International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 115–130
https://doi.org/10.21449/ijate.1321061
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A data pipeline for e-large-scale assessments: Better automation, quality assurance, and efficiency

Ryan Schwarz[1], **H. Cigdem Bulut**[2*], **Charles Anifowose**[3]

[1]Vretta Inc, Toronto, ON Canada
[2]Northern Alberta Institute of Technology, Education Insights, Data & Research, Edmonton, AB Canada
[3]Vretta Inc, Toronto, ON Canada

ARTICLE HISTORY

Received: June 30, 2023
Revised: Oct. 29, 2023
Accepted: Nov. 20, 2023

Keywords:
Data pipelines,
Psychometric analysis,
Large-scale assessments,
Data validation,
Reporting.

Abstract: The increasing volume of large-scale assessment data poses a challenge for testing organizations to manage data and conduct psychometric analysis efficiently. Traditional psychometric software presents barriers, such as a lack of functionality for managing data and conducting various standard psychometric analyses efficiently. These challenges have resulted in high costs to achieve the desired research and analysis outcomes. To address these challenges, we have designed and implemented a modernized data pipeline that allows psychometricians and statisticians to efficiently manage the data, conduct psychometric analysis, generate technical reports, and perform quality assurance to validate the required outputs. This modernized pipeline has proven to scale with large databases, decrease human error by reducing manual processes, efficiently make complex workloads repeatable, ensure high quality of the outputs, and reduce overall costs of psychometric analysis of large-scale assessment data. This paper aims to provide information to support the modernization of the current psychometric analysis practices. We shared details on the workflow design and functionalities of our modernized data pipeline, which provide a universal interface to large-scale assessments. The methods for developing non-technical and user-friendly interfaces will also be discussed.

1. INTRODUCTION

The field of education is significantly influenced by the impact of testing, as evidenced by the widespread adoption of national and provincial assessment levels by various countries, alongside their active participation in international large-scale assessments. National assessment programs mostly aim to understand how well students perform in terms of curriculum expectations and standards, as well as to promote performance accountability (Volante & Ben Jaafar, 2008). On the other hand, international assessments allow countries to compare across education systems or to identify their relative strengths and weaknesses based on student performance (Addey & Sellar, 2018). Despite their distinct purposes, both national and international assessments have emerged as crucial tools for enhancing educational systems (Kamens & McNeely, 2010).

*CONTACT: H. Cigdem Bulut haticeb@nait.ca Northern Alberta Institute of Technology, Education Insights, Data & Research, Edmonton, AB Canada

e-ISSN: 2148-7456 /© IJATE 2023

Both national and international measurement practices have changed significantly over the past two decades. Pushes towards modernization have been supported by recent advances in both online and offline technologies applicable to the education industry. Accordingly, assessment design, delivery, scoring, and reporting methods have evolved significantly (Zenisky & Sireci, 2002). Since the first decade of the twenty-first century, numerous large-scale tests have switched from paper to computer-based administration (i.e., online tests and online assessments), becoming the standard in modernized educational programs (Moncaleano & Russell, 2018). Online assessments have been adopted more rapidly due to increased access to information and communication technologies in classrooms, technological advancements in testing, and methodological improvements in psychometrics that enable efficient, personalized assessments (Moncaleano & Russell, 2018). Moreover, the recent safety and health concerns brought on by the pandemic (Lynch, 2022) have further prompted educational institutions to embrace online assessments, ensuring both test security and the well-being of students.

Online assessments have provided great advantages such as increasing test efficiency, enabling faster and more efficient scoring and reporting, as well as improving the standardization of assessments, and enhancing test security (Wise, 2018). The modernization of assessments has improved the efficiency of scoring not only for selected-response items but also for open-response items (Liu et al., 2014; Sung et al., 2017). Utilizing these advancements in assessments has led to a decrease in time, labour, and financial costs in scoring item responses (Moncaleano & Russell, 2018).

The adoption of computer-based administration of assessments has also led to the development of various new item types referred to as technology-enhanced items (TEI) (Scalise & Gifford, 2006; Bryant, 2019). These items allow educational practitioners to enhance the extent to which test tasks reflect the knowledge, skills, and abilities of interest and to be more flexible (Scalise & Gifford, 2006; Russell, 2019). These are especially useful as it can be difficult to measure complex and high-level capabilities with traditional paper-and-pencil assessments (Zenisky & Sireci, 2002).

Inevitably, online assessments and overall modernization have brought a number of challenges to educational organizations and testing companies. The complex designs of the assessments, scoring various types of items, and ensuring the validity, reliability, and security of the assessment results necessitate meticulous planning and execution in each step of the administration. The increasing volume of large-scale assessment data also challenges organizations to effectively manage, score, and analyze data (Rutkowski et al., 2010). The difficulties begin with data storage and extend all the way to sharing/transferring results. Furthermore, feeding the sheer size of large-scale assessment data for analysis makes it difficult to proceed timely and efficiently.

To address these challenges, we have designed and implemented a modernized data pipeline that allows psychometricians and statisticians to efficiently manage the data, conduct psychometric analysis, generate technical reports, and perform quality assurance to validate the required outputs. A data pipeline itself is a series of data processing steps that begins with extracting raw data sets, processing the information, and managing that data in a systematic way, and then generating outputs at the end (Skiena, 2017). In education, data pipelines are utilized in order to develop early warning systems, predict student performance, and in data modeling for educational stakeholders (Ansari et al., 2017: Bertolini et al., 2021; Bertolini et al., 2022; Schleiss et al., 2022). As of the time of this paper, to the best of the authors' knowledge, no publicly reported project has focused on the development of a comprehensive psychometric data pipeline for large-scale educational assessments. Our work seeks to address this gap by presenting a meticulously designed and well-documented pipeline solution that caters to the specific needs of this critical domain.

The data pipeline proposed in this paper offers a fully automated, end-to-end, configurable, and customizable application, delivering psychometric analysis and data quality verification to stakeholders. It provides the preparation of assessment data for psychometric analysis based on classical test theory (CTT) and item response theory (IRT), producing CTT and IRT reports. It has proven to scale with large databases, decrease human error by reducing manual processes, efficiently make complex workloads repeatable, ensure high quality of the outputs, and reduce overall costs of psychometric analysis of large-scale assessment data. The customizable and dynamic nature of the pipeline enables the standard analysis workflow to take place in a significantly reduced time as compared to traditional practices. Verification reports are also generated, providing quality assurance and flagging errors or warnings that are brought to the immediate attention of psychometricians and statisticians. Lastly, the pipeline empowers stakeholders by offering them an interface to independently execute the entire administration process. This interface enables stakeholders to navigate through the necessary steps and perform various tasks within the pipeline without requiring extensive technical expertise. By providing this capability, stakeholders gain greater control and autonomy over the administration process, facilitating efficient and independent management of reporting requirements.

In summary, our approach offers a valuable solution for researchers and practitioners seeking versatility, reproducibility, and rigorous documentation large-scale assessment data needs. It addresses a crucial need in operational settings where manual, fragmented processes are prevalent. Our data pipeline efficiently handles diverse large-scale assessments, producing detailed analyses, psychometric reports, verification reports, and scorecards within 40-50 minutes, streamlining the entire workflow.

1.1. Psychometric Analysis

Psychometric analysis can be considered one of the most technical aspects of assessments as it requires expertise and training in educational statistics and measurement, intensive and collaborative work with subject matter experts, and the ability to comprehend and reflect educational policies in assessments. The primary measurement frameworks for psychometric analysis are CTT and IRT (Lord & Novick, 1968; Embretson & Reise, 2000). These two frameworks differ significantly in terms of complexity, assumptions, and measurement precision (Hambleton et al., 1991). In CTT, all items make an equal contribution to student scores, and item and test-taker statistics are sample-dependent (Embretson & Reise, 2000; Reise et al., 2005). By contrast, IRT analysis estimates the probability of answering an item correctly by considering student latent abilities and item parameters (Hambleton et al., 1991; Embretson & Reise, 2000; Reise et al., 2005). Therefore, the resulting item and person statistics are sample-independent, especially in non-Rasch models (Hambleton et al., 1991; Embretson & Reise, 2000; Reise et al., 2005). An IRT model estimates abilities by utilizing the pattern of item responses, whereas CTT ignores these patterns. Therefore, measurement precision becomes higher in IRT models (Hambleton et al., 1991; Embretson & Reise, 2000; Zenisky & Sireci, 2002). Although CTT provides important information to evaluate and improve the items and tests, it falls short of meeting the needs of modernized assessments in many aspects (see for further discussion, Embretson & Reise, 2000).

With the modernization of assessments, methodological changes were made in the design of the assessments and item scoring. As larger and more detailed datasets allow for more complex psychometric analysis, IRT-based analysis has been commonly used in large-scale assessments and fulfills the criteria of large-scale assessments in terms of validity and fairness (Oranje & Kolstad, 2019; Camara & Harris, 2020). As tailoring administered items to each individual produces greater measurement precision (Hambleton et al., 1991; Embretson & Reise, 2000; Zenisky & Sireci, 2002), IRT-based assessments can yield more robust results owing to the

invariance assumptions inherent in IRT, as compared to assessments based on CTT. These invariance assumptions allow for a more precise understanding of the latent traits being measured. As item and person statistics are on the same scale in IRT models, IRT provides more flexibility to testing organizations in many steps, such as adaptive testing, form building, and the expansion and maintenance of item pools (Hambleton et al., 1991). Furthermore, considering test fairness and security, educational organizations and testing companies tend to generate IRT-based test forms (Oranje & Kolstad, 2019).

1.2. Psychometric Software and Programming Languages

As psychometric methodology increases in complexity, software programs must evolve to meet the changing criteria and demands stemming from educational policies, curriculum, and testing specifications. Many new tools have been built to better design assessments, as well as understand and analyze assessment data. Table 1 shows the most commonly used psychometric software and programming languages in testing companies and educational institutions.

Table 1. *Most common psychometric software and programming languages used.*

Software	Functionality	Open-source
BILOG, MULTILOG PARSCALE	IRT applications (calibration, equating, linking)	No
WINSTEPS, BIGSTEPS	Item calibration based on Rasch Measurement and Rasch Analysis	No
IRTPRO, flexMIRT	Item calibration using IRT	No
SAS	Item calibration and test scoring using IRT	No
Mplus	Item calibration using IRT, Structural Equation Modelling	No
R	IRT applications (calibration, equating, linking, form building, CAT applications) MIRT (and unidimensional mirt), GRM, CDM, SEM, SEM, DIF, EDM, Confirmatory and Exploratory Factor Analysis Automated Test Assembly	Yes
Python	Item calibration using IRT, MIRT, GRM, CAT, CDM, SEM, G-DINA	Yes
Julia	Structural Equation Modelling Automated Test Assembly Item calibration	Yes

As shown in Table 1, it is possible to conduct various psychometric analyses with different software or programming languages. However, not all of them are able to perform analyses based on different measurement frameworks, including CTT, IRT, generalizability theory, and Rasch measurement theories. Nor can they conduct every application of IRT, such as calibration, equating, multigroup analysis, and explanatory modeling.

The primary reason that R is currently at an advantage is due to its orientation towards data and statistical analysis (Desjardins & Bulut, 2018). Psychometric and statistical-oriented packages

are typically built off of academic research and provide a reference to associated documentation in the CRAN (Comprehensive R Archive Network) library or a peer-reviewed paper. Furthermore, R provides the most versatility in terms of measurement frameworks and IRT applications thanks to the numerous packages available (Schumacker, 2019). The R programming language has also grown in popularity in the field of educational measurement (Desjardins & Bulut, 2018). One possible reason for this is that R is free/libre software and therefore incurs no costs for its use (R Core Team, 2022). The trade-off with free and open-source software is the loss of technical support from purchasing licensed applications but gaining a great amount of customizability. Additionally, these applications are fairly rigid in how they require data to be input, whereas R can be customized at the ground level to data models.

Some software packages such as BILOG, MULTILOG, and PARSCALE (du Toit, 2003: Muraki & Bock, 2003; Thissen et al., 2003) necessitate a specific format for the input data for which users need to follow a guideline (Croudace et al., 2005). As a result, traditional psychometric software presents barriers, such as a lack of functionality for managing data. When dealing with large amounts of assessment data, it is possible to run into memory issues even in commonly used data management software such as Excel with a maximum limit of 4GB (Microsoft Corporation, 2018) or IBM SPSS Statistics (IBM, 2020). Secondly, none of these software packages provide the ability to perform all item- and test-level analyses required by modernized large-scale assessments (Rupp, 2003). This means that the data preparation process often requires the use of distinct software programs, each serving a specific purpose. This necessitates the creation of input data sets tailored to individual software requirements, as well as the careful formatting and customization of outputs to meet specific needs. These tasks demand meticulous attention to detail and consume valuable time. Moreover, the repetitive nature of these steps, coupled with the manual integration of various software programs, can result in time-consuming and inefficient workflows, hampering the completion of comprehensive analyses.

Standard assessment practices involve repeating psychometric processes numerous times until adequate results are achieved. However, the repetitive nature of these manual steps poses challenges for educational practitioners, particularly psychometricians, as it increases the risk of errors. Because assessments typically have tight deadlines, completing the psychometric work on time while allowing for quality control is essential to ensure that the results are technically accurate and reliable.

Another crucial aspect to consider is the cost associated with the use of property software, which can be quite expensive (Martinková & Drabinová, 2018). This becomes particularly significant when considering the need for multiple licenses to facilitate a comprehensive analysis. Additionally, there are various challenges involved in accommodating diverse assessment requirements, such as managing exceptions and addressing unforeseen data errors. Consequently, these challenges can contribute to substantial costs in order to attain the desired research and analysis outcomes.

1.3. Reporting

Reporting is another important aspect of large-scale assessments (Ysseldyke & Nelson, 2002). Once the analyses are complete, they should be reported and shared with different stakeholders. Reports may include raw scores, proficiency levels, percentiles, and standard scores, whereas reports related to items and tests provide statistics and information at the item and test level (e.g., Goodman & Hambleton, 2004). The main aims of these reports are to deliver student outcomes and evaluate the performance items and tests. Furthermore, these reports can be utilized in order to share information with students, teachers, families, and educational policymakers (Rutkowski et al., 2010).

Reports should include clear statements for the intended educational stakeholders (Ysseldyke & Nelson, 2002). Therefore, educators may be burdened by unclear and disorganized results. As traditional software programs print standard results in a text format or proprietary formats (du Toit, 2003: Muraki & Bock, 2003; Thissen et al., 2003), manually and separately preparing these reports would entail a laborious and time-intensive endeavor. Therefore, producing customized reports would be helpful in working efficiently with many internal and external stakeholders.

2. METHOD

2.1. Design Philosophy

The underlying design philosophy of the pipeline primarily adopts a stage-oriented approach, differing from a modular approach where each major functionality of the pipeline is treated as a distinct and independent element capable of operating autonomously. The sequential nature of data processing requirements lent itself to this method, as the pipeline would need to perform various tasks before analysis and reporting. We implemented a strategy of isolating stages to ensure the separation of processing rules and the preservation of original data for comparison and validation purposes.

Although this approach is sequential in nature, the philosophy behind the pipeline was still to be both dynamic and automated. Each function was designed to handle data from any assessment with any configurations. The code therefore incorporates adaptability and atomism, eliminating the need for code replication.

2.2. Tools Used

The language chosen at the outset of the project was R (R Core Team, 2022). The R language has a few advantages over other languages considered, including Python (Van Rossum & Drake, 1995) and Julia (Bezanson et al., 2012). The main rationale behind this is that R is developed by statisticians, and as a result, its user community predominantly consists of professionals and academics from relevant fields. Consequently, there exists a high level of support for psychometric tasks within the community. Furthermore, R offers robust reporting tools, such as RMarkdown (Allaire et al., 2022), which greatly enhance the language's capabilities in generating comprehensive reports. These advantages meant the project could more efficiently get up and running by using already built open-source packages. In this pipeline, we leverage several key R packages to enhance our data analysis and reporting capabilities. Packages used in the pipeline are shown in Table 2.

Table 2. *Commonly used packages in the pipeline and their purpose.*

Name	Description
dplyr	The "grammar" of data
mirt	Psychometrics
stringr	Processing strings
RMarkdown	Reporting and generation of HTML documents
openxlsx	Reporting and generation of Excel reports

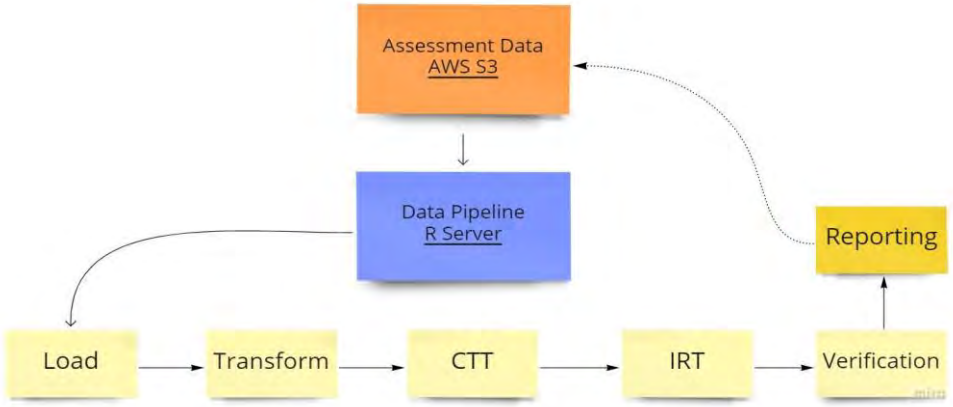
The *dplyr* package (Wickham et al., 2022) serves as the foundation for efficient data manipulation, allowing the pipeline to apply consistent and intuitive “grammar” very easily to incoming and outgoing datasets. For psychometric tasks, the project relied on the *mirt* (Chalmers, 2012) package, which offers a comprehensive suite of functions and tools

specifically tailored for psychometrics and item response theory (IRT) analysis. To handle string processing and manipulation tasks, we utilized the *stringr* package (Wickham, 2022). When it comes to generating high-quality reports, we used the powerful *RMarkdown* package (Allai*re* et al., 2022), which enables the pipeline to automatically produce dynamic HTML documents. The automated verification reports were one such document built with the package. Lastly, for generating professional-looking Excel reports, we used on the *openxlsx* package (Schauberger & Walker, 2022). This package also provided the ability to customize formatting for a specified range of cells and columns, including merging cells, bold, italics, underline, and creating borders.

2.3. Stages

The pipeline consists of several stages that must be completed successfully, from the beginning to the end, for it to run smoothly. Each stage is segmented by its purpose, with validations at each stage, so that troubleshooting is made easier. The flow of the stages in the pipeline is shown in Figure 1.

Figure 1. *A flow chart of the pipeline.*



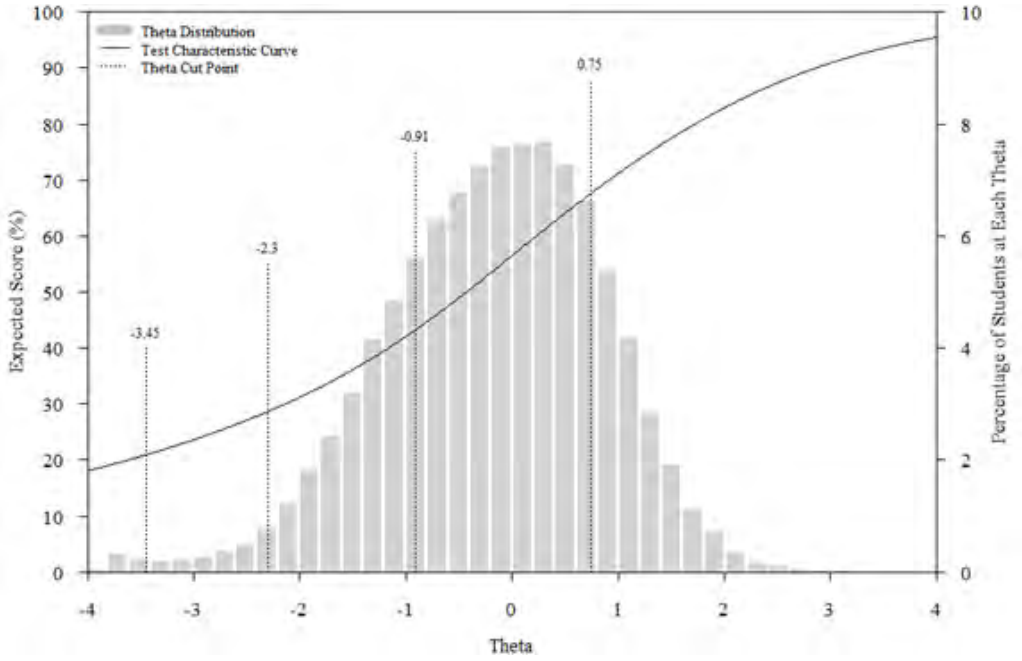
In the first stage, the standard process involves the pipeline retrieving data from Amazon S3 (Simple Storage Service), which is a cloud storage service provided by Amazon Web Services (AWS) where data is stored. If the data already exists in the local directory (the cache), the pipeline will import the data locally, saving time and processing power. We check whether the data meets the criteria for analysis and item/test specifications. At the end of this section, we create a single list of data frames that will be used in the next stages. During the transform stage, the pipeline will execute a series of processes including the application of business rules (including handling student exceptions and prorating), pivoting tables from long to wide format (for later use in IRT), and some aggregations for pre-analysis. Any kind of data cleaning is applied at this stage as well, which includes the filtering of invalid data. This stage ensures that the data frames are prepared for various types of analysis in subsequent stages.

In the CTT stage, the pipeline calculates item statistics (e.g., p, pbis, cbis) based on the CTT framework and conducts distractor analysis. These reports also include flags indicating if an item is too easy or too hard. Users have the flexibility to increase the number of flags or modify their values within the pipeline, not only in this particular section but also in other sections as per their decisions. The results of these sections help assessment teams and psychometricians to review item performance based on raw scores and frequency distributions (after the completion of the pipeline and the verification of results). For example, we can see how distractors or incorrect response options function in each item from distractor reports.

The pipeline moves to the IRT stage to conduct IRT-based analysis, including generating starting values, item calibration, equating (if necessary), and scoring (estimating thetas). In this

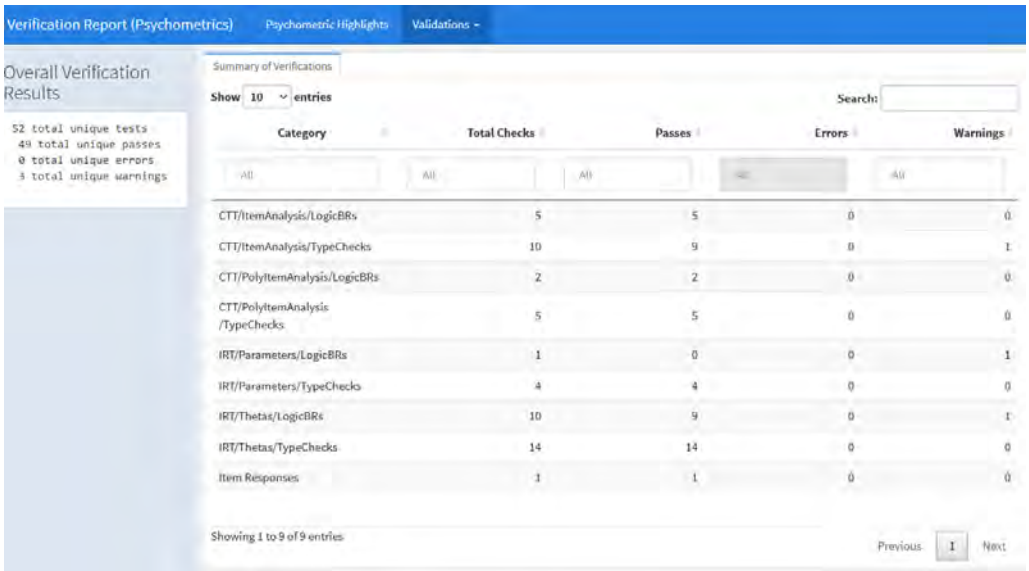
later stage, the pipeline estimates student abilities/thetas and assigns proficiency levels based on the cut scores that can be modified by the user. The pipeline produces all item- and test-level plots (see Figure 2) based on the IRT-based framework to examine individual items visually.

Figure 2. *Sample plot based on IRT framework.*



Following this stage, the pipeline generates a comprehensive report to verify the data and results prior to their reporting or publication. To ensure the accuracy and validity of the data provided to external services, the psychometrics pipeline has incorporated a robust set of data quality tests. These tests encompass the entire range of the datasets used, including both common tests applicable to all assessments and specific tests tailored to particular assessments. The tests include verifying the data format, structure of reports, constraints (nullable fields, primary and foreign keys), and business logic (consistency of the statistics reported both within a report and across several reports). The pipeline then generates an HTML report (see Figure 3) based on the outcomes of the data quality testing conducted at the conclusion of the analysis, ensuring that stakeholders and users receive accurate data.

Figure 3. *A sample page of a verification report.*



In the final stage, the pipeline generates various reports to psychometricians and content experts/assessment teams and generates database exports that are specifically designed for efficient and streamlined integration into the database, facilitating smooth and effective data transfer. Figure 4 shows examples of CTT and IRT reports.

Figure 4. *Sample pages of a CTT (below) and an IRT report (above).*

	A	B	C	D	F	G	H	I	J	K
1	label	id	ItemType	skill_category	a	b1	b2	b3	b	g
2	label.item1	item1	OR	A	1	-3.6869	-2.18543	0.35891	-1.83780667	
3	label.item2	item2	OR	A	1	-6.65895137	-2.2109488		-4.43495009	
4	label.item3	item3	MC	A	1				-1.84204	0.2
5	label.item4	item4	MC	A	1				-2.29955	0.2
6	label.item5	item5	OR	B	1	-3.19247701	-2.56037097		-2.87642399	
7	label.item6	item6	MC	B	1				-0.96892416	0.2
8	label.item7	item7	OR	C	1	-4.24080499	-2.44088468		-3.34084484	
9	label.item8	item8	MC	C	1				-0.04843	0.2
10	label.item9	item9	MC	A	1				-1.73279367	0.2
11	label.item10	item10	OR	A	1	-2.4403916	-2.69874177		-2.56956669	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	item_label	item_id	skill	lang	type	max_	num_responded	num_nr	item_mean	Test_Score_Means	p	pbis	cpbis	tooEasy	tooDifficult
2	label.item1	item1	A	en	MC	1	35463	20	0.958	55.683	0.958	0.345	0.324	TRUE	
3	label.item2	item2	B	en	MC	1	35460	23	0.809	56.845	0.809	0.419	0.381		
4	label.item3	item3	A	en	MC	1	35457	26	0.862	56.104	0.862	0.299	0.263		
5	label.item4	item4	B	en	OR	2	35450	33	0.873	56.215	0.873	0.345	0.311		
6	label.item5	item5	B	en	OR	3	35461	22	0.725	56.324	0.725	0.235	0.186		
7	label.item6	item6	C	en	OR	3	35454	29	2.43	56.039	0.81	0.461	0.359		
8	label.item7	item7	C	en	MC	1	64264	94	0.785	58.04	0.785	0.415	0.376		
9	label.item8	item8	A	en	MC	1	10507	8	0.934	55.099	0.934	0.321	0.295	TRUE	
10	label.item9	item9	A	en	MC	1	35463	20	0.958	55.683	0.958	0.345	0.324	TRUE	
11	label.item10	item10	A	en	MC	1	35460	23	0.809	56.845	0.809	0.419	0.381		
12															

Most of the stages in the pipeline are required, but some can be skipped. For example, the CTT stage is semi-required, as most of the analysis is not required to move on to IRT. On the other hand, the verifications stage is required to be performed and confirm the data quality is clean before the pipeline can begin the reporting stage. Lastly, it is worth noting that the pipeline can be utilized either locally or online, offering flexibility in its usage.

2.4. Architecture

The RStudio Server application was installed on AWS to provide an IDE from which to log in, view the code, make changes, and run the pipeline process (end to end). We also had the capability of synchronizing accounts to provide psychometricians with the same version of the code and data. Version control (with git) was implemented using multiple branches (for production, staging, and development) in order to ensure the ability to track changes, revert changes, and to stabilize the version of the code used to produce the results.

2.5. Performance Tuning

We carried out performance tuning and identified three main areas of improvement in the pipeline. Improving the storage aspect of the pipeline was the initial focus, and it proved to be a relatively straightforward task. To enhance efficiency, a local caching system was implemented, which allows data from S3 to be stored directly in the local directory of the RStudio account. Furthermore, we underwent a thorough review of the data model, enabling us to identify and exclude unnecessary fields and tables that were not relevant to the psychometrics pipeline. As a result, these redundant components were not stored locally, optimizing storage utilization and improving overall performance.

Next, the focus shifted toward optimizing memory usage in the pipeline. This involved reducing the size of tables that contained excessive or redundant data. For instance, one particular results

dataset did not require registration data until later stages of the pipeline when specific functions were invoked. To address this, objects or datasets were loaded or called only when necessary or as closely as possible to their required usage. Additionally, we proactively employed the rm() function to remove specific objects or a list of objects from memory, and the gc() function was utilized to enforce garbage collection in R. We implemented these to remove objects as soon as they were no longer needed. These measures effectively managed memory allocation and enhanced the efficiency of the pipeline's memory utilization. It is important to acknowledge that R exhibits a tendency to consume significant memory resources and may, at times, retain memory allocation without releasing it to the operating system until explicitly required (Morandant et al., 2012).

Lastly, the most significant improvement was achieved in terms of the pipeline's execution speed. In the initial iteration, the pipeline was constructed using base R functions, resulting in a relatively slow overall performance. Using the *dplyr* package (Wickham et al., 2022) (and its *tidyr* family of packages [Wickham & Girlich, 2022]) and the data.table (Dowle & Srinivasan, 2023) resulted in speed-ups in the range of 10-100 times faster, with some calculations taking minutes instead of hours.

2.6. Integration with a Larger Administration Environment

Various options for a user interface were considered that would avoid the complexity from having to access a large codebase in R. One such consideration, the Shiny package, provides a high-level package of modules from which to build a user interface in the R language itself, rather than using Javascript, HTML, and CSS. Projects using Shiny as a frontend are generally well suited for an isolated environment where the user uploads a file and analyzes the data, with user-friendly controls. However, given its availability, the already developed and available web portal was chosen as the interface with which to interact with the psychometrics pipeline.

The final stage of pipeline development involved its full automation as well as its activation from the web portal, which was accessible by various stakeholders. This would provide the ability for the psychometricians themselves to run the entire psychometric reporting process independently, without requiring any knowledge of the codebase itself or its configuration files. Accordingly, the pipeline code was modified to be initiated with Javascript, which itself would be activated based on user input coming from the web view (including which assessment and batch of data). The web view was also modified to provide a modifiable view of the configurations used in the pipeline (cut scores, items excluded, etc.). The version of the code used, time taken, and version of the underlying data would all be automatically recorded. With the above implemented, it was possible for psychometricians to log in to the web site, choose or edit the analysis configurations, run the psychometrics for a particular assessment, and have the results delivered in a data package all from the same portal.

3. FINDINGS

The psychometrics pipeline implemented for this project took a holistic approach to data processing. It was designed to be capable of integrating with various external sources of data, including databases and data lakes. It was further able to carry the data from import to transformation, to analysis, verification, and reporting without intervention on behalf of the user. We were able to integrate this end-to-end psychometric data pipeline into a larger ecosystem that includes the registration, testing, and reporting for an assessment administration. This is crucial as organizations are looking for ways to modernize their entire administration process, and that includes the statistical analysis and reporting thereof.

The pipeline successfully conducted CTT and IRT analyses, and the results were verified by multiple independent psychometricians. A key feature of the pipeline was the generation of analysis flags and highlighting of results that required the attention of psychometricians. These

flags proved invaluable for various teams as they helped to identify and address issues promptly, enabling operational improvements. Furthermore, the pipeline facilitated psychometric work before the main administration of assessments, allowing for item piloting and review of item changes. The psychometric-related sections of the pipeline were designed to accommodate different IRT models and mixed-format test designs. While primarily utilizing the *mirt* package, the pipeline remains flexible and open to incorporating other packages, offering the capability to perform a diverse range of psychometric analyses through coding and facilitating cross-validation of results. Notably, an extension was added to the pipeline to automate test form generation based on available item banks. This extension underwent rigorous testing and successfully generated parallel test forms. Another significant extension involved running simulation studies to test various test criteria, providing valuable insights to assessment teams and test designers. The pipeline efficiently executed these simulation studies, further enhancing its capabilities and utility in the assessment process.

We implemented a strategy of self-verification within the scope of the pipeline. Requirements and constraints were understood, and a suite of data quality tests was built accordingly. These tests enabled us to perform thorough testing on the psychometric results produced by the pipeline. In this fashion, every aspect of the data could be tested, including data types, data format, data length, and business rule constraints. As the number of tests increased, the need for a visualization of the results began to be apparent. We tapped into the power of RMarkdown, which enabled us to provide fully automated reports in the style of a flexible web dashboard. This also provided the ability to more easily share and report on data quality results with stakeholders, leading to increased transparency, trust, and oversight.

Overall, we proved that R could be used to integrate with a separate system utilizing a different language and server, providing a compatible external process. Furthermore, the R packages used were overall successful in meeting our requirements. While many software packages provide a "black box" situation, we were able to dig deeper into the code used for important packages such as *mirt*, allowing us to vet the processes underneath. Such transparency and control were instrumental in ensuring the reliability and validity of our psychometric analyses. We were also able to fine-tune the performance of our R-based processes, providing a much more rapid deployment of results.

4. DISCUSSION and CONCLUSION

Modern approaches to assessment have created new requirements that are now being supported by technological innovation (Moncaleano & Russell, 2018). As the industry is modernizing, analysis is following suit. Pushed forward by provincial, national level and international testing, the industry is also beginning to adopt new approaches to handle the incoming large-scale data (Rutkowski et al., 2010). This paper presented a comprehensive solution for end-to-end data processing in large-scale assessments, addressing a significant gap in the field. Our data pipeline offers numerous benefits for practitioners, psychometricians, educators, and researchers involved in testing. It has demonstrated the ability to handle large databases, minimizing human error by automating manual processes, enabling the replication of complex workloads, ensuring high-quality outputs, and reducing overall costs associated with psychometric analysis of large-scale assessment data. By following our approach, testing organizations can enhance automation, ensure quality assurance, and achieve greater efficiency in their own large-scale assessments.

This project provided important further developments on the topic of psychometrics, data processing, administration, reporting, and the combination thereof. We also learned several lessons in developing this project. One, an understanding of the requirements and constraints of the data analysis is fundamental. We should also draw attention to the importance of having

a clear vision for the overall architecture of the pipeline. Early implementations led to costly duplication of development, human errors, and inefficiencies in running the pipeline.

We also proved that R can be a flexible and powerful tool for constructing an end-to-end data pipeline. Python is frequently used for these purposes (Weber, 2020), but we accomplished a standardized data pipeline while using the strengths of the R language. Other languages, such as Python or Julia, were not needed to fulfill the requirements for the import, transformation, analysis, and export of data. However, in the future, it would be recommended to investigate a mixed-language approach. Given Julia has an advantage in speed and memory efficiency (Dogaru & Dogaru, 2015), the language could be used for the heavy lifting by pulling and transforming data, leaving R to do the analysis and reporting.

Further embracing technological advances in recent decades would also have been beneficial in this project. Version control, at first, was quite basic, which led to the implementation of branches later in the project. As well, containerizing through Docker (Merkel, 2014) would improve the portability of the project. Docker would encapsulate the entire environment and automate all the steps it takes to build the technology architecture, installation of packages and software, and possible simulation of datasets. This would provide the ability to use the project in various mirror and user acceptance testing environments with little to no error or additional work involved in setup (Azab, 2017).

Continuous integration and continuous deployment (CI/CD) pipelines would enhance quality assurance, ensuring that updates to the code are properly tested before being deployed into production (where official results are produced). Integrated with the git repository, these CI/CD pipelines can automatically test changes made to the code before deployment, integrating unit testing and sample data into code updates. Linting packages would provide additional oversight on code syntax and style during any further development.

While there were successes throughout this project, there are some key areas that deserve further research. We noted that aspects of the integration could be improved upon. One method of enhancing the integration with the web view portal would be to transform the R code into a fully-fledged REST (Representational State Transfer) API (Application Programming Interface), using the *plumber* package (Schloerke & Allen 2023). The pipeline was integrated as a sub-process that is (except for a few configuration options) independent of the parent process that called it. An API structure would allow the pipeline to receive requests in a standardized format (using GET requests) and return data in any number of formats (csv, JSON, etc.) directly to the caller process. This would facilitate a more customized activation of the pipeline, calling certain functions and not others (running the CTT and not the IRT). This would also allow the pipeline to run asynchronously, even enabling multiple psychometrics runs at the same time.

While developing the pipeline before and during the administration windows, we found that there was a need for the large-scale generation of student data that would match the constraints of valid psychometric analysis. To this end, the simulation of student data would be an improvement to the early testing of the pipeline but also a move forward in the portability of the pipeline. As such, the pipeline could be instantiated on a fresh server, generate simulated data, and run the analysis to show that the pipeline is functioning correctly.

Lastly, the project itself was custom-built from the ground up, rather than utilizing a pre-built pipeline or finished application software like the *ShinyItemAnalysis* package (Martinková & Drabinová, 2018). Although this increased the workload, it also provided the opportunity to construct a more adaptable and customizable codebase that provides much greater functionality, tailored to the needs of each individual client. This provides a greater ability to continue to extend the project into the future, with further functionality.

Acknowledgments

This study's initial results were previously presented at the IAEA 2022 Annual Conference, held from October 2nd to 7th, 2022, in Mexico City, Mexico.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

H.C.B discloses that she was employed as a psychometrician at Vretta Inc. while working on and completing this paper and is currently affiliated with Northern Alberta Institute of Technology.

Authorship Contribution Statement

Ryan Schwarz: Conception, Design, Supervision, Materials, Data collection and Processing, Analysis, Literature Review, Writing. **H. Cigdem Bulut:** Conception, Design, Supervision, Materials, Data collection and Processing, Analysis, Literature Review, Writing. **Charles Anifowose:** Conception, Design, Supervision, Materials, Critical Review.

Orcid

Ryan Schwarz https://orcid.org/0009-0004-5867-3176
H. Cigdem Bulut https://orcid.org/0000-0003-2585-3686
Charles Anifowose https://orcid.org/0009-0006-2524-9613

REFERENCES

Addey, C., & Sellar, S. (2018). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, H.K. Altinyelken, & M. Novelli (Eds.), *Global education policy and international development: New agendas, issues and policies* (3rd ed., pp. 97–117). Bloomsbury Publishing.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... & Iannone, R. (2022). *rmarkdown: Dynamic Documents for R*. R package version, 1(11).

Ansari, G.A., Parvez, M.T., & Al Khalifah, A. (2017). Cross-organizational information systems: A case for educational data mining. *International Journal of Advanced Computer Science and Applications, 8*(11), 170-175. http://dx.doi.org/10.14569/IJACSA.2017.081122

Azab, A. (2017, April). Enabling docker containers for high-performance and many-task computing. In *2017 ieee international conference on cloud engineering (ic2e)* (pp. 279-285). IEEE.

Bezanson, J., Karpinski, S., Shah, V.B., & Edelman, A. (2012). *Julia: A fast dynamic language for technical computing.* ArXiv Preprint ArXiv:1209.5145.

Bertolini, R., Finch, S.J., & Nehm, R.H. (2021). Enhancing data pipelines for forecasting student performance: Integrating feature selection with cross-validation. *International Journal of Educational Technology in Higher Education, 18*(1), 1-23. https://doi.org/10.1186/s41239-021-00279-6

Bertolini, R., Finch, S.J., & Nehm, R.H. (2022). Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines. *Computers and Education: Artificial Intelligence,* 3, 100067. https://doi.org/10.1016/j.caeai.2022.100067

Bryant, W. (2019). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation*, *22*(1), 1. https://doi.org/10.7275/70yb-dj34

Camara, W.J., & Harris, D.J. (2020). Impact of technology, digital devices, and test timing on score comparability. In M.J. Margolis, R.A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 104-121). Routledge.

Chalmers. R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Croudace, T., Ploubidis, G., & Abbott, R. (2005). BILOG-MG, MULTILOG, PARSCALE and TESTFACT. *British Journal of Mathematical & Statistical Psychology, 58*(1), 193. https://doi.org/10.1348/000711005X37529

Desjardins, C.D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Dogaru, I., & Dogaru, R. (2015, May). Using Python and Julia for efficient implementation of natural computing and complexity related algorithms. In *2015 20th International Conference on Control Systems and Computer Science* (pp. 599-604). IEEE.

Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of 'data.frame'*. https://r-datatable.com,https://Rdatatable.gitlab.io/data.table.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Scientific Software International.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Erlbaum.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Kamens, D.H., & McNeely, C.L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative education review, 54*(1), 5-25. https://doi.org/10.1086/648471

Goodman, D.P., & Hambleton, R.K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145-220. https://doi.org/10.1207/s15324818ame1702_3

Liu, O.L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M.C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19-28. https://doi.org/10.1111/emip.12028

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Addison Wesley, Reading, MA.

Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research, and Evaluation, 27*(1), 22.

IBM (2020). *IBM SPSS Statistics for Windows,* Version 27.0. IBM Corp.

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *R Journal, 10*(2), 503-515.

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.

Microsoft Corporation. (2018). *Microsoft Excel.* Retrieved from https://office.microsoft.com/excel

Moncaleano, S., & Russell, M. (2018). A historical analysis of technological advances to educational testing: A drive for efficiency and the interplay with validity. *Journal of Applied Testing Technology, 19*(1), 1–19.

Morandat, F., Hill, B., Osvald, L., & Vitek, J. (2012). Evaluating the design of the R language: Objects and functions for data analysis. In *ECOOP 2012–Object-Oriented Programming: 26th European Conference, Beijing, China, June 11-16, 2012. Proceedings 26* (pp. 104-131). Springer Berlin Heidelberg.

Muraki, E., & Bock, R.D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales* [Computer software]. Scientific Software International, Inc.

Oranje, A., & Kolstad, A. (2019). Research on psychometric modeling, analysis, and reporting of the national assessment of educational progress. *Journal of Educational and Behavioral Statistics, 44*(6), 648-670. https://doi.org/10.3102/1076998619867105

R Core Team (2022). *R: Language and environment for statistical computing.* (Version 4.2.1) [Computer software]. Retrieved from https://cran.r-project.org.

Reise, S.P., Ainsworth, A.T., & Haviland, M.G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current directions in psychological science*, *14*(2), 95-101.

Rupp, A.A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3*(4), 365-384. https://doi.org/10.1207/S15327574IJT0304_5

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, *17*(1), 20-32.

Rutkowski, L., Gonzalez, E., Joncas, M., & Von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151. https://doi.org/10.3102/0013189X10363170

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing" intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, *4*(6).

Schauberger, P., & Walker, A (2022). *openxlsx: Read, Write and Edit xlsx Files*. https://ycphs.github.io/openxlsx/index.html, https://github.com/ycphs/openxlsx

Schleiss, J., Günther, K., & Stober, S. (2022). Protecting student data in ML Pipelines: An overview of privacy-preserving ML. In *International Conference on Artificial Intelligence in Education* (pp. 532-536). Springer, Cham.

Schloerke, B., & Allen, J. (2023). *plumber: An API Generator for R*. https://www.rplumber.io, https://github.com/rstudio/plumber

Schumacker, R. (2019). Psychometric packages in R. *Measurement: Interdisciplinary Research and Perspectives, 17*(2), 106-112. https://doi.org/10.1080/15366367.2018.1544434

Skiena, S.S. (2017). *The data science design manual*. Springer.

Sung, K.H., Noh, E.H., & Chon, K.H. (2017). Multivariate generalizability analysis of automated scoring for short answer items of social studies in large-scale assessment. *Asia Pacific Education Review*, *18*, 425-437. https://doi.org/10.1007/s12564-017-9498-1

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *MULTILOG 7 for Windows: Multiple category item analysis and test scoring using item response theory* [Computer software]. Scientific Software International, Inc.

Van Rossum, G., & Drake Jr, F.L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Volante, L., & Ben Jaafar, S. (2008). Educational assessment in Canada. *Assessment in Education: Principles, Policy & Practice, 15*(2), 201-210. https://doi.org/10.1080/09695940802164226

Weber, B.G. (2020). *Data science in production: Building scalable model pipelines with Python*. CreateSpace Independent Publishing.

Wickham, H. (2022). *stringr: Simple, consistent wrappers for common string operations*. https://stringr.tidyverse.org.

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation*. Retrieved from https://dplyr.tidyverse.org.

Wickham, H., & Girlich, M. (2022). *tidyr: Tidy messy data.* https://tidyr.tidyverse.org

Wise, S.L. (2018). Computer-based testing. In *the SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 341–344). SAGE Publications, Inc.

Ysseldyke, J., & Nelson, J.R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. (pp. 467-483). Routledge

Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362. https://doi.org/10.1207/S15324818AME1504_02