AQuAA: Analytics for Quality Assurance in Assessment

Manqian Liao, Yigal Attali, Alina A. von Davier
Duolingo, Inc.
mancy@duolingo.com, yigal@duolingo.com, avondavier@duolingo.com

ABSTRACT

High-stakes digital-first assessments are assessments that can be taken anytime and anywhere in the world and their scores impact test takers' lives. Computational psychometrics, a blend of theory-driven psychometrics and data-driven algorithms, provides the theoretical underpinnings for these data-rich assessments. The unprecedented flexibility, complexity, and high-stakes nature of these digital-first assessments poses enormous quality assurance challenges. In order to ensure these assessments meet both "the contest and the measurement" requirements of high-stakes tests [5], it is necessary to conduct continuous pattern monitoring and be able to promptly react when needed. In this paper, we illustrate the development of a quality assurance system, Analytics for Quality Assurance in Assessment (AQuAA), for a high-stakes and digital-first assessment. To build the system, educational data from continuous administrations of the assessments are mined, modeled and monitored via an interactive dashboard.

Keywords

high-stakes assessment, digital-first assessment, quality assurance

1. INTRODUCTION

Digital-first assessments are based on artificial intelligence (AI) tools that direct and optimize test-takers' experience. These digital tools include automatic systems for test development, scoring, and test delivery. In contrast to traditional large-scale assessments that are based on in-person administration to large groups of test takers in fixed locations, digital-first assessments are administered continuously to individual test takers, thus allowing for unprecedented flexibility. The advantages of the digital-first assessments have manifested themselves during the pandemic when traditional group assessments in brick-and-mortar test centers became impractical.

When digital-first assessments are used for high-stakes purposes (for example, for admissions or employment purposes), they, as any traditional high-stakes assessments, have a significant potential impact on test takers' lives. Thus, the digital-first high-stakes assessment also need to meet both "the contest and the measurement" requirements of high-stakes tests [5], where the "contest" here refers to the expectation that the test gives everyone a fair chance; the "measurement" refers to the requirement that the test is accurate and valid.

Quality assurance refers to a systematic process to maintain the high quality of the test and assessment scores and to prevent errors from all stages of the test, including test design, item design and development, test scoring, test analysis and score reporting [7]. Its complement, quality control, refers to a set of methods and statistics to evaluate the quality of the test. Many of the statistics and methods employed for quality assurance and quality control are similar, with quality control being part of the quality assurance overarching system. The International Test Commission Guidelines have articulated step-by-step procedures for quality control of general educational assessments but many of the steps are more applicable to traditional assessments, that is, "large-scale testing operations where multiple forms of tests are created for use on set dates."[7]

Since digital-first assessments differ from traditional assessments in many respects (e.g., administration frequency, item bank size), it is necessary to develop quality assurance procedures that are tailored for digital-first assessments. Developing such systems also requires research into the appropriate methodology to identify the most relevant statistics to be monitored for such new type of assessment, which is the focus of this paper.

In order to conduct quality assurance for digital-first high-stakes assessments, we developed a monitoring system named Analytics for Quality Assurance in Assessment (AQuAA), which is a blend of psychometrics and educational data mining packed into a dynamic and interactive dashboard-based system. AQuAA was designed to accommodate at least two unique characteristics of the digital-first assessments. On one hand, many key aspects of digital-first assessments, such as item generation and scoring, are automatically accomplished by machine. Therefore, compared to traditional assessments, the quality assurance of the digital-first assessments requires more extensive data mining techniques.

Computational psychometrics [18, 17] is leveraged to mine and model educational data in order to develop the statistics included in AQuAA. On the other hand, as a consequence of the continuous nature of administration, the quality assurance activities for digital-first assessments need to be conducted more frequently with a flexible timeline. In addition, tools that facilitate swift and efficient communication are indispensable so that prompt actions can be taken when issues are detected. In AQuAA, a variety of statistics are updated regularly and are integrated into an interactive dashboard for continuous pattern monitoring and timely communication purposes. AQuAA is also symbiotic with other activities (such as item development) given the fact that conclusions drawn from AQuAA could be used to direct the maintenance and improvement of the assessment.

This paper elaborates the development of AQuAA and aims to address three research questions: 1) What statistics should be used as indicators of test quality and score validity of digital-first assessments? 2) How to identify patterns and irregularities relevant to test quality of digital-first assessments? and 3) How to communicate the findings from the quality assurance process to stakeholders? This paper is focused on the the quality assurance of the test administration activities.

2. RELATED WORK

Quality assurance plays an important role in maintaining test score validity. [1] indicated that mistakes that jeopardize the assessment score validity could occur at all stages of assessment development and administration and that the mistakes could accumulate since many stages are contingent on previous stages. Therefore, quality control guidelines and step-by-step procedures [1, 2, 7] have been developed to help test developers identify possible mistakes as well as the causes of these mistakes, thereby helping them to identify solutions to fix the mistakes and prevent the mistakes from happening again.

Quality control procedures were mostly designed for traditional large-scale assessments that are administered in only a few test dates and have large test volumes in each administration [7, 1], with [2] being an exception. [2] recommended a quality control procedure for continuous mode tests (i.e., tests that are administered to small groups of test takers on many test dates) which share some similarities with digital-first assessments. Moreover, [2] have demonstrated an automated quality control system for continuous mode tests and the system consists of both an automatic part and a human review part. These two parts also apply to the quality assurance of digital-first assessments. In the automatic part, a number of steps that need to be conducted recurrently and can be implemented programmatically are packed into an automatic procedure with the use of digital tools. Steps in such an automatic procedure may include fetching the data from the database, conducting a variety of quality control analyses (see [9] for a review of quality control methods) and generating statistical reports. In the human review part, human experts are trained to review the statistical reports generated from the automatic procedure in order to identify potential irregularities or outliers, and determine whether or what actions need to be taken to handle these irregularities.

The foundation of an automated quality assurance procedure consists of a wide range of data mining and data visualization techniques. In the realm of quality assurance, the data mining and data visualization techniques serve two major purposes: First, to describe the trends and seasonal patterns of the assessment statistics; Second, to detect abrupt changes in the relevant assessment statistics. [9] have summarized a number of statistical methods and data visualization techniques for score quality assurance purposes. Various time series techniques can be chosen to describe trends or seasonal patterns, which include linear ANOVA models [4], regression with autoregressive moving-average [10], harmonic regressions [8] and dynamic linear models [19]. The Shewhart chart is a useful data visualization tool for continuous of the test score characteristics [9, 12, 14]. In terms of detecting abrupt changes in the assessment statistics, some model-based approaches have been applied to mine the data and identify abrupt changes in score time series, such as change-point models and hidden Markov model [9]. A data visualization techniques for detecting abrupt changes is cumulative sum (CUSUM) charts [13].

The products of the automated quality assurance procedure may include summary tables of the statistics, graphs and statistical testing results [2]. These statistical products could be organized into different formats, such as reports [2] and dashboards [11]. Since the products of the automated quality assurance procedure will serve as the starting point of the human review process [2], the choice of organizing format should be determined by the ease of communication to the targeted stakeholders.

3. MAJOR COMPONENTS OF AQUAA

This section illustrates how several key components of AQuAA address the research questions mentioned above. AQuAA has been launched as a minimum viable product (MVP) and additional features and statistics are being added to the system. This paper demonstrates the application of AQuAA the Duolingo English Test, a digital-first assessment. In order to help readers understand the context from which the AQuAA is developed, this section will start with a brief overview of the Duolingo English Test. However, the methodologies for designing AQuAA and the statistics considered for evaluation are intended to be adaptable to other digital-first assessments.

3.1 Overview of the Assessment

The Duolingo English Test is a high-stakes computerized adaptive test that is designed to be accessible anywhere and anytime [15]. Thus, it also falls under the category of continuous mode assessments [2]. The Duolingo English Test is an adaptive test, with a very large item bank that has been designed by subject matter experts (SMEs) and produced automatically by the machine. The items are reviewed by panels of SMEs to ensure quality and cultural fit. The items are scored automatically and the scoring methods are reviewed periodically by SMEs. Each individual test is proctored remotely using a complex and innovative asynchronous system that involves both AI-based tools and human proctors. Discrepancies or unusual situations are adjudicated by SMEs. Test results are reviewed through the quality assurance process in AQuAA. As part of this process, a wide range of process information related to test takers' behavior

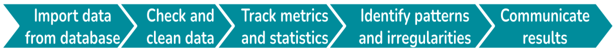

Import data from database
Check and clean data
Track metrics and statistics
Identify patterns and irregularities
Communicate results

Figure 1: AQuAA updating procedure

(e.g., time per item response, length of responses, etc.) is analyzed and monitored for quality assurance. The amount of data and the multiple sources and types of data are significantly more demanding of sophisticated analytics than is the case in more traditional assessments.

3.2 Overview of AQuAA

An overview of the procedure of developing and updating AQuAA is shown in Figure 1. Except for the first step (i.e., importing the data) that is relatively straightforward, the design of each step requires deliberation and, thus, is elaborated in the following sections. The steps in Figure 1 are scheduled to be automatically implemented on a daily basis (and in some cases more frequently). R [16] is the major programming tool used to develop AQuAA and automate the AQuAA updating process.

3.3 Checking and Cleaning Data

In general, the assessment data used for AQuAA can be separated into two types: Person-level data and item-response-level data. Person-level data contain variables that describe the overall person/session information, such as test takers' overall test score, sub-scores, test dates, and background characteristics. Item-response-level data contains variables that delineate information about each item the test taker responded to, such as item IDs, item difficulty levels, item responses and item scores, and other process information such as time duration test takers spent on each item.

After the data are imported, the integrity of the data is inspected to ensure that the data used for subsequent analyses are accurate and of high quality. For example, data are inspected for irregular values (e.g., negative values in time duration variables), and the causes of any such values are further investigated to identify any potential threats to the integrity of the data collection process.

3.4 Tracking Metrics and Statistics

The first research question is to determine what metrics and statistics are most relevant to monitor over time in order to evaluate the health of a continuous assessment. In order to support a statistical quality assurance system, AQuAA monitors results in the following five categories across time, adjusting for seasonality effects.

1. **Scores**. Test scores are directly used by test users (e.g., test takers, institutions), thus important indices at the level of test scores, including overall scores, sub-scores, and item type scores, are tracked in AQuAA. Score-related statistics include the location and spread of scores, inter-correlations between scores, bivariate or multivariate outliers, person fit, internal consistency reliability measures and standard error of measurement (SEM), and validity coefficients (e.g., correlation with self-reported external measures).

2. **Test taker profile**. The composition of the test taker population is tracked over time, as it could be used to explain the variability in test scores to some extent. Specifically, the (percentage) volume of test takers in the important population categories, such as country, native language, gender, age, intent in taking the test, and other background variables, are tracked. In addition, many of the score statistics are tracked across major test taker groups.

3. **Repeaters**. Repeaters are defined as those who take the test more than once within a 30-day[1] window. The prevalence, composition, and performance of the repeaters are tracked. The composition of the repeater population is defined with respect to the same test taker profile categories discussed above. The performance of the repeater population is tracked with many of the same test score statistics identified above, with additional statistics that are specific to repeaters: location and spread of both the first and second tests, as well as their difference, and test-retest reliability (and SEM).

4. **Item analysis**. As tests consist of items, ensuring that items are of high quality and that the item quality is stable over time are the prerequisites of maintaining the validity of the test scores. In AQuAA, item quality is quantified with four categories of item performance statistics: Item difficulty, item discrimination, item slowness (response time), and differential item functioning (DIF). Tracking these statistics would help test developers to develop expectations about the item bank with respect to item performance, flag items with extreme and/or inadequate performance, and detect drift in measures of performance across time.

5. **Item exposure**. The item exposure statistics concern how frequent each item (or each group of items) are used. An item being used either too frequently (over-exposure) or too infrequently (under-exposure) are undesirable for maintaining the item quality. An important statistic in this category is the item exposure rate, which is calculated as the the number of test administrations containing a certain item divided by the total number of test administrations. Tracking the item exposure rates can help flag under- or over-exposure of items.

3.5 Identifying Patterns and Irregularities

The second research question concerns the identification of patterns and irregularities in the data, which involves the development of the alarming mechanism of AQuAA. Developing the alarming mechanism in AQuAA is challenging partly due to the fact that the population of test takers is evolving and changing constantly, and, thus, many of the tracked metrics cannot be assumed to be stationary over time. Instead, the tracked metrics are often prone to systematic variation over and beyond predictable changes due to seasonality effects, thereby making it complicated to set an appropriate alarming criteria for the alarming mechanism.

[1] The day threshold of determining repeaters could be adjusted based on the test taking policy and the research purpose

The alarming mechanism in AQuAA is intended to detect persistent but smaller trends as well as alert large and abrupt changes that may be due to a problem in the assessment. To achieve these goals, we combined model-based psychometric analyses method with the time series and control charts techniques, both of which are useful for distinguishing systematic changes from chance variation in outcome processes.

The psychometric model-based methods allow us to track metrics after adjusting for certain factors (e.g., test takers' background characteristics), thus increasing the metrics' comparability over time. Specifically, in AQuAA, the item statistics and metrics are adjusted for test taker ability and background variables, and test taker statistics and metrics are adjusted for item characteristics.

3.6 Communicating Results

Our third research question involves how to communicate the information to the operational analysts as well as to the business unit. To visualize the trends and patterns of the statistics and facilitate the communication in the human review process, statistics are plotted using the ggplot2 R package [20]. Line plots are one of the most basic tools to visualize the time-series data. For example, Figure 2 demonstrates the stable trend of the mean of the overall test score during the Fall of 2020. Each dot in these figure represent a statistic calculated using a day worth of data; the lines are smoothed lines created by the locally weighted scatterplot smoothing (LOWESS) [3] method in order to represent the trends of the statistics.

Plots are also used to visualize the alerts raised by the AQuAA alarming mechanism introduced in Section 3.5. In AQuAA, the alerts are classified into three severity categories which are represented by different color codes. Specifically, yellow, orange and red represent low, medium and high levels of severity, respectively. For example, Figure 3 displays a monitoring plot for the daily median response time a few alerts in low severity. Once an alert is raised by AQuAA, messages are automatically sent to inform all the relevant stakeholders via email and the organization communication tool.

Various statistics and figures are integrated into an interactive dashboard using the flexdashboard [6] package. Figure A.1 demonstrates the layout of the dashboard. At the top of the dashboard (i.e., Section 1), there are five tabs corresponding to the five categories of statistics articulated in Section 3.4. Within each tab, the relevant statistics are arranged into storyboards: The statistics could be further classified into subcategories and allocated into different pages (i.e., Section 2); figures are displayed at the major section of the dashboard (i.e., Section 3); text description and some numerical results are displayed in the commentary section (i.e., Section 4).

4. THE APPLICATION OF AQUAA

As the quality assurance of digital-first assessments is a combination of automatic processes and human review processes, the AQuAA system is used as the starting point for the human review process, and the human review process, in turn, helps AQuAA to evolve into a more powerful tool to detect assessment validity issues. Figure B.1 demonstrates an example human review process following every week's updates of AQuAA: SMEs meet to review the alerts raised by AQuAA alarming mechanism and review for any anomalies that are suggested by the AQuAA figures but have not been caught by the AQuAA alarming mechanism. The SMEs review each individual alert and determine whether it is an actual sign of a validity issue or it is a false alarm. If the alarm is believed to be caused by a validity issue, follow-up actions are taken to determine the severity and urgency, fix and document the issue. If the issue had not been caught by the AQuAA alarming mechanism, improvements would be made to the AQuAA functionality such that AQuAA would be more sensitive in detecting the issue.

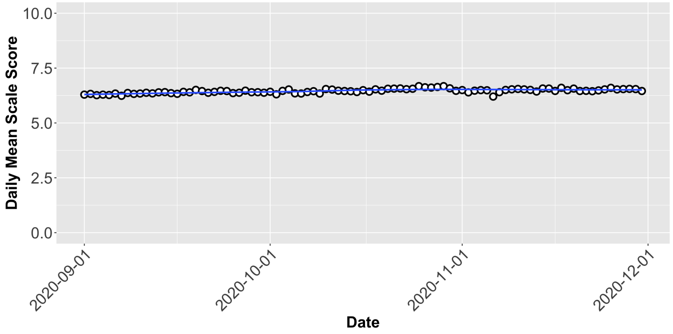


Figure 2: Trend of daily mean overall scores

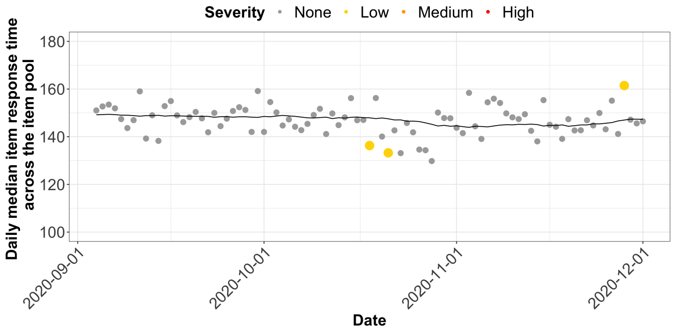


Figure 3: Trend of daily median response time with alerts.

5. DISCUSSION

This paper demonstrates the development of a quality assurance system that is tailored for digital-first assessments that are continuously administered. Several research questions motivated many of these approaches, as very few of the traditional methods apply to the digital-first assessments. The steps and considerations for building the quality assurance system have been elaborated, so that test developers could adapt the methodologies in this paper to their own assessments. It should be noted that the list of quality assurance statistics presented here is not exhaustive. Instead, due to the data-rich nature of the digital-first assessment, the list of monitoring statistics is expected to be lengthened and improved as the research in statistical techniques advances. The list of monitoring statistics should also be customized to the purposes and characteristics of the assessment. Hence, the infrastructure of AQuAA is designed to be so flexible as to incorporate and monitor additional statistics.

6. REFERENCES

[1] A. Allalouf. Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, 26(1):36–46, 2007. Publisher: Wiley Online Library.

[2] A. Allalouf, T. Gutentag, and M. Baumer. Quality Control for Scoring Tests Administered in Continuous Mode: An NCME Instructional Module. *Educational Measurement: Issues and Practice*, 36(1):58–68, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12140.

[3] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979. Publisher: Taylor & Francis.

[4] S. J. Haberman, H. Guo, J. Liu, and N. J. Dorans. Consistency of SAT® I: Reasoning test score conversions. *ETS Research Report Series*, 2008(2):i–20, 2008. Publisher: Wiley Online Library.

[5] P. W. Holland. Measurements or contests? Comments on Zwick, bond and Allen/Donoghue. In *Proceedings of the social statistics section of the American Statistical Association*, volume 1994, pages 27–29. American Statistical Association Alexandria, VA, 1994.

[6] R. Iannone, J. J. Allaire, B. Borges, RStudio, K. I. D. CSS), A. A. D. CSS), J. Mosbech (StickyTableHeaders), N. Bossart (Featherlight), L. Verou (Prism), D. Baranovskiy (Raphael.js), S. Labs (Raphael.js), B. Djuricic (JustGage), T. Sardyha (Sly), B. Lewis (Examples), C. Sievert (Examples), J. Kunst (Examples), R. Hafen (Examples), B. Rudis (Examples), and J. Cheng (Examples). flexdashboard: R Markdown Format for Flexible Dashboards, June 2020.

[7] International Test Commission (ITC). ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. *International Journal of Testing*, 14(3):195–217, July 2014. Publisher: Taylor & Francis Ltd.

[8] Y.-H. Lee and S. J. Haberman. Harmonic regression and scale stability. *Psychometrika*, 78(4):815–829, 2013. Publisher: Springer.

[9] Y.-H. Lee and A. A. von Davier. Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, 78(3):557–75, 2013.

[10] D. Li, S. Li, and A. A. von Davier. Applying time-series analysis to detect scale drift. In *Statistical models for test equating, scaling, and linking*, pages 327–346. Springer, 2009.

[11] L. Mohadjer and B. Edwards. Paradata and dashboards in PIAAC. *Quality assurance in education*, 2018. Publisher: Emerald Publishing Limited.

[12] M. H. Omar. Statistical process control charts for measuring and monitoring temporal consistency of ratings. *Journal of Educational Measurement*, 47(1):18–35, 2010. Publisher: Wiley Online Library.

[13] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. Publisher: JSTOR.

[14] W. D. Schafer, B. J. Coverdale, H. Luxenberg, and J. Ying. Quality control charts in large-scale assessment programs. *Practical Assessment, Research, and Evaluation*, 16(1):15, 2011.

[15] B. Settles, G. T. LaFlair, and M. Hagiwara. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263, 2020.

[16] R. D. C. Team. R: A language and environment for statistical computing. 2013.

[17] A. A. von Davier. Virtual and collaborative assessments: Examples, implications, and challenges for educational measurement. In *Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning 2015*, 2015.

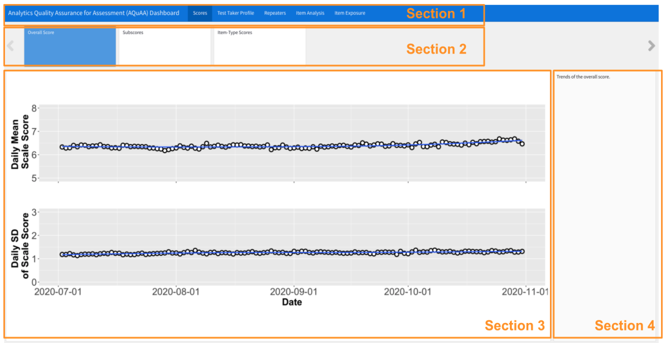
[18] A. A. von Davier. Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, 54(1):3–11, 2017.

[19] R. G. Wanjohi, P. W. van Rijn, and A. A. von Davier. A state space approach to modeling irt and population parameters from a long series of test administrations. In *New developments in quantitative psychology*, pages 115–132. Springer, 2013.

[20] H. Wickham. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2):180–185, 2011. Publisher: Wiley Online Library.

APPENDIX

A. DEMO OF AQUAA



Trends of the overall score.

Figure A.1: Demo of AQuAA with annotations. Section 1 is the navigation bar containing five tabs corresponding to the five categories of statistics monitored in AQuAA. Within each tab, the relevant statistics are grouped into subcategories and are arranged into storyboards. Section 2 display the pages that correspond to the subcategories of statistics. Section 3 is the major section of the dashboard where figures are displayed. Section 4 is the commentary section that display the text description and numerical results.

B. SUBJECT MATTER EXPERT REVIEW PROCESS

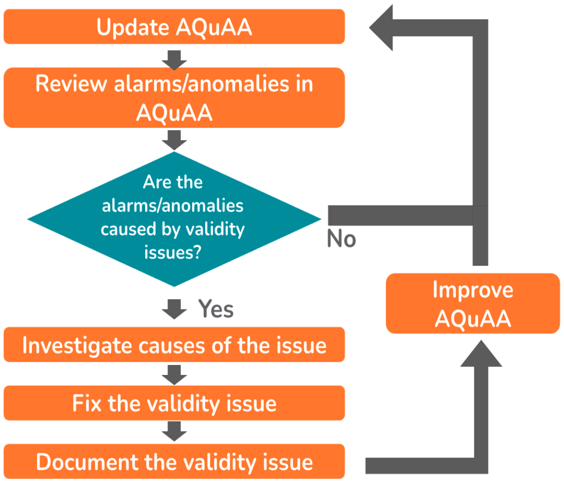

Update AQuAA
Review alarms/anomalies in AQuAA
Are the alarms/anomalies caused by validity issues?
No
Yes
Investigate causes of the issue
Fix the validity issue
Document the validity issue
Improve AQuAA

Figure B.1: Subject Matter Expert (SME) review process.