A HUMAN FACE TO TECHNOLOGY? DIGITAL EXAMS ARE COMING

doi.org/10.61425/wplp.2020.15sp.1.13

Gergely A. Dávid
Department of English Language Pedagogy
Eötvös Loránd University, Budapest
david.gergely@btk.elte.hu

Abstract: Technological development has reached a point where computerised foreign language testing is possible as well as desirable. It was not caused by the pandemic, which only speeded up development as all the elements of digital, then online testing were already in place. Technological developments are doubly interesting because they also appear to militate against communicative language testing, the prevailing orthodoxy in the field for the past forty years. At the same time, the field of language testing also faces other challenges, such as a language tests becoming less of a social act, stakeholders functioning in changing roles and last but not least, measurement itself is facing a challenge from business logic.

Key words: computer-based tests, computer adaptive testing, communicative language testing, authenticity

1 Introduction

Technological developments are changing the face of language testing. The past two years have seen the rapid rise of computerised language testing in Hungary. The change was not a surprise, but was pending for a number of years. For example, an early feasibility study indicates that the idea for a fully computerised foreign language examining system was engendered as early as the end of the millennium (Huszty & Dávid, 2000), but the subsequent gestation period was almost two decades. It was only recently, it appears, that the time was ripe for the introduction of digitalised language testing systems, especially if one considers comparable and timely developments abroad. In March 2019, the International Test of Language Competence (iTOLC) language exam was accredited as the first fully digital foreign language testing system in Hungary, following as second only to one other language exam, two of whose skills components were not digital. Since then, there has been news of other exam providers following suit and launching their own digital exams. It could not have been accidental that after the initial and tentative concepts, as alluded to above, more and more foreign language testing systems transition to digital in a short period of a few years. Indeed, digital exams are coming. It would have happened without the pandemic too, perhaps, in less of a haste.

In the discussion below, some clarification is provided first, for a range of relevant terminology and kinds of digital tests, stating that computer-based testing systems are not a monolithic group and there is room for an intermediate technology. Then the future is investigated in the form of the various challenges that the foreign language testing as we have known it for the past forty years is facing: The final section is devoted to the challenge to measurement itself.

2. A review of relevant terminology: CBTs, CATs, PBTs and IBTs

Digital technology in the field of language testing brought about a range of solutions that may be sorted into a small number of categories, all acronyms. As opposed to the traditional paper-based tests (PBT), the newer category of computer-based tests (CBT) gradually emerged. As a special CBT category, one might continue distinguishing computer-adaptive tests (CAT), in which test items are automatically selected by the computer (algorithm) on the basis of statistical calibrations from a set of previous item performances, all available in an item-bank (Davies et al., 1999). The most recent CBT category is that of internet-based tests (IBTs), where the test on a central server is accessed by test-takers through the web.

2.1 The mirage of CAT solutions

The computer-adaptive solution may still be considered to be the "high-end" of a range of computer-based solutions (Antal & Erős, 2010; Csapó et al., 2008), due to the technical advances it can demonstrate, but it appears rather inflexible and unwieldy because of the sheer size of the bank of items it is supposed to operate from. In order for it to work properly, a very large number of items are needed, all having been previously calibrated. The item-bank of a CAT should also include the statistical information as well as the items themselves. Moreover, it should be able to make a selection from all items that are calibrated at the targeted level and have comparable, equally acceptable statistical properties, all of which is clearly unattainable, at least in Hungary, both for real technical-logistical reasons and the constraints that follow from the regulations in the Hungarian foreign language accreditation scheme (Akkreditációs Kézikönyv [Accreditation Handbook], 2021). Take the example of a computerised language exam in one of the frequently taught languages, where technology allows the delivery of at least about 40 tests per calendar year. The construction of either the listening or reading "paper" would demand the writing, moderation and pretesting of at least 80 tasks (assuming a minimum of two in this case), including their texts (written and/or recorded) and each with 10 items. In this way, the number of items that need to be written should come to 800 in either of the skills, not counting items that get screened out in the pretest and are best left as draft. In order for an adaptive system to work well, the item banks need to be considerably larger than that, since the calculations above only describe the bare minimum of foreign language examinations on the basis of the official accreditation requirements of representativeness, as published in Akkreditációs Kézikönyv (p. 23). If a large pool of items is not available, the very act of computerised selection becomes impossible. In order to illustrate the magnitude of the resources needed for adaptive systems, the writer of this paper can confirm that only about 80% of this target range was in any way achievable for a two-year test development project that involved, to varying degrees, 6-8 or more trained item writers for English.

It is thus fair to say that the amount of labour to go into the development of a CAT, so that it can really work, is daunting. By comparison, the CAT Antal and Erős (2010) designed, only contained about 180 items in its item-bank. It is not surprising that, as a result, Antal and Erős were only able to statistically simulate the functioning of their design on the computer. Although Csapó et al. (2008) make reference to the need for a large initial investment of setting up and launching a CAT system being very expensive, they do not discuss how large the item bank must be in order for it to properly operate from the beginning. As a result, it is not

surprising, once again, that CATs are a rarity and barely more than detailed plans (Antal & Erős, 2010).

2.2 CBTs are not a monolithic group

Computerised tests that are not computer-adaptive (CAT) should all still fit into the broader category of being computer-based. On closer inspection, however, one might observe that a number of different computer-based solutions exist, asking for further differentiation since this category is far from homogeneous. This view is shared by Csapó et al. (2008), who identify different levels of solutions to the category CBTs being heterogeneous. The simplest of this kind is not more than paper-based test material committed to the screen. The test material, the sequence of items or tasks that test-takers must do, is mediated to the test-taker by the computer on analogy of a PBT. A more highly developed solution is when the CBT is essentially a database and a digital platform with interfaces for the item-writer, the moderator, examiner and, last but not least, the test editor, whose job is to turn the items/tasks into computer files, not necessarily similar in appearance to the creation of the writer of the items. One special category is when only one or more of the "papers", e.g. the reading or listening components of the exam, are mediated through the computer, rather than the whole examination. Such a combination might be referred to as a semi-computerised solution.

2.3 CBTs and „enhanced" CBTs

A more complex version of a CBT is a test that goes beyond managing and standardising the appearance of the test tasks and computerised delivery. Some of its features may be directly appreciated by the candidates. For example, it can indicate word limits for the candidate, or facilitate truly independent and unbiased rating by not having raters make notes or write anything, for example, on the pages of the test of writing. In a paper-based test, such comments would most likely be seen by the other rater, unless a separate form is used for reporting observations and comments. The computer can also keep track of the time elapsed (or remaining), allow candidates to do the tasks in their own preferred order and greatly shorten the time until the publication of the results.

CBTs may also offer benefits to examiners and language testing specialists, which are not necessarily appreciated by the candidates. These CBTs can combine test delivery with data management, test analysis and measurement. The computer can be programmed to record all candidate responses, assist the rating process and to facilitate item and test analysis by collating response data and actually producing the data files for the analyses. However, on top of a host of improvements, CBTs can, if so designed, also offer the use of powerful statistical analytical software, integrated with the CBT interface and database software that manages the CBT. In this way, the whole process of testing, from the commencing step of test delivery (or even that of pretesting) to the final step of the publication of the results and validation studies can be considerably shortened while still maintaining professional and quality management standards. "Enhanced" CBTs are thus possible, through the intensive use of software, bridging the gap between fully automated, "high tech" CATs and the "low-tech" of basic CBTs and traditional PBTs.

One further advantage to be appreciated by test providers might be the capability of CBTs to monitor how candidates use one or more integrated digital dictionaries and to prevent

the use of any non-integrated online source. The reason why dictionary use is important is because it is highly doubtful whether it is a true factor of language proficiency. Some of the older literature that discuss dictionary use show that researchers (examiners, ushers, clerical workers) had to observe the examinees in the exam room as they were doing the PBT tests, in order to be able to make observations about dictionary use. Many researchers had to be asked to be observers (Bensoussan et al., 1984). Observing examinees while the test is being done clearly has its limitations such as the low ability of the observer to watch many candidates reliably, to pay attention evenly and constantly, etc. To make matters worse, dictionary use was first allowed in Great Britain in 1998, for unclear reasons. Worse, permission was withdrawn a few years later, casting some doubt whether the original permission was given with adequate research and enough circumspection (Hurman & Tall, 2002; Tall & Hurman, 2002). The use of computers can thus provide us with much needed accurate information concerning dictionary use, which can yield potential validity information. In addition, CBTs have potential in preventing cheating that occurs either through the misuse of printed dictionaries or by allowing test-takers to log into unauthorised online sources for assistance.

2.4. Intermediate technology

The unfeasibility of adaptive solutions in most situations may give rise to the concept of intermediate technology in language testing. PBTs or basic CBTs could be enhanced by experts so that items would not be included randomly (chosen from the pool of items calibrated) in the test sequence, but tasks follow each other according to a planned sequence and are connected in an anchored design. Similarly, test (pretest) response data could be generated from the database to form the right input files for the analytical software to process and test results generated from the output files of the analysis. It appears, thus, that enhanced CBTs can provide the "intermediate technology" that this country can sustain at present, which in turn should go a long way to satisfy professional and process requirements in foreign language testing. To be fair, it should be added, software-assisted language testing is also possible on the basis of PBTs (Dávid, 2014).

3. Threats and challenges: A leap in the dark?

There is still so little we know about computerised (electronic) testing. The ideal would be an electronic assessment system that is both fully digitalised and also communicative, in which the stumbling block clearly seems to be speaking proficiency. While rating by humans still seems to be indispensable, with artificial intelligence still in need of 5-10 years' development at the level of daily practice, interacting with a machine appears to be antithetic to some. It is also a possibility that computerised assessment systems will mark the end of communicative language testing as we know it, if appropriate digital solutions cannot be provided soon.

In a number of ways, transitioning to computerised testing of the foreign language is a leap in the dark and presents challenges to language testing as we know it.

1. Computer-based testing is likely to impact the concept that testing is done by people to people.
2. Stakeholders of language testing might also change, affecting the roles and experiences of the expert professional, individually or in groups.

3. Computerised language testing may well challenge the dominance of the construct of Communicative Language Competence and through that Communicative Language Testing, still the prevailing orthodoxy in the field.
4. Finally, computerised testing may mount a challenge to measurement in the sense that we might see an end to testing in the largest possible groups.

The predictions articulated below in more detail are tentative, of course, as more evidence and experience might expose them as wrong. However, as with all new developments, threats and challenges open up opportunities and possible benefits as well, as is appropriately put by Henry Miller, the controversial American writer: "All growth is a leap in the dark, a spontaneous unpremeditated act without the benefit of experience" (1941, p. 91).

3.1 "I want my examiner": A social act

Computerised exams (some, at least) pose a challenge to the experience of being involved as test-takers and raters or "interlocutors". McNamara (2000) reminded of the social character of language tests, although what he appealed to was for language test developers to take into account the consequences of testing, or more broadly, the ethical issues that spring from tests that assume a gate-keeping role. His argument is an argument for consequential validity, following Messick (1981,1989), in that ethical considerations should bear upon the results of the measurement per se. An alternative way of interpreting the label "language testing as a social act" might be the need to create human contact (or maintain some of it) in the face of advancing technology, much like the need for human contact in a pandemic. In terms of popularity with test-takers, some CBTs, especially the speaking tests where human contact is never actually made, may be at a disadvantage since there is no examiner present in person at the time of taking the test.

Computer-based solutions appear have a certain non-human quality about them. Technology has enormous appeal, but it may also be off-putting. The writer of this paper is aware of a test development project where feedback from pretests indicated that test-takers missed the human element having to take a fully computerised test. One other instance, of personal experience, was when a pretest candidate refused the computerised speaking part, claiming they would not talk to a machine. In general, there seems to be a discernible relationship of love and hate, even if impartial assessment is virtually guaranteed in a CBT solution, where there is only an audio recording made and then rated.

A 'technological' test is nothing new, of course. The magnetic tape mediated interview test, known as the Semi-direct Oral Proficiency Interview (SOPI) had been available for decades, until its very medium, magnetic recording became obsolete. According to the literature, as early as 1988, Shohamy et al. compared the SOPI with another direct test of oral proficiency, a traditional one, in which the examiner and the candidates interacted in person (face to face) and human contact was made. Shohamy (1994) found that while concurrent validity of a SOPI was high in light of the statistical analyses, other aspects of the tests, such as their discourse features and communicative strategies as well as the nature of elicitation tasks showed differences on the basis of qualitative data (p. 99). At the moment, it is not known whether the fully computerised CBT will become popular enough, to match the popularity of the PBT and face-to-face exams. The question of what their future prospects are hangs in the balance.

3.2. Stakeholders change

What professionals in language education like to believe is that test-takers need to be educated towards the language test they are to be taking, but not just so that they can do well on the test, but also that they understand what is happening to them. This is also true for computerised exams, both according to authorities in the assessment profession (Csapó et al., 2008) and the writer of this paper. It is also the case that their teachers need to be educated (Csépes, 2014).

Another category of stakeholders that appears to be on the rise is that of clerical workers. To initially handle complaints from candidates or decide about pretests seem to be professional matters, but managers of exam providers appear to be tasking their clerical staff with these jobs. At the moment, it is not clear whether this is poor organisation that should simply be rectified, or there is a trend towards a decreasing role for the examiner/measurement specialist. This is a useful question since automated systems might call less and less for the intervention of the measurement specialist, and depending on the level of the digital solution, for the examiner's role as well, if speech recognition is applied for the assessment of speaking. It might also be the case that automation, taken further and further in computer-based solutions, sends the unintended message that there is now little room for the measurement specialist. Last but not least, it is also the possibility that the manager sees the opportunity for cost cutting in this way and in order to pay less in salaries, hires a less well-trained work force or hire those who have not been trained at all.

3.2.1 Communicating from different professional backgrounds

Although language testing professionals (examiners, applied linguists, language teachers) and information technology (IT) personnel are both categories of stakeholders of digitally delivered exams, one of the greatest difficulties, as it seems, is the communication between these two groups, i.e. those who design the digital testing platform, the interfaces, etc. and those who are responsible for the measurement aspects (test design, materials writing, rater training, rating etc.). Having gone through years of test development, the writer of this paper can only underscore the importance of good communication between IT experts and testing professionals so that the best possible results are obtained. Developing CBTs should be an interdisciplinary venture. IT experts and applied linguists should collaborate in developing such exams.

However, experience from the past three years has shown that the work cultures in the above two groups are very different, including work habits such as what time they typically begin work in the day or what time they are available for consultation on a typical working day. Apart from differing work cultures and habits, these two groups also know rather little about each other's professions. As an applied linguist (language teacher), the writer of this paper was perhaps more sensitive to what appears to be the IT experts' awareness of testing foreign language competence. Having gone to school, they have all done classroom tests, took the school leaving examination and sat for tests and examinations later at university, all of which forms their experiential bases. Some have a good predisposition to inferential statistics and the applied mathematics therein, matching comparable expertise in word processing, spreadsheets, cloud-based applications (e.g. Google Drive) and test statistics among the applied linguists, but

experience clearly shows that it is difficult to communicate across the barriers created by divergent experiential bases.

One other difficulty sprang from differing concepts of collaboration and software development. As it turned out at one of the first project meetings (same project, see above), the IT personnel wanted the applied linguists to draw up the specifications precisely in advance of the development of the item-banking software, which then they would implement. This was a concept of a single round of development, in which the software was to be responsible for uploading test material, storing a large number of items and tasks and delivering a selection of these, and finally for the management of scores and results. The single round type of development is often referred to as the *waterfall* type of software development, on the basis of it being one-directional and single cycle. In this traditional approach, a typical way to work is to ask the client to draw up specifications of requirements, which will be followed by the design of the requested software and concluded by handing it over for use (Wikipedia, n.d.). For the language teachers in this project, it was patently obvious that such an approach would hardly work, given their familiarity with collaborative processes and the reflective cycle in their own profession, as articulated successfully, for example, by Wallace (1991) for teacher training.

It appeared to this writer that waterfall was inappropriate for contexts in which communication must be ensured between different groups and in which there is no precedent for the participants, i.e. no previous experience of having developed a CBT. Digital exams presented such an uncharted territory to which everyone involved was a newcomer. Thus, instead of the concept of a single round of development, convenient as it may be, a more recent approach, called *agile software development* should be followed (Edge, 2020; Martin, 2002; Tabaka, 2006). The agile approach is iterative and places emphasis, most importantly, on the cooperation (of two disparate groups of professionals, in this case) as on "responding to change over following a plan" through, as was stated in the *Agile Manifesto* by leading IT professionals. (Beck et al., 2001). Apart from responding to change, the manifesto includes other important features of the agile approach, there being a preference for

1. individuals and interactions over processes and tools,
2. working software over comprehensive documentation,
3. customer collaboration over contract negotiation.

3.3 A challenge to Communicative Language Teaching and Testing?

The introduction of computerised foreign language testing may well pose a challenge to Communicative Language Testing as we know it, especially but not exclusively to the testing of speaking/oral skills. Whereas the end of the 1970s saw the rise of Communicative Language Teaching and later Communicative Language Testing (CLT), the coming decade might see the end of it. There has been discussion of the future of CLT recently. As McNamara (2014) pointed out, citing new developments of technology, it may be difficult, or deemed unnecessary, to stage a fully communicative speaking test on the computer, when the live participation of the interlocutor, modelled on face-to-face interaction cannot (would not) be achieved. Prompts, of course, may be recorded voices by one or more speakers, all recorded prior to the testing session, but, except for the purely technological sense, this would not be any more advanced than the Semi-direct Oral Proficency Interviews (SOPI) of the 1980s were, since it would only amount to substituting computer media for the now outdated magnetic tape of forty years ago.

The early theorists of CLT articulated an argument for the fluid nature of real communication, based on contingency (Morrow, 1979, 1982, 1983, 1986; Weir 1990). What speaker B might respond to A is never actually known, perhaps only suspected and expected by speaker A. Speaker B's response in turn will be followed by the identically contingent response of speaker A. This contingency marks two of the seven crucial features that Morrow (1979) identifies as communication being interaction-based and unpredictable (p. 149). If these two are missing from the test task, full (any?) communicativeness cannot be expected.

One would think that until artificial intelligence (AI) becomes sufficiently more broadly available and considerably cheaper, a limited version of communicativeness is to be expected and accepted. However, it is also possible – in fact it is already happening – that examination providers part with Communicative Language Testing so that they can more freely use digital technology (voice recognition at the stage of development it is now) to test language proficiency without any human involvement. As McNamara (2014) correctly observes, the task types used in the Versant Test (Pearson Education, 2019) may indicate a break with communicativeness. It has task types of *reading out sentences*, *short answer questions*, *building up sentences* from words, *story retelling*, and *open ended questions*, of which all but the last one have little to do with communicative language testing as we have known it in the past decades. The reading test is essentially the reading aloud of sentences presented to the candidate, while in the building up exercise phrases and words are presented to the candidate, from which they construct sentences. As such exercises work not at the text level, but only at the level of sentences, so they would not pass Morrow's criteria (1979) for a communicative test. Thus, as it seems, language testing is being steered away from a communicative model of language proficiency, which has essentially been a sociolinguistic model, a legacy of Hymes (1972) and Canale and Swain (1980). At the same time, AI does not seem to be at an appropriate stage of development just yet so that tests can be designed that are both communicative *and* digital. AI does not seem capable of rating content and the appropriacy of the message automatically in productive components of a test. As a result, AI is made to focus on speech production, harking back to earlier psycholinguistic or even structuralist approaches to testing language proficiency (McNamara, 2014, p. 231).

3.3.1. Open ended task types

To compensate for the technology-related limitations in the communicativeness that CBTs can offer, an increased use of open-ended task types might be recommended in the testing of reception (reading, listening), such as answering open-ended questions, to which test-takers either respond with complete sentences or with completing the answers started for them. However, it seems to be a matter of judgement whether the use of such open-ended items will actually increase the communicative nature of the test. At the moment, it may be argued that they do contribute to communicativeness, likely increase the validity of candidate responses, assuming the test provider validates the key before automatic scoring begins, but research still needs to provide support for the claim of improved communicativeness.

3.3.2 Authenticity

Another focus of the challenge to communicative language testing is to do with the claim of authenticity by proponents of CLT (Morrow, 1979; Weir, 1990). Authenticity has been seen by them as an important feature of communicative tests. One way to conceptualise achieving authenticity is if published pedagogical (teaching) materials are not used in the

construction of items and tasks (Akkreditációs Kézikönyv, 2021, p. 24). As far as texts go, this principle only permits the use of input texts originally not meant for pedagogical purposes. This is a limited, practical conceptualisation of authenticity. The reason that might be given for it is that teaching materials cannot be conceived as authentic because they must have been rewritten to suit the pedagogical purpose. A further -- broader -- concept of authenticity is that authenticity of the situation (context) by replicating a real-life situation in the testing context. This is the real-life approach as Bachman (1990, p. 302) dubbed it, which continues to be unrealistic because the many specific situations are bound to be too specific or unique for generalised conclusions (test results) to be drawn on their basis. Bachman identifies it with what he called the interactional/ability approach to authenticity. In this approach, it is the language tester who analyses the target language performance to extract distinguishing characteristics of communicative language use, which are then included in the specifications and used as a basis to generate tasks. As may well be observed here, the language tester goes a long way generalising target language use to obtain the focuses for their tests.

Whichever conceptualisation teacher and expert might accept, the test will have its own authenticity, as an authentic testing experience, but not as real-life authenticity or some non-testing situation. It will predictably be treated by test-takers as an authentic testing situation, irrespective of the issue whether candidates engage in genuine communication in the context of, say, a shopping situation with an information gap or not, as they would in a truly communicative test. Candidates will be painfully aware of the testing context, i.e. that they are being tested rather than being in of the context simulated for them in the test task (Lewkovitz, 1997, 2000). Ultimately, it will be seen what players in the language testing field will do. It will be revealed through what CBTs will be designed in the future and whether authenticity of context and situation will be seen as an integral part of foreign language testing or not. On the basis of present experience, CBTs may actually sharpen the contrast, the already present tension between the authenticity requirements of CLT and the capabilities of the technology available for CBTs at present.

3.4 The challenge to measurement

It seems to be fairly clear to this author, on the basis of first-hand experience, that computerized test delivery allows (literally invites) the fragmentation of the test-taker populace, which is a real threat for measurement. Test providers will administer exams more frequently because CBTs allow the testing cycle to be shorter since the testing material does not have to be printed on paper at a printing press, nor does it have to be transported in secure envelopes to the exam centres. Once the tests have been administered, the scripts do not have to transported back to the exam provider and redistributed among raters, nor do they need to be recollected from raters once the rating has been done. Separately organised oral testing occasions, often on different dates, do not have to be continued either, but computerisation allows them to be pulled together into one sitting, with candidates working in booths. It seems fairly clear that advancing technology has lifted the logistical constraints on the frequency of examination dates that can be planned for. Test providers will follow their business interests in administering their exams as frequently as possible.

Measurement specialists and lawmakers should take note of the threat that this development poses to reliability and validity, which are much harder to ensure (sustain) between a large number of test administrations to smaller groups of test-takers in each administration (dataset) than to ensure the same between large datasets from fewer

administrations. While the significant and positive outcome of the computerisation we are experiencing is the considerable shortening of the complete cycle of testing, from planning and organising the test to announcing the results, the downside and threats are already plain to see. The fragmentation of the candidates into smaller, often quite small, groups -- not to mention the possibility of individual, single candidate administrations -- is likely to bring about comparability problems. Comparability will certainly be an issue for the simplest type of CBTs that only deliver the tests via the computer and can offer nothing in terms of software assisted language testing as described above.

With the coming of the age of CBTs, the question for the quality control manager, or the measurement specialist who plans the testing procedure is no longer how high they can go in terms of the number of responses in a dataset and the number of respondents (candidates for a single examination), although administering tests to a larger number of candidates always remains desirable. The issue in the future seems to be just how low one can go in terms of the number of candidates in any one administration, while keeping the principle of putting candidates into groups for testing. With lower and lower numbers, statistical programmes increasingly produce extreme, volatile and therefore unpredictable (or unreliable) results, rendering, especially, software based on Classical Test Theory (CTT) unusable.

It will probably be raised at this point that if quantitative solutions were unworkable, a qualitative approach to quality management should be substituted for it, as is proposed (mandated) in AK (2021). However, qualitative methods typically take a long time to produce results. Some of the methods proposed in AK (2021, p. 85), such as the use of verbal protocols, for example, will certainly take longer than the reduced time the computer-based solutions imply for a whole testing cycle from the assembly of test material to the publication of results. There is no doubt that once shorter cycles are made possible by technology, the business community is not going to accept time-consuming methods of quality management.

Computer-based solutions can better respond to the challenge, provided they are supported by probabilistic measures (IRT or Rasch software) from an item-bank. In such analyses, the technical minimum number of candidates seems to be as low as 2 because the calibrations are based on differences – between candidates, items and raters. It therefore follows logically that a single candidate, item or rater cannot show differences and thus cannot be processed with probabilistic software either (Linacre, 1994, 2014). Even small datasets, larger than the above, are prone to malfunctioning because with low numbers there is a good chance that, for example, raters of writing might award the same score for the same assessment point of view across all, say, three candidates and both tasks. In the same vein, the principle may be formulated as ratings are needed for a minimum of two candidates, by at least two raters, for at least two items or tasks – with the wording reminiscent of “government of the people, by the people, for the people”, a line in the equally fundamental Gettysburg Address (Lincoln, 1863).

In cases of low count, or numbers, the unorthodox solution available for the specialists is running the analyses in conjunction with pretests and previous exams with the same tasks. However, the solution demands experience as it is technically advanced. Moreover, the fragmentation described above can also bring about a temptation to “cut corners” in quality control (management) by skipping the analyses, out of a lack of appropriate training or the lack of will to live up to the professional standards. Thus, the computerised medium, it may be stated, does not guarantee quality in terms of testing methodology or quality control. The business mind will certainly perceive the opportunity to make more money by increasing the frequency of exam dates, whereas for low numbers of test-takers the measurement profession will only be

able to offer qualitative solutions that are slow and quantitative solutions that fail to work with low numbers. Less popular CoE levels, at which there are fewer test-takers, seem especially vulnerable (practically all the CEFR levels, except B2 and C1, in Hungary).

3.5 Cost-efficient language testing?

Finally, it might be asked what provides the impetus for the various challenges to the field of language testing as outlined above, or whether examinations and experts can provide an appropriate response.

It is hard not to notice that privately owned companies (and not state-owned/state-run operations such as state-run higher education) seem to take the lead in development towards digitalised exams. As is well-known, language testing is a field where investment is more strongly made for financial gain than elsewhere in education. Consequently, it should also be the case that cost-efficient thinking will come to play a more prominent role in the future. Thus, it might be proposed that impetus for the development of digital exams comes from business; to affect various aspects of language testing.

As is known by many in the field, we owe it to Morrow (1979), following Spolsky (1976), that the three consecutive ages of language testing are distinguished as the "Garden of Eden", the "Vale of tears" and the era of Communicative Language Testing as the "Promised land". But what is going to follow after the "Promised Land"? Because of the strong presence of the business interest and seeing just how powerful it is, it is suggested that the next era of language testing be called Cost-efficient Language Testing.

Proofread for the use of English by: Andrea Thurmer, Department of English Language Pedagogy, Eötvös Loránd University, Budapest

References

Agile Software Development. (2020, December 29). In *Wikipedia*. https://en.wikipedia.org/wiki/Agile_software_development

Akkreditációs Kézikönyv 2021 [Accreditation Handbook 2021] (2021). Nyelvvizsgáztatási Akkreditációs Központ, Oktatási Hivatal [Educational Authority Accreditation Centre for Foreign Language Examinations.] https://nyak.oh.gov.hu/nyat/doc/ak2021/word/Akkreditacios_Kezikonyv_2021.pdf

Accredition Handbook 2021. (2021). (English version). Nyelvvizsgáztatási Akkreditációs Központ, Oktatási Hivatal [Educational Authority Accreditation Centre for Foreign Language Examinations].

Antal, M., & Erős, L. (2010). Item-válasz-elmélet alapú adaptív tesztelés [Item response theory based adaptive testing]. In K. Á. Bíró & Gy. Sebestyén-Pál (Eds.), *XI. ENELKO - XX. SzámOkt Nemzetközi Energetika - Elektrotechnika és Számítástechnika Konferencia* (pp. 101 - 106). Erdélyi Magyar Műszaki Tudományos Társaság (EMT).

https://nyak.oh.gov.hu/nyat/doc/AH2021-eng/EN_Accreditation_Handbook_2021.pdf

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

Beck, K., Grenning, J., Martin, R. C., Beedle, M., Highsmith, J., Mellor, S., van Bennekum, A., Hunt, A., Schwaber, K., Cockburn, A. Jeffries, R. Sutherland, J. Cunningham, W. Kern, J. Thomas, D., Fowler, M., & Marick, B. (2001). *Manifesto for Agile Software Development. Agile Alliance.* http://agilemanifesto.org/

Bensoussan, M., Sim, D., & Weiss, T. (1984). The effect of dictionary usage in EFL test performances compared with student and teacher attitudes and expectations. *Reading in a Foreign Language, 2*(2), 262-276.

Csapó B., Molnár Gy., & R. Tóth, K. (2008). A papíralapú tesztektől a számítógépes adaptív tesztelésig: A pedagógiai mérés-értékelés technikájának fejlődési tendenciái [From paper-based tests to computer-based adaptive testing: Trends in the development of pedagogical assessment techniques]. *Iskolakultúra, 18*(3-4), 3-16.

Csépes, I. (2014). Language assessment literacy in English teacher training programmes in Hungary. In J. Horváth & P. Medgyes (Eds.), *Studies in Honour of Marianne Nikolov* (pp. 399-411). Lingua Franca Csoport

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to language learning and testing. *Applied Linguistics, 1*(1). 1-47. http://dx.doi.org/10.1093/applin/I.1.1

Dávid, G. A. (2014, November 20). Software-assisted measurement and validity: Performance testing [Powerpoint slides]. Habilitation Lecture. Eötvös Loránd University,

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of language testing. In M. Milanovic (Ed.), *Studies in language testing* (Vol.7). Cambridge University Press.

Edge, J. (2020). *Agile: A guide to agile project management with scrum, kanban, and lean, including tips for sprint planning and how to create a hybrid waterfall agile software development methodology.* Bravex Publications.

Hurman, J., & Tall, G. (2002). Quantitative and qualitative effects of dictionary use on written examination scores. *The Language Learning Journal, 25*(1), pp. 21-26. https://doi.org/10.1080/09571730285200061

Huszty, A, & Dávid, G. A. (2000). *Megvalósíthatósági tanulmány az Első Magyar Nyelvvizsgacentrum projektről* [Feasibility study for the first Hungarian foreign language testing centre]. Mimeo.

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

Lewkowicz, J. A. (1997) Authentic for whom? Does authenticity really matter? In A. Huhta, V. Kohonen, & S. Luoma, (Eds.) *Current developments and alternatives in language assessment* (pp. 165-184). University of Jyvaskyla.

Lewkowitz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing, 17*(1), 43–64. ttps://doi.org/10.1177/026553220001700102

Linacre, J. M. (1994). *Many-facet Rasch Measurement*. Mesa Press.

Linacre, J. M. (2014). A user's guide to facets. Rasch-model computer programs. Program Manual 3.71.4. https://www.winsteps.com/manuals.htm.

Lincoln, A. (1863). *The Gettysburg address* [Transcript]. Retrieved from https://www.gilderlehrman.org/sites/default/files/inline-pdfs/06811_FPS.pdf

Martin, R. C. (2002). *Agile software development, principles, patterns, and practices*. Pearson Education (US).

McNamara, T. (2000). *Language testing.* Oxford Universty Press.

McNamara, T. (2014). 30 years on—Evolution or revolution? *Language Assessment Quarterly, 11*(2), 226-232. https://doi.org/10.1080/15434303.2014.895830

Messick, S. (1981). *Evidence and ethics in the evaluation of tests: Research report.* Educational Testing Service.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). American Council on Education/Macmillan.

Miller, H. (1941). *The wisdom of the heart*. New Directions Publishing.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C.J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143-159). Oxford University Press.

Morrow, K. (1982). Testing spoken language. In J. B. Heaton (Ed.), *Language testing* (pp. 56-58). Modern English Publications.

Morrow, K. (1983). Some comments on issues. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 115-119). Academic Press.

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing* (pp. 1-13). NFER-Nelson.

Pearson Education Inc. (2019). *Versant English test. Test description and validation summary*. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-English-Test-Description-Validation-Report.pdf

Shohamy, E. (1988). A proposed framework for testing the oral language of second/ foreign language learners. *Studies in Second Language Acquisition, 10*(2), 165-180. https://doi.org/10.1017/s0272263100007294

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99-124. https://doi.org/10.1177/026553229401100202

Shohamy, E., Gordon, C., Kenyon, D., & Stansfield, C. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education*, *4*, 4-9.

Spolsky, B. (1976). *Language testing: Art or science?* [Conference Paper]. The 4th International Congress of Applied Linguistics. Stuttgart, Germany.

Tabaka, J. (2006). *Collaboration explained: Facilitation skills for software project leaders.* Pearson Education (US).

Tall, G. & Hurman, J. (2002). Using dictionaries in modern languages GCSE Examinations. *Educational Review, 54*(3), 205–217. https://doi.org/10.1080/0013191022000016275

Wallace, M. J. (1991). *Training foreign language teachers: A reflective approach*. Cambridge University Press.

Weir, C. (1990). *Communicative language testing.* Prentice Hall.