A Data Science Approach to Item Analysis in Higher Education

Michael Joseph Ennis – Free University of Bozen-Bolzano, IT

Abstract

Item analysis and the reporting of test statistics are crucial to ensuring the reliability, validity, and fairness of high-stakes language exams. For practical reasons, the performances of language exams, especially traditional pen and paper exams, are often evaluated using samples of responses and ratings manually collected during exam administrations or pretesting stages of exam development. Due to known constraints, the traditional approach to data collection can be time consuming and tedious, and the resulting samples are not always sufficiently large, random, or representative for the purpose of evaluation, monitoring, and decision making. This paper presents the data science approach to item analysis and test statistics employed at a small multilingual university which administers four to six thousand computer-based language exams per year. After summarizing the rationale for a data science approach, the paper describes how the complete set of responses and ratings from all exam sessions are extracted from the online exam platform (OWL Testing Software) and then queried into Microsoft Power BI to build a data model and design interactive, automated reports that can be shared online with stakeholders. The paper describes recent and future enhancements to existing reports, in particular a novel automated approach to estimating the test-retest reliability of writing and speaking exams and concludes by reflecting on the advantages and disadvantages of a data science approach within this context.

Part of
Bonetto, E., Unterkofler, D., & Wimhurst, A. (Eds.). (2026). Teacher Researchers and Researcher Teachers: Bridging the Research–Practice Gap at University Language Centres. bu,press.
https://doi.org/10.13124/9788860462077

1. An Introduction to Educational Data Science

Data science can be described as an interdisciplinary paradigm which synthesizes principles and methods from computer science, statistics, and scientific visualization with the expressed aim of extracting knowledge from all types of digitally stored, machine-readable data and presenting that knowledge in a manner that is accessible and informative for data-driven decision-making (Cao, 2020, Hey et al., 2009). The datasets which data scientists work with are often referred to as *big data*, that is, datasets containing thousands or millions of individual observations and dozens or more quantifiable variables (e.g., Mayer-Schönberger & Cukier, 2013; Williamson, 2017). One advantage of data science is that it is not limited to the analysis of primary data—samples collected for research purposes—but can handle secondary data and real-world data collected from private and public repositories, such as organizational/institutional databases or the internet (e.g., Mayer-Schönberger & Cukier, 2013). To extract and visualize actionable insights from such data, the data scientist relies upon programing languages (most notably R and Python), artificial intelligence, computer software (such as Hadoop, SAS, and MATLAB), and domain-specific knowledge (Cao, 2020, Hey et al., 2009; Estrellado et al., 2020). Potentially, data science can generate insights in any field of inquiry or praxis, whether humanistic, social-scientific, formal-scientific, or natural-scientific (e.g., Cao, 2020; Hey et al., 2009; Mayer-Schönberger & Cukier, 2013).

In addition to fields such as business analytics, economic research, medicine, public health, and politics, data science is increasingly being applied in educational research and praxis. Two interconnected but distinguishable branches of educational data science are *learning analytics* and *educational data mining* (EDM). The former, which has evolved out of online learning since the 1990s, typically involves the analysis of structured data routinely stored in virtual learning environments (VLEs) and learning management systems (LMSs), such as those used for massive open online courses (MOOCs) (Dülger, 2020; Estrellado et al., 2020). Learning analytics may, for example, entail the measurement and comparison of learner engagement via user logs, such as views, clicks, completion times, completion status, scores, etc., to make inferences and predictions about individual learners or cohorts (Baig et al.,

2020; Estrellado et al., 2020; McFarland et al., 2021; O'Neil, 2016; Romero and Ventura, 2020; Salas-Pilco, Xiao, & Hu, 2023).

EDM, on the other hand, can extend to the automated extraction and exploration of semi-structured and unstructured data from more obscure sources such as institutional databases, institutional websites, institutional documents, and institutional social media accounts, often using machine learning algorithms (e.g., web scraping and text mining algorithms) to extract the data and detect hidden and unexpected patterns. An example would be obtaining comments and associated metadata from public posts to a university's social media page for the purpose of *natural language processing* (NLP), like *sentiment analysis*, *keyword analysis*, or *thematic analysis* (Aljawarneh, & Lara, 2021; Estrellado et al., 2020, Benelli, Desimoni, & Montecchiari, 2022; McFarland et al., 2021; Romero and Ventura, 2020).

Combined, these educational applications have been utilized to:

- track and predict student behavior and achievement;
- understand student backgrounds and needs;
- group students based on common characteristics;
- develop and improve learning tools such as adaptive learning platforms and AI tutors;
- identify recurring themes in course evaluation responses and reflective writing;
- evaluate educational programs and staff;
- inform administrative and policy decisions;
- support teacher education and reflective practice; and
- contribute to theories of teaching and learning (Aljawarneh, & Lara, 2021; Baig et al., 2020; Dülger, 2020; Estrellado et al., 2020; Kessler, 2018; Romero and Ventura, 2020; Salas-Pilco, Xiao, & Hu, 2023; Williamson, 2017).

These and many other examples demonstrate that educational data science has the potential to empower teachers (and their learners) to make informed, evidence-based decisions in complex educational contexts (Estrellado et al., 2020).

1.1 Data Science in Language Education

Data science has been applied in the specific context of language teaching and learning in similar ways as in other subjects. For example, learning analytics and data mining algorithms are central to most contemporary language learning platforms, especially adaptive ones, where big data is harnessed to personalize language instruction and monitor language acquisition (Ashrafimoghari, 2022). In the case of popular commercial language learning apps, like Duolingo, user data not only drives instructional design but also user retention and the company's business model (Orbey, 2023). The adaptive and intelligent tools produced by educational data science can make language learning more accessible and motivating, especially when used as supplements rather than replacements for structured, communicative learning (Orbey, 2023).

But the application of data science in applied linguistics and language education is also distinctively domain specific. For example, NLP algorithms have been used to measure syntactic complexity in second language writing and quantify language development (Kyle & Crossley, 2018), predict reading comprehension and processing difficulty based on syntactic and semantic features (Crossley, Kyle, & McNamara, 2017), and simulate human-like comprehension of written dialogues (Sun et al., 2019). Such applications have implications for language pedagogy, curriculum design, and the development of intelligent language learning applications. Regarding language assessment and testing, educational data science has produced automated feedback and assessment tools (Dülger, 2020; Kessler, 2018; Romero and Ventura, 2020) and has practical utility for test development such as the identification and selection of appropriately difficult reading texts used in test item writing (Kyle & Crossley, 2018).

1.2 The Risks of Educational Data Science

While data science has been recognized as a potentially transformative paradigm with profound implications for scientific discovery and human progress (e.g., Cao, 2020; Domingos, 2015), there are also potential risks associated with the acceptance of data as a new form of "capital" or data science

as a flawlessly efficient and objective data-driven approach. Overreliance on big data can negatively impact society, especially education, by amplifying inequality, reducing transparency and accountability, undermining privacy and research ethics, and entrenching systemic biases (Mayer-Schönberger & Cukier, 2013, O'Neil, 2016). While educational data science has demonstrated its immense potential to revolutionize teaching and learning, there remain many concerns, including:

- the ethics of accessing certain types of student data;
- the implicit bias of algorithms used for prediction, adaptation, and evaluation due to the limitations of their training data;
- the educational value of algorithmically optimized content, especially when behind paywalls (e.g., shallow learning, lack of human interaction, and cultural and linguistic simplification);[1]
- the fairness and equity of evaluating learners, teachers, and institutions with automated metrics and basing consequential decisions on those metrics alone, such as passing or failing, retaining or firing, and funding or defunding;[2] and
- the generalizability and quality of big data which was not collected explicitly for research (or evaluative) purposes (see Aljawarneh, & Lara, 2021; Ashrafimoghari, 2022; Baig, et al., 2020; Benelli, Desimoni, & Montecchiari, 2022; McFarland et al., 2021; O'Neil, 2016; Orbey, 2023; Romero and Ventura, 2020; Williamson, 2017).

To mitigate risks, researchers and practitioners recommended investment in infrastructure and teacher and administrator training (Aljawarneh, & Lara, 2021; Dülger, 2020; Salas-Pilco, Xiao, & Hu, 2023), interdisciplinary cooperation (Benelli, Desimoni, & Montecchiari, 2022; McFarland et al., 2021), reproducibility of data science projects (Estrellado et al., 2020), and *algorithmic*

1 For example, see Orbey's (2023) critical review of Duolingo, in which the author concludes that gamification is not the same thing as deep learning, algorithms are not the same thing as pedagogy, and knowing how to use an app is not the same thing as possessing communicative competence.

2 O'Neil (2016) provides examples where schools have been defunded or closed because algorithms analyzing standardized test scores identified these schools as "underperforming". O'Neil argues that these algorithms perpetuate educational inequality because they ignore test bias and socioeconomic disparities and thereby penalize the under-resourced communities they purport to help.

accountability, that is, transparency in how data is collected, stored, processed and used in the interest of ethics and equity (Williamson, 2017).

2. The Potential for Data Science in Item Analysis

To fully appreciate the potential of data science in item analysis, it is crucial to outline the fundamental concepts, procedures, and role of the latter in test development. In addition to applied linguistics and pedagogy, language assessment theory borrows many core concepts from the field of psychometrics (Bachman & Palmer, 2010; Coombe et al., 2012; Green, 2014; Hughes, 2002), namely the branch of psychometrics which focuses on educational and psychological assessment (Cooper, 2021; Crocker & Algina, 2006). Psychometrics as applied to language assessment concerns itself with the construction and validation of instruments employed to assess linguistic or communicative competence, especially in the context of standardized, high-stakes testing. From the perspective of *classical test theory*, one important procedure in the validation of tests is item analysis, which is the application of statistical methods to evaluate the quality of tests and test items based on their performance in *pretesting* (trialing/piloting) or testing situations. Item analysis constitutes an essential step in the *test design cycle* (e.g., the one summarized by Green and Fulcher, 2020), especially informing decisions about item revision, *retention*, or *removal*.

From a psychometric perspective, a test is an instrument that is designed to measure a *latent construct* (i.e., one that cannot be directly observed) (Cooper, 2021; Crocker & Algina, 2006), such as ability to read for gist or ability to listen for specific details. Item analysis estimates the precision of the resulting measurement and item quality, providing direct evidence relevant to reliability and contributing to broader validity and fairness claims (Cooper, 2021; Crocker & Algina, 2006). Within Classical Test Theory, core concepts informing item analysis include *difficulty*, *discrimination*, and *distractor analysis*, while *reliability*, *validity*, and *fairness* serve as overarching measurement principles supported by these and other analyses (Cooper, 2021; Crocker & Algina, 2006).

A valid item—and by extension a valid test—is one whose scores can be meaningfully interpreted as reflecting the construct it is intended to assess

(Cooper, 2021; Crocker & Algina, 2006). There are many ways of slicing up the concept of validity, but the two most frequently mentioned in the language assessment literature are *face validity* and *construct validity* (Bachman & Palmer, 2010; Brown, 2004; Coombe et al., 2012; Green, 2014; Hughes, 2002). An item has face validity when subject experts (e.g., language teachers, test item writers, and test developers) judge the item to be appropriate for testing the stated construct (Brown, 2004; Hughes, 2002). In other words, face validity rests on informed yet inherently subjective impressions. An assumption of construct validity, on the other hand, requires additional objective evidence (Brown, 2004; Cooper, 2021; Crocker & Algina, 2006; Hughes, 2002).

Further support for validity claims could come from confirming that the test is reliable. Reliability is the extent to which variability in test takers' *observed scores* reflects variability in their *true scores*. (i.e., their actual quantifiable ability related to the test construct) (Brown, 2004; Coombe et al., 2012; Cooper, 2021; Crocker & Algina, 2006; Green, 2014; Hughes, 2002). Theoretically, one could subtract each test taker's true score from their observed score and calculate the variance of these *errors*, which in statistics is commonly called *noise* (Bachman & Palmer, 2010; Cooper, 2021; Crocker & Algina, 2006; Green, 2014; Hughes, 2002). The less noise contributes to observed score variability, the more reliable the assessment instrument. The psychometrician's dilemma is, of course, that true scores remain unknown. Therefore, reliability must be estimated indirectly by measuring the consistency of observed scores (Bachman & Palmer, 2010; Cooper, 2021; Crocker & Algina, 2006; Green, 2014; Hughes, 2002).

There are several statistical methods for estimating consistency, but two common procedures used in language assessment are the *split-half* method and *Cronbach's alpha* (Bachman & Palmer, 2010; Cooper, 2021; Crocker & Algina, 2006; Hughes, 2002). Both methods estimate *internal consistency*, that is, the tendency of item responses (or right-wrong scores) to show consistent relationships across items when compared across test takers (Brown, 2004; Crocker & Algina, 2006; Hughes, 2002). The split-half method divides the test into two equal parts (e.g., all even-numbered questions versus all odd-numbered questions, or the first half of the test versus its second half), effectively treating the two halves as separate tests by calculating each test taker's score on each half. If there is strong correlation between test takers' scores on both

halves of the test, then it can be concluded that there is statistical evidence of internal consistency. Cronbach's alpha can, under the right assumptions, be understood as the equivalent to the mean consistency of all possible split-halves of the test and is therefore a more robust and often preferred estimate of reliability (Cooper, 2021; Crocker & Algina, 2006).

Importantly, a valid test is, by definition, reliable, but a reliable test is not necessarily valid (Bachman & Palmer, 2010; Brown, 2004; Cooper, 2021; Crocker & Algina, 2006; Green, 2014; Hughes, 2002). This is because a test may consistently measure an unintended construct (e.g., a supposed reading comprehension test which really tests lexical knowledge). In praxis however, if a test is grounded in a sound theoretical framework (i.e. demonstrates *content validity*), it exhibits face validity, and it provides evidence of reliability, it is often treated as being valid, particularly in low-stakes contexts (Hughes, 2002). However, validity is a multifaceted concept, and the higher the stakes of an examination, the more evidence required to support the intended interpretation and use of test scores (see Messick, 1989).[3]

Item difficulty and item discrimination are statistics which can have effects on internal consistency measures (Bachman & Palmer, 2010; Cooper, 2021; Crocker & Algina, 2006; Green, 2014). Item difficulty is simply the percentage or proportion of test takers who selected the correct response of a test item and thus estimates how easy or difficult an item is (Bachman & Palmer, 2010; Brown, 2004; Crocker & Algina, 2006; Green, 2014; Hughes, 2002). Item discrimination is the correlation between test takers' scores on a test item with their total scores on the entire test and thus estimates an item's ability to differentiate test takers with greater ability from test takers with lesser ability in the construct (Bachman & Palmer, 2010; Brown, 2004; Crocker & Algina, 2006; Green, 2014; Hughes, 2002).[4] If a test has too many easy or too many difficult items, then estimates of internal consistency reliability can be distorted. If too many test takers with more ability tend to perform poorly on

3 For example, evidence based on relationships with external criteria (e.g., *gold standard* reliability) or internal structure (e.g., factorial analysis), when appropriate, may further strengthen claims of validity but fall outside the scope of the present paper.

4 The term "discrimination" has a positive connotation in this context. It should not be conflated with terms such as "racial discrimination" or "gender discrimination". In cases where certain classes of individuals underperform in comparison to other classes on average, the test should be evaluated for potential biases.

easy items and/or if too many test takers with less ability tend to perform well on difficult items, then split-half or Cronbach's alpha coefficients can be negative in value, which is a mathematically feasible result which may suggest an issue with item writing or a false assumption about the test design (e.g., testing two unrelated constructs). Therefore, revising, replacing, or removing items with unacceptable levels of difficulty and discrimination can improve the accuracy of reliability estimates and, possibly, reliability itself (Bachman & Palmer, 2010; Cooper, 2021; Crocker & Algina, 2006).

If an item exhibits unacceptable difficulty or discrimination, *key* and *distractor analysis* may help detect the root cause (Brown, 2004; Cooper, 2021; Crocker & Algina, 2006; Hughes, 2002). The response options for multiple-choice items, for example, include the key (i.e., the correct response) and the distractors (i.e., the incorrect options). Analyzing the proportion of test takers who select each option is called key and distractor analysis, or often simply distractor analysis.[5] Distractor analysis considers the proportion of test takers who select each incorrect option. This information can quickly reveal problems with an item. Generally, all incorrect options should be selected in similar proportions. Given a sufficiently large sample, if a distractor is not selected at all or very infrequently in comparison to the other distractors, then it is not performing its job because most test takers, even less competent ones, are easily eliminating it as an option (Crocker & Algina, 2006; Hughes, 2002). If a distractor is selected more often than the correct answer, making the item appear too difficult, then this may suggest that the item in fact has two correct responses or that the correct response has been *miskeyed* (Crocker & Algina, 2006).

Fairness is arguably the most important criterion for the quality of a language test (Bachman & Palmer, 2010; Coombe et al., 2012; Council of Europe, 2003; Green, 2014; Hughes, 2002). Within *classical test theory*,[6] if a test exhibits

5 Note that the proportion of test takers who select the correct response is the item difficulty if each correct response is awarded one point and each incorrect response is awarded zero points.

6 There are other models for evaluating tests, such as *item-response theory*, which also considers test takers' respective abilities when calculating difficulty and reliability, and *generalizability theory*, which accounts for the fact that error between observed scores and true scores is not static because there are many sources of error (Crocker & Algina, 2006). But these are beyond the scope of the present paper, as these models have not yet been integrated into the item analysis report presented here.

reliability, validity, and sufficient difficulty and discrimination across items, and this data was collected administering the test to a sample that is representative of the target test-taking population, then the prerequisites for fairness exist (Cooper, 2021). However, fairness also requires examining whether test scores function equivalently across relevant subgroups, so it may also be advisable to analyze test performance by subsets of the population (e.g., gender, race, socioeconomic status, etc.) to detect potential *differential item functioning* (DIF), which may indicate discriminatory bias (Bachman & Palmer, 2010; Brown, 2004; Coombe et al., 2012; Council of Europe, 2003; Green, 2014; Hughes, 2002).

Item and test analysis are generally performed with the aid of a computer (Bachman & Palmer, 2010; Cooper, 2021; Crocker & Algina, 2006; Green, 2014). They can be done with Microsoft Excel (see Zaiontz, n.d.) or most statistical analysis software programs, such as IBM SPSS. There are also numerous commercial and open-source packages designed specifically for item analysis and other psychometrics, including SITA (Sistema per l'Item Analysis), Iteman, The Test Analysis Program (TAP), and Xcalibre. Although these options can easily handle large datasets and offer certain automation features, they generally require manual data cleaning and manual importing of each new dataset. Because the manual manipulation of data is time-consuming—especially with pen and paper exams—in university contexts, item analysis is traditionally performed periodically and/or using samples, and commonly only during the pretesting stage of exam development and not for continuous monitoring and evaluation purposes. Some scholars (e.g., Okan Bulut et al., 2024) caution that item analysis should not be fully automated, because the statistics alone are rarely sufficient for decision making. Yet when viewed as a tool to support human judgement, there is great potential for data science in item analysis. Computer-based language testing, in particular, presents numerous opportunities for data science approaches, as responses are by default machine-readable.

Although data science is not being applied regularly in item analysis in language testing praxis in localized educational contexts, its theoretical potential is being explored with increasing frequency in research. A large number of recent studies and patents on this application have originated at organizations which develop standardized language exams, like Educational

Testing Service in the United States and Cambridge English Assessment in the United Kingdom. Some theoretical applications which have been investigated include, among others:

- automated item generation (Sayin, & Gierl, 2024; Shin & Lee, 2024);
- pretesting items with AI test takers (Maeda, 2024);
- estimating item difficulty using NLP algorithms (Settles et al., 2021);
- estimating difficulty and discrimination using R packages (Estrellado et al., 2020);
- automated distractor analysis (Raina et al., 2023); and
- using AI to detect DIF which may disadvantage some test takers (Belzak et al., 2023; Liao, & Yao, 2021; Maeda & Lu, 2025).

The remainder of this paper reports on the exploration of data science for the purpose of continuous evaluation and re-evaluation of high-stakes, standardized language tests at a small university, which is a novel approach in the Italian higher educational context.

3. Language Assessment at the Free University of Bozen-Bolzano

The Free University of Bozen-Bolzano (unibz) is a trilingual university with three official languages of instruction: English, German, and Italian. As such, the university has strict language requirements for matriculation and graduation based on the Common European Framework of Reference (CEFR, Council of Europe, 2003) (see Table 1 for the current requirements). The unibz Language Centre contributes to the implementation of this language policy by supporting language learning (e.g., language courses, language advising, etc.) and by certifying the entry and exit language requirements. The language in which prospective students completed high school—or a previous university degree—can be recognized as their first language at C1 level, if that language is one of the official languages of unibz. All other entry and exit requirements must be certified by submitting a recognized international language certificate (e.g., IELTS, TestDaF, PLIDA) or by passing an in-house language proficiency examination.

The Language Centre has developed a separate in-house exam for each official language (English, German, Italian) at each target level (B1, B2, C1, according to the CEFR), that is, nine exams in total. The exams are divided into three modules: 1) Reading and Listening, 2) Writing, and 3) Speaking (see Table 2). All three modules are computer-based, using the web-based exam management system, OWL Testing Software (OWLTS, https://owlts.com/). Students complete the exams on university-provided notebook computers in a large lecture hall (called the "Aula Magna"), while prospective students complete the exams remotely on their personal devices. The first and second modules are typically offered in the morning of an exam date, while the third module is offered in the afternoon. The first module is instantly machine marked by the platform, whereas the second and third modules are marked by human raters using a rubric. Students who pass the first module immediately advance to the second module and are then invited to return in the afternoon to attempt the third module. Students must pass all three modules to pass the full exam. However, under current unibz regulations, students have 18 months to pass the second and third modules after passing the first module, and they may attempt the same exam up to three times per academic year. After 18 months have lapsed, they must repeat the first module.

Table 1 – Language requirements for BA programs

Language	Entry Requirement	ASAP[7]	Exit Requirement
First Language[8]	B2		C1
Second Language	B2		C1
Third Language	A0	B1	B2

7 Students must demonstrate a minimum of B1 proficiency in the language of instruction before registering for the final exam of any course offered in their degree program. For example, if students are expected to complete a statistics course, and that course is taught in Italian, then students can attend the lectures, but they cannot attempt the final exam for their statistics course until they have certified a minimum of B1 proficiency in Italian at the Language Centre. Thus, students are encouraged to focus on achieving B1 (and then B2) in their third language before working toward the C1 exit level for their second language.

8 We use the terms "first", "second", and "third" in this context with complete awareness of broader debates on multi- and plurilingualism. "First language" in this context does not necessarily mean "native language" or "mother tongue", but simply the first and presumably easiest language for students to certify at least the C1 level. For the vast majority of students, this is the language in which they attended high school, which typically, but not always, was their dominant home language since birth. The "second language" is the second language which students certified at the C1 level, and the "third language" is simply the other language. We use an algorithm to automatically assign these labels.

Although unibz is considered a "small" university with a "small" language center for the Italian context,[9] due to its trilingual language policy and the current exam regulations, the Language Centre administers a disproportionately large number of exams: between 4,000 and 6,000 per year. This range is inflated because many students opt to repeat a failed exam module multiple times before seeking the support of the Language Centre, especially with respect to the certification of their so called "second" language (see Ennis, 2020; Ennis et al., 2022). Due to limited human resources in the face of this daunting task, the Language Centre must improvise somewhat in its test design cycle. For example, item writing follows a strict and rigorous procedure guided by test specifications, but for several years now, the evaluation of new items has been mostly limited to trialing among internal staff, with occasional external validation. It is for this reason that a novel approach to item analysis is being explored. The approach which is emerging is one that borrows inspiration from data science to enable the continuous monitoring of item and test performance by extracting the results of all administered exams.

Table 2 – Language exam structure

Module	Constructs	Duration	Task Types	Notes
1	Reading & listening	50 minutes	multiple choice; matching; true-false-not given	
2	Writing	60 minutes	comment; text message; correspondence; essay; report	must pass module 1 first
3	Speaking	10-15 minutes	voice message; response (to question); presentation	must pass module 1 first

9 There are approximately 4100 students currently enrolled at unibz. The Language Centre consists of a director, six language specialists (two per language), two testing specialists, and a secretariat with a staff of five.

4. Toward a Data Science Approach to Item Analysis at the unibz Language Centre

As previously mentioned, the platform used to administer exams at the unibz Language Centre is called OWL Testing Software (OWLTS), which is a web-based testing management system. Within OWLTS, the Language Centre can create, administer, and manage all language exams, including placement tests and mock exams. Most importantly for the present paper, the system supports the retention of all data related to exam results, including responses, item scores, task scores, test scores, start and end times, metadata, etc.

Exam results are exported into two institutional information systems at the conclusion of each exam session, both of which were designed by the unibz ICT Department. The first, which was developed according to Language Centre specifications, is called LCIS (Language Course Information System) and is used to store, edit, access, and view data related to language courses and language exams by Language Centre staff. The second is called AIS (Academic Information System) and is used by staff across the university to store, edit, access, and view student records. The overall results of exam modules are imported into LCIS and AIS not only so that unibz staff can view exam histories for any student, but also to automate administrative processes such as limiting students' access to language courses and language exams based on their "current" and "next" CEFR levels for each language, or blocking enrollment in degree course exams until a student has attained at least B1 proficiency in the language in which those exams are conducted.

All three of these systems—OWLTS, LCIS, and AIS—were designed for institutional and administrative processes, and not specifically for data analysis, although all allow users to export data into CSV or Microsoft Excel files. Thus, a separate software package is necessary for more advanced analytics, like item analysis. When the Language Centre originally decided to build data models to monitor all aspects of the unibz language curriculum (e.g., language courses, language exams, and the satisfaction of language requirements), multiple software options were discussed, including SPSS, SAS, MATLAB, and even a custom interface based on R. The unibz ICT Department, however, recommended Microsoft Power BI as an option because it had

recently been selected as the preferred software for designing institutional reports and because licenses were therefore already available.

4.1 Basics of Power BI

Power BI (PBI) is designed for business analytics, and not for learner analytics or educational data mining. The BI, in fact, stands for "business intelligence". As such, PBI has many built-in functionalities for analyzing sales data, for example. Its functions for statistical analyses, on the other hand, are somewhat limited in comparison to alternatives. The advantage it has over other software considered in this context is that Power BI can easily automate the extraction and visualization of very large datasets (i.e., big data).

PBI has two main components: Power BI Desktop and Power BI Service. PBI Desktop is a Windows desktop application used for creating reports with four main user interfaces. The first interface is the Power Query Editor (Microsoft, 2024c), which relies on M formula language (Microsoft, n.d.-c) to *extract*, *transform*, and *load* (ETL) data (Microsoft, n.d.-d). Data is first extracted from any number of data sources, such as Excel files, websites, or institutional databases. The data can then be transformed, that is, cleaned, restructured, and combined into a more usable format in preparation for *data modeling*.[10] Finally, the data is loaded into the application. The ETL process is completed in clearly documented steps, the end result of which is an automated *query* which reapplies the same steps every time the data is loaded or refreshed with updates. For instance, if a step is applied to convert a column of numbers to be processed as text, then this transformation will also be applied to new rows of data subsequently added to the dataset, without the need for further human manipulation. The steps can be deleted, reordered, or edited as needed in the future. Automated ETL with the Query Editor is thus a time-efficient way of manipulating datasets that are expected to grow over time.

Once data is loaded, a data model can be built from the second user interface, the *model view* (Microsoft, 2024b). In PBI, data models are built according to an underlying *schema* involving *dimension tables* and *fact tables*, often queried from multiple data sources (Microsoft, n.d.-b). Dimension tables contain

10 Power BI creators and users employ the term transformation, but data scientists sometimes use the term *data wrangling.*

descriptive information, such as the three exam languages (English, German, and Italian), the three exam modules (first, second, and third), exam sessions (e.g., September 2024), or a list of test takers and their demographic information. Fact tables contain quantitative (or quantifiable) information, such as the responses submitted by all test takers (e.g., option 1, 2, or 3) and whether these responses were correct (e.g., 1 for correct, and 0 for incorrect). The data in fact tables can be connected to data in dimension tables via *keys*, or unique identifiers, like numerical or alphanumerical codes. In the model view, data analysts can use the keys to create *relationships* between data tables so that dimension tables can filter or segment the data in fact tables for improved *data analysis* and *visualization*. For example, exam responses for each task or item can be filtered or segmented according to the language of the exam, the exam module, the exam session, and/or the age of test takers. Modeling data in this way results in a *star schema* (multiple dimension tables related to a central fact table) or, for more complex datasets, a *snowflake schema* (multiple dimension tables related to each other as well as to a single or to multiple fact tables).[11] Both types of schemas improve the efficiency of computations and thereby end-user interactions with the data by eliminating or minimizing data redundancies, which reduces the size of the data model.

The third and fourth interfaces in PBI desktop are the *table view* (or data view) (Microsoft, 2025b) and the *report view* (Microsoft, 2025a), which are where data analysis and visualization take shape. The table view provides a list of all tables in the data model and displays a selected table as a spreadsheet, with similar functionalities as Excel spreadsheets for exploring the data. The report view consists of a canvas and panes which allow users to select various types of visuals (e.g., graphs, tables, and charts) and data fields to populate those visuals. Through these interfaces, DAX (Data Analysis Expressions), which is a formula expression language similar to Excel formulas (Microsoft, 2023), can be used to create *calculated columns*, *calculated tables*, and *measures* (Microsoft, n.d.-a) to fine tune the data model and enhance data analysis and visualization. For instance, raw exam data queried into PBI may not include test takers' total scores, so the user may want to create a new column or even a new table which sums the total points scored at the task or test

11 In data science, the combined ETL and data modeling processes are often called *data warehousing*.

level for each test taker. As mentioned before, PBI does not include the standard statistical capabilities for item analysis, so a user may also want to create a measure which dynamically recalculates a reliability statistic (e.g., Cronbach's alpha) based on applied filters or data segmentation. As visuals are added to the canvas, and new pages are added, an interactive report emerges. Once completed, the report can be uploaded to the PBI Service (Microsoft, 2024a) so that it can be shared online with selected end-users. The PBI Service can also be used to schedule automatic refreshes of the data models so that reports based on those models are always up to date.

4.2 Item Analysis in Power BI

The unibz Language Centre has been using PBI to generate automated reports for language curriculum monitoring since 2018. The data model pools information collected and stored in LCIS, AIS , and other sources to connect data related to the satisfaction of language requirements (e.g., students' current and next CEFR levels, language course completion, and exam results) to variables related to student demographics (e.g., gender, age, and nationality) and academic careers (e.g., degree program and cohort). This framework has also been used to design reports to monitor general exam results, including exam enrollments, pass and fail rates, and correlations between language exam scores and language course participation. In 2023, a new project was initiated to expand this data science approach to item analysis to enhance the development and monitoring of our in-house language exams.

At the time of writing, the item analysis data model and report remain in the development phase. For example, data queries are only semi-automated at present, in that data is not extracted directly from OWLTS, but instead from Excel files which have been exported from OWLTS and saved to a local server. Currently, the files must be updated manually—via copy and paste—with new data after each exam session, although transformation and loading are automated in PBI. In addition, the item analysis model has been built independently of the much larger and more complex curriculum monitoring model, so that it is not yet possible to filter and segment item data based on defined dimensions, like gender, age, or degree program. Finally, the current design of the report is purely functional and not aesthetic. Still, the automat-

ed item analysis report in its current state has proven to be decidedly useful within the test design cycle at the unibz Language Centre.

Figure 1 presents the landing page of the item analysis report as end-users see it, filtered for an example task. The data on this page refers to the first module—reading and writing—only. The report contains the main features of classical item and test analysis, including:

- item difficulty and discrimination scores;
- response rates for each item;
- the number of test takers and tests in the sample;
- the mean score (difficulty) for all tests, tasks, and items in the selection;
- the standard error of the mean score;
- the split half and Cronbach alpha coefficients; and
- mean completion times.

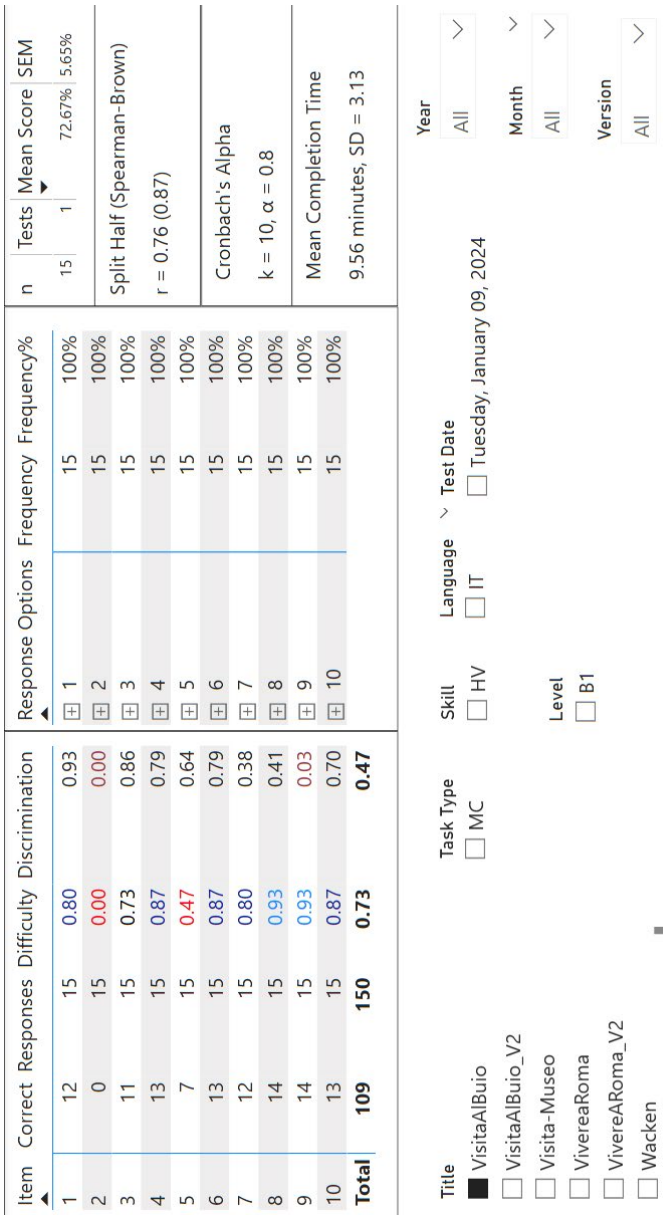
The statistics are calculated with DAX measures (see Appendix 1 for an example).

The first report page also contains multiple *slicers* used to filter the data, which are:

- the year, month, and date of the testing session;
- the language and CEFR level of the exam;
- the skill (reading or listening);[12]
- the task type;[13]
- the task title; and
- the different versions of the task.

12 The codes used for the skills are abbreviations of the German *Leseverstehen* (LV) and *Hörverstehen* (HV), which are reading and listening comprehension, respectively.

13 MC is the internal abbreviation for multiple choice.



Item	Correct	Responses	Difficulty	Discrimination
1	12	15	0.80	0.93
2	0	15	0.00	0.00
3	11	15	0.73	0.86
4	13	15	0.87	0.79
5	7	15	0.47	0.64
6	13	15	0.87	0.79
7	12	15	0.80	0.38
8	14	15	0.93	0.41
9	14	15	0.93	0.03
10	13	15	0.87	0.70
Total	**109**	**150**	**0.73**	**0.47**

Response Options	Frequency	Frequency%
1	15	100%
2	15	100%
3	15	100%
4	15	100%
5	15	100%
6	15	100%
7	15	100%
8	15	100%
9	15	100%
10	15	100%

n	Tests	Mean Score	SEM
15	1	72.67%	5.65%

Split Half (Spearman-Brown)

r = 0.76 (0.87)

Cronbach's Alpha

k = 10, α = 0.8

Mean Completion Time

9.56 minutes, SD = 3.13

Title
- VisitaAlBuio
- VisitaAlBuio_V2
- Visita-Museo
- VivereaRoma
- VivereARoma_V2
- Wacken

Task Type
- MC

Skill
- HV

Level
- B1

Language
- IT

Test Date
- Tuesday, January 09, 2024

Year: All

Month: All

Version: All

Fig. 1 – Item analysis report page 1 (reading and writing module)

The slicers allow the user to compare the statistics for a task to the averages for all tasks of the same skill, level, and/or type, as well as to explore changes over time and across different versions of the same task. Figure 1, for example, has been filtered to display the data for a specific B1, listening comprehension, multiple-choice task which was administered only once to 15 test takers. This example was chosen to demonstrate the functionalities and potential practical applications of the report.

Although the reliability measures (split half and Cronbach's alpha) and the overall difficulty of this example are within the respective "good" ranges, there are some red flags with this task. Two items (2 and 5) appear to be too difficult, while two other items (8 and 9) may be too easy. Due to the very small sample size, these results should be interpreted cautiously. But the difficulty score of 0 for the second item, even with only 15 responses, necessitates further investigation and reflection. This is where the report's capabilities for basic distractor analysis can be useful. By drilling down on the responses, as done in Figure 2, the user can visualize the selection rates for each response option. The fact that 66% of test takers selected option A and 33% selected option B, while none selected the correct option C, suggests the possibility of a miskey, that is, the wrong answer set as the correct answer. The distractor data for item 5, meanwhile, suggests that its distractors may be performing too well, meaning that they may be plausibly correct answers in need of rewording. In this particular case, the task was revised with the aid of this information and the second version of the task is thus far performing much better, as shown in Figure 3, although the revised task may be slightly too easy. However, this determination will not be made until at least 100 responses have been collected and other factors considered.

Item	Correct	Responses	Difficulty	Discrimination	Response Options	Frequency	Frequency%
1	12	15	0.80	0.93	⊞ 1	15	100%
2	0	15	0.00	0.00	⊟ 2		
3	11	15	0.73	0.86	A	10	67%
4	13	15	0.87	0.79	B	5	33%
5	7	15	0.47	0.64	⊞ 3	15	100%
6	13	15	0.87	0.79	⊞ 4	15	100%
7	12	15	0.80	0.38	⊟ 5		
8	14	15	0.93	0.41	A	7	47%
9	14	15	0.93	0.03	B	4	27%
10	13	15	0.87	0.70	C	4	27%
Total	**109**	**150**	**0.73**	**0.47**	⊞ 6	15	100%
					⊞ 7	15	100%

n	Tests	Mean Score	SEM
15	1	72.67%	5.65%

Split Half (Spearman-Brown)

r = 0.76 (0.87)

Cronbach's Alpha

k = 10, α = 0.8

Mean Completion Time

9.56 minutes, SD = 3.13

Title
- VisitaAlBuio
- VisitaAlBuio_V2
- Visita-Museo
- VivereaRoma
- VivereARoma_V2
- Wacken

Task Type
- MC

Skill
- HV

Level
- B1

Language
- IT

Test Date
- Tuesday, January 09, 2024

Year: All

Month: All

Version: All

Fig. 2 – Basic distractor analysis

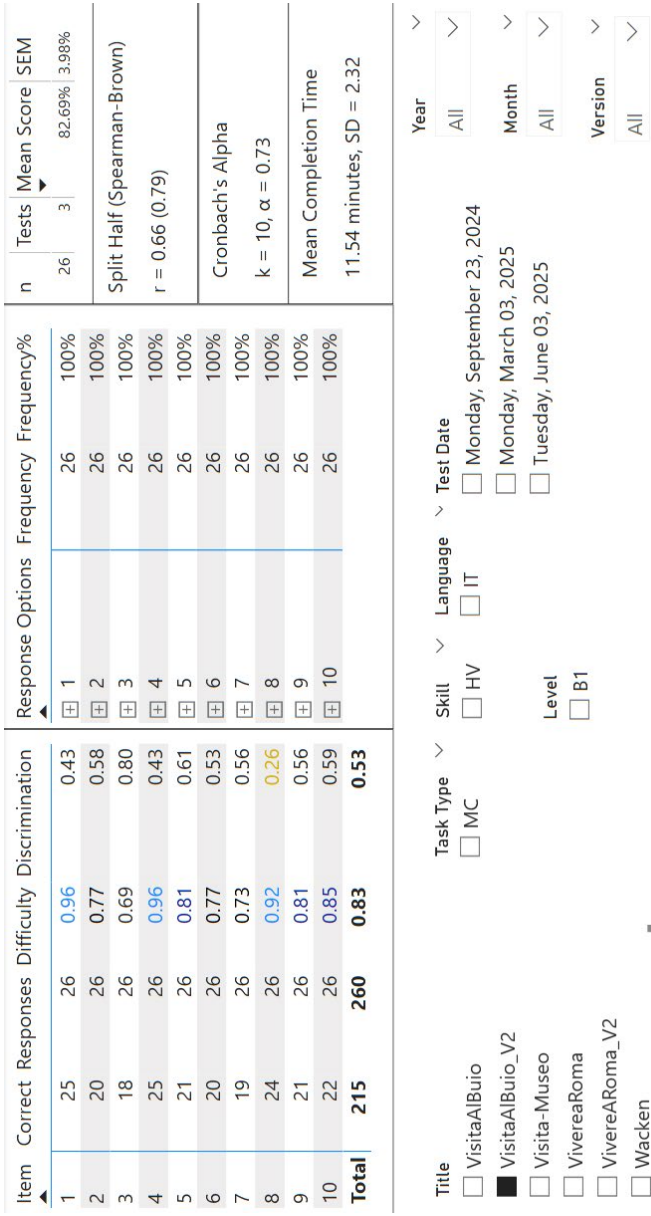


Fig. 3 – Initial data for the second (revised) version of the same task

As demonstrated by this simple example, having continuous access to data on the performance of all tasks and items has many advantages for exam development. First, it enables the Language Centre to quickly identify errors made when creating exams in OWLTS, including misspelled titles, incorrect labels, miskeyed responses, etc. The report has also been relied upon to revise tasks and items and to monitor the performance of revised tasks and items in comparison to previous versions. Finally, the underlying database, which contains detailed item-by-item data for every test taker, is now used to automate the provision of detailed feedback to test takers who submit an online request form, thereby reducing the time spent on an otherwise cumbersome administrative task by many hours each exam session.

4.3 Future Plans

The current data model contains 239,904 item-level responses, submitted by 6964 test takers, who completed 622 reading and listening tasks on 209 exams administered over 20 exam sessions since February 2023. As the data model continues to grow with each new exam session there is potential for many new developments and applications. For example, one near-term goal is to enhance the report with elements of *item response theory*, which is a statistically more complex model than classical test theory because it accounts for the interaction between test takers' abilities and item difficulties. A long-term goal is to integrate the item analysis model into the language curriculum monitoring model in order to connect item performance data with the many dimensions found in the larger model. Gradually, the automated item analysis report could be used for more advanced administrative procedures, such as selecting the best-fit tasks for each exam session. The writing and speaking modules have also produced a large corpus of written and oral responses, which in the future will be analyzed with corpus linguistics methods, and potentially NLP, to explore features of test taker language. But another project which has already been initiated is a novel approach to estimating the reliability of the productive skills modules.

As mentioned above, the writing and speaking modules of the unibz language exams are graded by humans. Each response is graded by two members of a nominated exam commission who use a rubric provided by the Lan-

guage Centre. The textbook methods for monitoring the reliability of exams marked in this manner are *interrater reliability* and *intrarater reliability* (Bachman & Palmer, 2010; Brown, 2004; Coombe et al., 2004; Crocker & Algina, 2006; Cooper, 2021). Interrater reliability refers to the consistency in scores between two or more raters, whereas intrarater reliability refers to the consistency in scores given by the same rater at two different points in time. Both are estimated using correlation coefficients, of which multiple options are mentioned in the literature, each with advantages and disadvantages (Bachman & Palmer, 2010; Coombe et al., 2004; Crocker & Algina, 2006; Cooper, 2021).

The examiners who mark the unibz language exams are required to attend mandatory training sessions twice per year. In addition to other topics, these sessions review the concept of reliability and include practical exercises which help the examiners reflect on the consistency of their ratings, such as asking participants to re-rate a sample of responses they had rated during the previous academic year. In addition, at the beginning of each exam session raters begin their work by calibrating their scores with their co-raters. However, for many practical reasons, insufficient data is collected to apply the same data science approach to interrater and intrarater reliability monitoring at the unibz Language Centre. Most importantly, only the final consensus scores for each response are saved in OWLTS, which prevents an automatic calculation of an interrater reliability coefficient. Due to the number of exams which must be marked within very tight deadlines, it would also be an inefficient allocation of resources to require examiners to re-rate all exams at a later date for the purpose of intrarater reliability estimates. Therefore, a different method must be explored.

Another method for estimating the reliability of exam scores is the *test-retest* method (Bachman & Palmer, 2010; Crocker & Algina, 2006; Cooper, 2021). Rather than having two raters rate the same response or having one rater re-rate the same response, the test-retest method prescribes administering the same test to the same test takers on two different occasions. In many contexts, this method is impractical, especially when applied to a high-stakes, standardized exam which requires significant time, space, and human resources. However, one of the biggest challenges at unibz also presents a unique opportunity to apply a proxy for test-retest reliability without the necessity of expending additional resources. Since many students reattempt the second

and third modules, potentially multiple times, before eventually passing the full exam, the data model already contains an ever-growing sample of paired test and retest scores. Most interesting, each test and retest response was produced under nearly identical testing conditions. The only caveat is that students are likely to receive a different task prompt each time they re-sit the exam, and their responses are likely marked by a different team of examiners each time. But, ideally, the objective of test development is to produce exams that exhibit acceptable levels of reliability in spite of variations in task or raters. While this approach does not meet the strict assumptions of classical test–retest reliability, it provides an ecologically valid estimate of score stability in this operational context.

The data model has already undergone additional transformations in PBI to explore the application of a test-retest approach. The first step was the creation of a calculated table containing all instances in which a test taker reattempted the written or speaking module within 60 days of the first attempt. Then additional measures and visuals were used to create a template for an automated test-retest report. Figure 4 is a screenshot of the first page of this new report, which displays:

- a scatterplot of all test and retest scores to visualize the correlation between the two in comparison to the 45-degree perfect *agreement line* (indicated in the graph where pink shading meets green shading);
- Pearson's correlation coefficient for the relationship between test and retest scores;
- the appropriate *intraclass correlation* coefficient, which unlike Pearson's r, increases toward the maximum of 1 as the relationship between scores approaches perfect agreement (Bachman & Palmer, 2010; Crocker & Algina, 2006);
- basic slicers to filter for level, language, skill, number of days between test and retest, and the first test date.

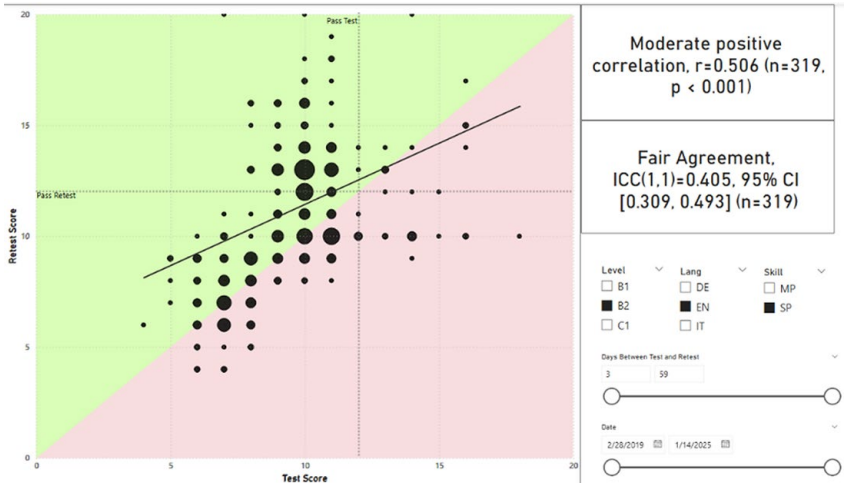


Fig. 4 – Test-retest scores for the B2 English writing exam

Figure 4 shows the report page filtered for the B2 English writing exam. Since February 2019, there have been 319 instances of test takers repeating this exam within sixty days of the previous attempt. The scatter plot and the coefficients suggest that the test and retest scores exhibit moderate correlation and fair agreement. However, sixty days is a significant amount of time in the lives of university students. While language attrition is perhaps not a major concern within that timeframe, test takers who had practiced more for their first attempt may have forgotten some test-taking skills over that period. Other students, who went abroad or engaged in intensive language learning—a modality emphasized at the unibz Language Centre, may have in fact improved their writing or speaking skills during that short time.

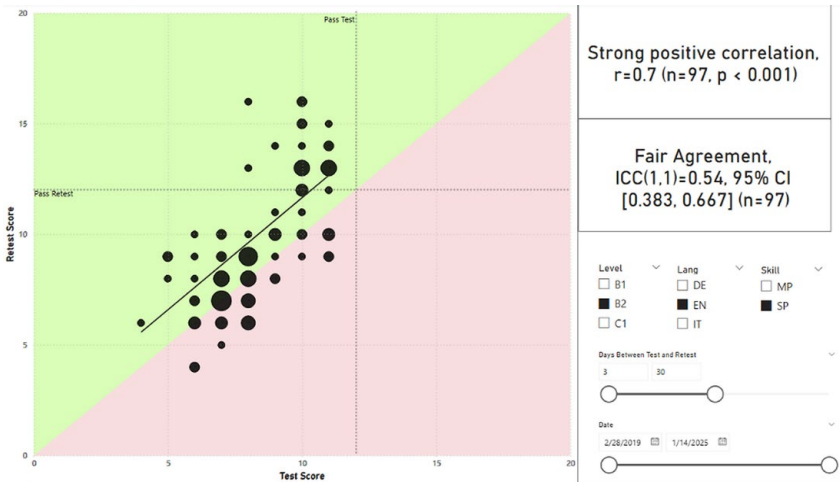


Fig. 5 – Test-retest scores filtered for thirty or fewer days between test and retest

Figure 5 shows the same report page filtered for thirty or fewer days between test and retest in an attempt to eliminate some of the "noise", that is, some of the extraneous variables which are leading to measurement error. Although the sample size is much smaller (n=97) , the correlation is much stronger and the intraclass correlation (which considers agreement), is somewhat stronger. Notably, the regression line now runs almost perfectly parallel to the agreement line. If this relationship holds as the data model grows, then it could be interpreted positively. The primary objective of the Language Centre is to foster language learning and the satisfaction of the unibz exit requirements; high-stakes testing is a secondary objective which serves to confirm that we (and the students) are meeting objectives. If our students tend to perform slightly better on a second attempt of the productive skills modules after one month, and there is still a strong correlation between their first attempt scores and their second attempt scores, despite all the "noise" which unavoidably occurs in higher education, then this could be interpreted as evidence that both the primary and the secondary objectives are being met. Moreover, this initial exploration of a novel approach to test-retest reliability demonstrates potential in this context.

5. Conclusion

The key advantage of this data science approach is that it transforms item analysis from a periodic or one-off task—typically conducted during pretesting or validation—into a continuous, scalable process. Traditional item analysis relies on samples to generalize results. In contrast, this solution enables semi-automated reporting based on all available exam data. This system can provide stakeholders—test developers, item writers, test raters, etc.—with ongoing access to dynamic, interactive reports that remain up to date. The reports can support a variety of functions, including item writing and revision, quality control, and administrative tasks such as generating automatic exam feedback.

To-date, the report mitigates some of the risks associated with data science. Since the data model includes all test results and relies upon established psychometrics, generalizability is not doubted. The fairness and equity of the approach are not a major concern, either, as the aim of item analysis is not to evaluate test takers, but to judge the effectiveness of items and tests to ensure that they are fair to all test takers, and there are no plans to use the reports to evaluate the performance of item writers or exam raters. Once the item analysis data model is connected to dimension tables containing students' demographic information, it will also be possible to detect potential biases which might disadvantage some test takers. Finally, data science ethics and accountability are partially assured by the EU's General Data Protection Regulation (GDPR), which unibz has adopted and applied in full, and partially by the fact that the shared reports are anonymized, aggregated, and never reveal personal data.

However, this approach also comes with limitations. Power BI, while powerful for visualization and reporting, is not a statistical analysis tool specifically designed for item analysis. Leveraging its full potential for this purpose requires intermediate to advanced understanding of item analysis, psychometrics, and coding languages, including DAX, M, and, optionally, R or Python. That necessitates constant training of staff and/or interdisciplinary cooperation. Additionally, maintaining the system demands regular and sometimes intensive review of the steps applied to data queries to ensure reliability and accuracy in reporting.

References

Aljawarneh, S., & Lara, J. A. (2021). Data science for analyzing and improving educational processes. *Journal of Computing in Higher Education*, *33*(3), 545–550. https://doi.org/10.1007/s12528-021-09285-z

Ashrafimoghari, V. (2022). Big data and education: Using big data analytics in language learning. *arXiv*. https://arxiv.org/abs/2207.10572

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: A state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education*, *17*, Article 44. https://doi.org/10.1186/s41239-020-00212-2

Belzak, W. C. M., Naismith, B., & Burstein, J. (2023). Ensuring fairness of human- and AI-generated test items. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Communications in computer and information science: Vol. 1831. Artificial intelligence in education* (pp. 701-707). Springer. https://doi.org/10.1007/978-3-031-36336-8_108

Benelli, G., Desimoni, M., & Montecchiari, A. (2022). Data science and machine learning in education. *arXiv*. https://arxiv.org/abs/2207.09060

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.

Cao, L. (2020). Data science: A comprehensive overview. A*CM Computing Surveys (CSUR), 50*(3), 1-42. https://doi.org/10.1145/3076253

Coombe, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (2012). *The Cambridge guide to second language assessment*. CUP.

Cooper, C. (2021). *An introduction to psychometrics and psychological assessment* (2nd ed.). Routledge.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, *54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Cengage Learning.

Council of Europe. (2003). *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Dülger, E. (2020). Big data technology in today's education systems: Learning analytics. *European Journal of Science and Technology, special issue 2020*, 353-361. https://doi.org/10.31590/ejosat.824182

Ennis, M. J. (2020). Motivating Italian university students by meeting their needs: The introduction of a business English track at the Free University of Bozen-Bolzano Language Centre. In E. Bonetto, M. J. Ennis, & D. Unterkofler (Eds.), *Teaching languages for specific and academic purposes in higher education*, 243-264. bu,press. https://doi.org/10.13124/9788860461551_14

Ennis, M. J., Barchi, K. A., Merello Astigarraga, A., & Wimhurst, A. (2022). A pilot course with project-based learning in an intensive English program. *Language Learning in Higher Education*, *11*(2), 57-85. https://doi.org/10.1515/cercles-2022-2047

Estrellado, R. A., Freer, J., Rosenberg, J. M., & Velásquez, I. C. (2020). *Data science in education using R* (2nd ed.). Routledge.

Green, A. (2014). *Exploring language assessment and testing: Language in action. Routledge*. https://doi.org/10.4324/9781003105794

Green, A., & Fulcher, G. (2020). Test design cycle. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 69-77). Routledge.

Hey, T., Tansley, S., Tolle, K., & Gray, J. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Hughes, A. (2002). T*esting for language teachers*. CUP. https://doi.org/10.1017/CBO9780511732980

Kessler, G. (2018). Technology and the future of language teaching. *Foreign Language Annals*, *51*(1), 205–218. https://doi.org/10.1111/flan.12318

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in General English Proficiency Test-Kids listening. *Frontiers in Psychology*, *12*, 1-9. https://doi.org/10.3389/fpsyg.2021.767244

Maeda, H. (2024). Field-testing multiple-choice questions with AI examinees: English Grammar items. E*ducational and Psychological Measuremen*t, *84*(2), 345–362. https://doi.org/10.1177/00131644241281053

McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open*, *7*, 1–19. https://doi.org/10.1177/23328584211052055

Maeda, H., & Lu, Y. (2025). *Finding words associated with DIF: Predicting differential item functioning using LLMs and explainable AI*. arXiv. https://arxiv.org/abs/2502.07017

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Microsoft. (n.d.-a). *Create measures for data analysis in Power BI Desktop.* Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-measures

Microsoft. (n.d.-b). *Model data with Power BI.* Microsoft Learn. https://learn.microsoft.com/en-us/training/paths/model-data-power-bi/

Microsoft. (n.d.-c). P*ower Query M formula language*. Microsoft Learn. https://learn.microsoft.com/en-us/powerquery-m/

Microsoft. (n.d.-d). *Clean data in Power BI.* Microsoft Learn. https://learn.microsoft.com/en-us/training/modules/clean-data-power-bi/

Microsoft. (2023, October 20). *DAX overview.* Microsoft Learn. https://learn.microsoft.com/en-us/dax/dax-overview

Microsoft. (2024a, August 7). *What is the Power BI service?* Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-service-overview

Microsoft. (2024b, September 3). *Work with Modeling view in Power BI Desktop*. Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-modeling-view

Microsoft. (2024c, September 4). *Query overview in Power BI Desktop*. Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-query-overview

Microsoft. (2025a, February 28). *Work with Report view in Power BI Desktop*. Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/create-reports/desktop-report-view

Microsoft. (2025b, April 2). *Work with Table view in Power BI Desktop*. Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/connect-data/desktop-data-view

Bulut, O., Beiting Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim Erbasli, S. N., Liu, X., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). Application of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Journal of Educational and Behavioral Statistics*, *49*(5), 715-722. https://doi.org/10.3102/10769986241248771

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

Orbey, E. (2023, April 24). How much can Duolingo teach us? *The New Yorker*. https://www.newyorker.com/magazine/2023/04/24/how-much-can-duolingo-teach-us

Raina, V., Liusie, A., & Gales, M. (2023). Assessing distractors in multiple-choice tests. *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems* (pp. 12-22). Indonesia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eval4nlp-1.2

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), 1-21. https://doi.org/10.1002/widm.1355

Salas-Pilco, S. Z., Xiao, K., & Hu, X. (2023). Artificial intelligence and learning analytics in teacher education: A systematic review. *Education Sciences, 13*(2), 123. https://doi.org/10.3390/educsci13020123

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. E*ducational Measurement: Issues and Practice*, *43*, 5-18. https://doi.org/10.1111/emip.12590

Settles, B., LaFlair, G. T., & Hagiwara, M. (2021). Machine learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, *9*, 1306–1323. https://doi.org/10.1162/tacl_a_00310

Shin, D., & Lee, J. H. (2024). AI-powered automated item generation for language testing. *ELT Journal*, *78*(4), 446–452. https://doi.org/10.1093/elt/ccae016

Sun, K., Yu, D., Chen, Y., & Cardie, C. (2019). DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, *7*, 217–231. https://doi.org/10.1162/tacl_a_00264

Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*. Sage. https://doi.org/10.4135/9781529714920

Zaiontz, C. (n.d.). *Reliability. Real Statistics Using Excel*. https://real-statistics.com/reliability/

Appendix 1.
Three Measures to Calculate Split-Half Coefficient

```
Odd Item Scores = CALCULATE(SUM('All Responses'[Item Points Scored]),
       FILTER('All Responses',
              'All Responses'[Responses] <> 0
              && ISODD('All Responses'[Item Order])))


       Even Item Scores = CALCULATE(SUM('All Responses'[Item Points Scored]),
              FILTER('All Responses',
                     'All Responses'[Responses] <> 0
                     && ISEVEN('All Responses'[Item Order])))


       Split Half (Spearman-Brown) =

       VAR Correlation_Table =
           FILTER (
               ADDCOLUMNS(
                   DISTINCT(SELECTCOLUMNS( FILTER('All Responses',
                   'All Responses'[Responses] <> 0),
                   "ExamineeID", 'All Responses'[Examinee ID])),
                   "Value_X", [Odd Item Scores],
                   "Value_Y", [Even Item Scores]
               ),
               AND (
                   NOT ( ISBLANK ( [Value_X] ) ),
                   NOT ( ISBLANK ( [Value_Y] ) )
               ))

       VAR Count_Items =
           COUNTROWS ( Correlation_Table )
       VAR Sum_X =
           SUMX ( Correlation_Table, [Value_X] )
       VAR Sum_X2 =
           SUMX ( Correlation_Table, [Value_X] ^ 2 )
       VAR Sum_Y =
           SUMX ( Correlation_Table, [Value_Y] )
       VAR Sum_Y2 =
           SUMX ( Correlation_Table, [Value_Y] ^ 2 )
       VAR Sum_XY =
           SUMX ( Correlation_Table, [Value_X] * [Value_Y] )
       VAR Pearson_Numerator =
           Count_Items * Sum_XY - Sum_X * Sum_Y
       VAR Pearson_Denominator_X =
           Count_Items * Sum_X2 - Sum_X ^ 2
       VAR Pearson_Denominator_Y =
           Count_Items * Sum_Y2 - Sum_Y ^ 2
       VAR Pearson_Denominator =
           SQRT ( Pearson_Denominator_X * Pearson_Denominator_Y )

       VAR CORR = IF(ISBLANK(DIVIDE ( Pearson_Numerator, Pearson_Denominator )), 0, DIVIDE (
       Pearson_Numerator, Pearson_Denominator ))
```