
SD and mean err. %		2.90	0.75	1.53	0.28	4.83	0.92	3.23	0.25
Chi-square and d.f.		187.2	2	302.8	2	197.8	2	2736.0	2
Sig:		0.00		0.00		0.00		0.00	

In Tables 3 and 4, all logit values, fair ability scores, and error values included were converted to a scale of 100 units so that they could be thought of in an accessible way as percentages. The test component columns, such as Reading, Writing, etc., contain the means for the three test-taking mediums. The asymptotic error values are provided on the right of each test component column, which was

also converted into the percentage metric. If the error values are considered "halos" around the means of mediums and the halos fail to extend into each other, the chances are that the means represent a separate facet.

Below the three mediums, the mean standard deviations for the mediums by the test component are given with the related mean error. Further below, chi-square values with their degrees of freedom are provided, which form the basis of the significance. According to the designer of *Facets*, this "chi-squared distribution (also chi-square or $\chi 2$-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables" (J. M. Linacre, personal communication). These are provided for each set of medium means in an exam component that tests the reported means' distinctness. As shown in Tables 3-4, all the mediums represent significant differences between the medium mean values. At the level of 0.00l, the significance values also summarise whether the means of mediums in the relevant component is sufficiently spaced apart from each other to be considered a facet of performance.

The different test-taking mediums thus seem to bring construct-irrelevant variance (Messick, 1989; 1995) into the scores and test results. Construct-irrelevant variance is not desirable, given that it is the test provider's duty to measure competence irrespective of the test method or other circumstances that contaminate the assessment. Construct-irrelevant variance is not controlled if the results are reported in raw scores, as raw scores do not show the effect of a host of performance variables. However, if test results are computed (calibrated) with the use of probabilistic software, and if the medium of test-taking is thus taken into account, the software will compensate for the effect of the medium, eliminate the construct-irrelevant variance from the scores and the results will have a stronger claim to validity.

Conclusions

The most important outcome of this research seems to be the emergence of a facet of test-taking mediums resulting from technological developments and, in this country, at least, the pandemic. This article should take us closer to a more comprehensive view of performance factors and contribute to a possible model of actual performance in the future.

If the facet of test-taking mediums can be demonstrated in the iTOLC test data, it should also be present in other examinations. Whether it can be shown or not depends on the specific measurement technology used. Suppose they use software that can calculate the effect of the test-taking medium, compensating for a higher difficulty of CBTs, or whichever medium is more difficult. In that case, they can rid the ability scores from construct-irrelevant variance. If, however, they continue to calculate abilities in terms of raw scores, the construct irrelevant

variance will not be separated and will only increase the measurement error of the ability estimates.

For iTOLC, the above findings justify using MFRM technology to calculate the scores. iTOLC can legitimately continue to use the facet of mediums in their measurement of language competence and continue compensating test-taker scores in relation to the medium of their test-taking.

It is to be seen in the future whether taking the CBT version is more difficult for the candidates at the other three CEFR levels as well. There seems to be a strong indication that results will be similar for level C1. The fact that the number of test-takers is much smaller at levels B1 and A2 will naturally cast some doubt on those statistical outcomes until more data can be collected.

The philosophy of validation demands that rival interpretations also be investigated and rejected if possible (Messick, 1989). When all the facets are active, the relative easiness of the IBT medium may or may not be blurred by their interactions; therefore, the facet of test-taking mediums needs to be investigated in every assessment context. In this case, the most apparent rival interpretation, which would explain higher IBT means as an alternative, is that remote online testing can induce cheating more strongly than personally supervised CBT or PBT sessions. While cheating can never entirely be excluded, and indeed there have been a handful of problem cases reported since accreditation, iTOLC is equipped with the technology and observation facilities, dubbed online proctoring (Atoum et al., 2017), which can reduce the possibility of cheating. iTOLC has made a serious effort to exclude the possibility of cheating as much as possible by marshalling what technology could offer and providing online proctoring. The technology mandates the use of a second camera that can "look around" the room, wherever the test is taken, and the iTOLC interface also prevents "escaping" through switching to another browser or using a second keyboard, all this information being logged. Online proctors carefully monitor a maximum of 15 test-takers at a time, listening in and watching for unexplained conversations, strange response patterns, etc., and possible violations of security rules that ban headphones and certain hairstyles that can camouflage cordless earpieces.

Apart from the threats above, the action of the test-taking medium variable can offer advantages as well. Avoiding the pitfalls of the test-taker's limited, modified choice is one strong argument for why appropriate score calculation procedures were developed and, finally, the validity of test scores increased. Appropriate score calculation procedures, away from raw scores, are needed not only because construct-relevant variance should be increased and construct-irrelevant variance decreased but because measurement error can also be cut if the right facets are active in the dataset.

The outcome of this research also implies that if the exam provider uses raw scores, CBT and IBT mediums should not be used simultaneously, or if there is a clear need to administer both exam mediums, appropriate MFRM software should

be used, compensating for the relative difficulty of CBT and the easiness of IBT. Otherwise, the fairness of the test might be called into question. Test providers who administer both mediums ought, first of all, to research their own tests. If they find similar differences, they should allow compensation in calibrating abilities.

References

Akkreditációs Kézikönyv 2023. (2023). [Accreditation Handbook.] Oktatási Hivatal. Nyelvvizsgáztatási Akkreditációs Osztály. https://nyak.oh.gov.hu/nyat/doc/ak2023/ak2023.htm

Atoum, Y., Chen, L, Liu, A. X., Hsu, S. D. H. & Liu, X. (2017). Automated Online Exam Proctoring. *IEEE Transactions on Multimedia*. 19(7). 1609-1624. https://10.1109/TMM.2017.2656064.

Adams, R. J., Wu, M. L, Cloney, D., & Wilson, M. R. (2020). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 5. Camberwell, Victoria: Australian Council for Educational Research.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Babos, K., Gróf, Sz., Dömők, Sz., Kissné Adorján, J., Lehmann, M., Lukácsi, Z., Märcz, R., & Soproni, Zs. (2022). Az akkreditált számítógépes nyelvvizsgák tapasztalatai a 2018−2021 közötti időszakban. [Lessons from accredited languages examinations from 2018 to 2021.] *Új köznevelés*, *78*(7) 20–24.

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

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Baker, D. (1989). *Language Testing. A Critical Survey and a Practical Guide.* Sevenoaks, Kent: Edward Arnold.

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Brumfit, C. J. & Johnson, K. (Eds.). (1979). *The Communicative Approach to LanguageTeaching* (143-159). Oxford: Oxford University Press.

Canale, M. (1983a). From communicative competence to communicative language pedagogy. In J. C. Richards, J. C. & Schmidt, R. W. (Eds.). *Language and Communication* (2-27). London: Longman.

Canale, M. (1983b). On some dimensions of language proficiency. In Oller, J. W. Jr. (Ed.), *Issues in Language Testing Research* (333-342). Rowley, Mass.: Newbury House.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to language learning and testing. *Applied Linguistics.* 1(1). 1-47.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass.: M.I.T. Press.

Chomsky, N. (1980). *Rules and representations*. Oxford: Blackwell.

Dávid Gergely. (2014). Mitől jó egy nyelvvizsga? *Iskolakultúra*. 24(4). 3-25.

Eckes, T. (2015). Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments. 2nd Revised and Updated Edition. Frankfurt am Main: Peter Lang.

Goffman, E. (1967). *Interaction ritual.* New York: Doubleday.

Hymes, D. H. (1972). On communicative competence. In Pride, J. B. & Holmes, J. (Eds.). *Sociolinguistics: selected readings* (269-293). Harmondsworth, Middlesex: Penguin.

Kenyon, D., M. & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology*. 5(2). 60-83. http://llt.msu.edu/vol5num2/kenyon/default.html.

Kiszely, Z. (2022). Államilag elismert nyelvvizsgák tudományos feldolgozottsága magyarországi alkalmazott nyelvészeti és pedagógiai szakfolyóiratokban. *Iskolakultúra: pedagógusok szakmai-tudományos folyóirata*. 32(12) pp. 22-40.

Kunnan, A. J. (1995). *Test Taker Characteristics and Test Performance. A Structural Modeling Approach*. Cambridge: Cambridge University Press.
Linacre, J. M. (1989). *Many-facet Rasch Measurement*. Chicago: Mesa Press.
Linacre, J. M. (2014a). *Facets: Many-facet Rasch Measurement Computer Program.* Version 3.71.4 [Computer software] Winsteps.com.
Linacre, J. M. (2014b). *A User's Guide to Facets. Rasch-Model Computer Programs. Program Manual 3.71.4*. Letöltés: www.winsteps.com.
Littlewood, W. (1981). *Communicative Language Teaching: An introduction.* Cambridge: Cambridge University Press.
McNamara, T. (1996). *Measuring second language performance.* Harlow: Longman.
Messick, S. (1989). Validity. In Linn, R. L. (Ed.), *Educational Measurement* (13-103). New York: American Council on Education/Macmillan.
Messick, S. (1995). Validity of psychological assessment. *American Psychologist*. 50/9. 741-749.
Morrow, K. (1979). Communicative Language Testing: Revolution or Evolution? In Brumfit, C. J. & Johnson, K. (Eds.). *The Communicative Approach to Language Teaching* (143-159). Oxford: Oxford University Press.
Morrow, K. & Johnson, K. (1981). *Communication in the classroom.* Essex: Longman.
O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*. 19(2). 169-192.
Oliveri, M. E., Lawless, R. & Young, J. W. (2015). *A Validity Framework for the Use and Development of Exported Assessments*. Princeton: Educational Testing Service.
Oller, J. W. Jr. (1976). Evidence of a general language proficiency factor. *Die Neueren Sprachen* 76. 165-174.
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and achievement tests*. Chicago: University of Chicago Press.
Saussure, Ferdinand de. (1997). Bevezetés az általános nyelvészetbe. Lőrinczy Éva, Bokorné (ford.) Budapest: Korvina kiadó. (1916). *Cours de linguistique générale* ,
Stern, H.H. (1978). 'The formal-functional distinction in language pedagogy: a conceptual clarification'. paper read at the 5th AILA Congress. Montreal, August. Mimeo.
Weir, C. J. (1990). *Communicative Language Testing*. Hemel Hempstead, Herts: Prentice Hall.
Wind, S. & Hua, C. (2022). *Rasch Measurement Theory Analysis in R*. New York: Routledge. https://doi.org/10.1201/9781003174660
Widdowson, H. (2001). Communicative language testing: The art of the possible. In Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (Eds.) *Experimenting with Uncertainty: Essays in honour of Alan Davies* (12-21). Cambridge: Cambridge University Press.