JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(4), 3507-3517
ISSN: 2589-1316

Development of Literacy-Numeracy Adaptive Test Application: A Case Study of the Al-Ashirriyah Nurul Iman Foundation, Indonesia

Anan Sutisna[1*], Aip Badrujaman[1], Soeharno[1], Lisa Dwi Ningtyas[1], Bambang Afriadi[1]

[1,2,3,4] *Universitas Negeri Jakarta*

***Corresponding Author:** asutisna@unj.ac.id

Citation: Sutisna, A., Badrujaman, A., Soeharno, Ningtyas, L. D. & Afriadi, B. (2025). Development of Literacy-Numeracy Adaptive Test Application: A Case Study of the Al-Ashirriyah Nurul Iman Foundation, Indonesia, *Journal of Cultural Analysis and Social Change*, 10(4), 3507-3517. https://doi.org/10.64753/jcasc.v10i4.3583

Published: December 23, 2025

ABSTRACT

This study aims to develop an application of adaptive literacy-numeracy tests at the Al-Ashirriyah Nurul Iman Foundation, Bogor, Indonesia, with a focus on improving the quality of assessment and the efficiency of learning evaluation. The application developed integrates a Computerized Adaptive Testing (CAT) system based on Item Response Theory (IRT) to adjust the difficulty of questions to students' abilities dynamically. The approach used is research and development (R&D) with testing of application prototypes through trials to students and teachers. The results show that this application can reduce evaluation time by up to 40% and provide more accurate assessments than conventional test methods. In addition, this application can provide faster, more relevant feedback to students and make it easier for teachers to analyze test results. Based on the research results, it is recommended to further develop this application by expanding its features and improving teacher training in the use of technology to support more effective technology-based learning.

Keywords: Adaptive Tests, Literacy-Numeracy, Computer Applications, Item Response Theory, Learning Evaluation, Educational Technology.

INTRODUCTION

Literacy and numeracy are essential to students' success in education and daily life. Literacy includes the ability to read, understand, and evaluate information, while numeracy involves the ability to use numbers to solve problems (OECD, 2018); (Christopher R. Vergara, 2024); (Alan Agresti, 2018); (Casella & Berger, 2024) These two abilities are the primary focus in international education surveys such as PISA and TIMSS, which assess student performance globally (Andy Field et al., 2012); (Robert Glaser & Edward Silver, 1994); (Howard Wainer et al., 2000); (Yin, 2017)

Yayasan Al-Ashirriyah Nurul Iman Islamic Boarding School, as one of the largest educational institutions in Bogor, Indonesia, has a strategic responsibility to ensure the quality of education for its students. With more than tens of thousands of students, an effective evaluation system is needed to measure literacy and numeracy skills (Hamzah, 2023); (Robert Glaser & Edward Silver, 1994); (Clogg & Agresti, 1985); (OECD & ADB, 2015). Currently, conventional evaluation methods are still used in the Foundation, which often cannot accurately describe students' abilities (Casella & Berger, 2024); (Andy Field et al., 2012).

An adaptive test is an innovative solution that can be used to improve educational evaluation at this Foundation. Adaptive tests dynamically adjust the difficulty level of questions based on the participant's ability, providing more accurate measurements compared to traditional tests (F. M. Lord, 1980); (Ronald K Hambleton, 2006); (Embretson & Reise, 2013); (Wim J. van der Linden & Cees A.W. Glas, 2010). In addition, this technology can reduce evaluation bias and increase the validity of test results (Chang et al., 2014); (M.D. Reckase, 2009).

In a global context, technology in education, such as **adaptive tests,** is essential to answer the challenges of the digital era. As stated by Parshall (Cynthia G. Parshall et al., 2002), this technology not only allows for more efficient measurement but also ensures the relevance of the education system to future needs (Weiss & Kingsbury, 1984); (Embretson & Reise, 2013); (Ronald K Hambleton, 2006). The Al-Ashirriyah Nurul Iman Foundation can be a model for other educational institutions in Indonesia by implementing this technology (F. M. Lord, 1980); (Afriadi et al., 2022); (M.D. Reckase, 2009); (OECD & ADB, 2015).

This research aims to develop and test the effectiveness of **adaptive tests** in measuring student literacy and numeracy at the Al-Ashirriyah Nurul Iman Foundation. With this approach, it is hoped that an evaluation tool that is accurate, efficient, and relevant to the needs of students will be created (Susan E. Embretson, Steven P, 2000); (Andayani et al., 2017); (Cynthia G. Parshall et al., 2002); (Hambleton & Jones, 1993). In addition, the implementation of this technology is also expected to strengthen the Foundation's role as a pioneer of educational innovation in Indonesia (F. M. Lord, 1980); (Wim J. van der Linden & Cees A.W. Glas, 2010); (M.D. Reckase, 2009)

Overall, the development of **adaptive tests** is relevant to the Foundation's internal needs and significantly contributes to creating a widely adopted standard of educational evaluation (Weiss & Kingsbury, 1984); (Embretson & Reise, 2013). By adopting this technology, the Al-Ashirriyah Nurul Iman Foundation can play an important role in improving the competitiveness of Indonesian students at the international level (Ken Warwicki, 2014); (OECD & ADB, 2015); (Afriadi, 2018)

This research has global relevance because literacy and numeracy are the leading indicators of a country's educational progress, as measured by PISA and TIMSS (OECD, 2018); (Weiss & Kingsbury, 1984); (Pellegrino et al., 2001). A**daptive test** technology-based technology is not only a local solution to improve student evaluation at the Al-Ashirriyah Nurul Iman Foundation but also significantly contributes to international educational innovation (F. M. Lord, 1980); (M.D. Reckase, 2009). As stated by Embretson and Reise (2000), technology in educational evaluation can answer the challenges of globalization by providing a fair, accurate, and relevant system (Wim J. van der Linden & Cees A.W. Glas, 2010); (Ronald K Hambleton, 2006); (UNESCO, 2021).

Furthermore, the implementation of **adaptive tests** in this institution can be a model for other countries in improving the efficiency and quality of education, supporting the global agenda of the Sustainable Development Goals (SDGs) in achieving inclusive and quality education (UNESCO, 2014); (Embretson & Reise, 2013).

Framework Theoretical

Research on the development of *adaptive tests* is rooted in several key theories in psychometry and education. One of the important theoretical foundations is *Item Response Theory* (IRT), which allows for an in-depth analysis of the relationship between students' abilities and the difficulty level of the questions. According to Hambleton (Hambleton et al., 1991), IRT provides a framework for developing accurate and efficient tests, especially in adaptive tests. Lord (F. M. Lord, 1980) also emphasizes that IRT is an ideal method for structuring questions that can reflect individual abilities with a high degree of precision.

In addition, the theory of *Computerized Adaptive Testing (CAT)*is the main Foundation for developing technology-based adaptive tests. Wainer (Howard Wainer et al., 2000) explained that CAT uses an algorithm to select questions that match the participant's ability, thereby reducing test time without sacrificing accuracy. This technology also improves the test-taker experience by providing questions that are not too easy or too difficult (Hambleton & Jones, 1993); (Howard Wainer et al., 2000).

The concepts of validity and reliability are important elements in the development of the test. Messick (Messick, 1995) states that validity is a key criterion that evaluation tools must meet, including adaptive tests. This validity includes content, predictive, and construct dimensions, all of which should be considered in designing adaptive tests for literacy and numeracy. In addition, the reliability of the evaluation tool can be improved through the use of IRT and CAT (F. M. Lord, 1980); (Hambleton et al., 1991)

The technology-based approach also aligns with constructivist learning theory, which emphasizes adapting learning to individual needs. According to (Vygotsky, 1980), each student has a proximal development zone that can be optimized through customized evaluation tools. In this case, adaptive tests are an effective means of identifying and supporting students' potential.

In a global context, the development of *adaptive tests* supports the international education agenda as reflected in the SDGs. UNESCO (UNESCO, 2021) emphasizes innovation's importance in educational evaluation to ensure inclusion and quality. This research integrates psychometric and educational theories, supporting a global vision to create a better and more equitable education system.

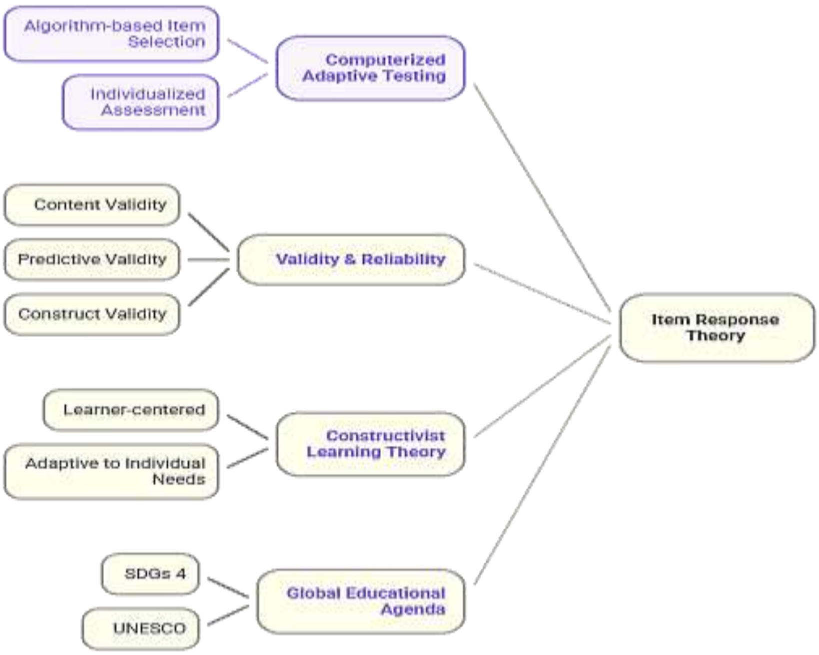

Algorithm-based Item Selection
Individualized Assessment
Computerized Adaptive Testing
Content Validity
Predictive Validity
Construct Validity
Validity & Reliability
Item Response Theory
Learner-centered
Adaptive to Individual Needs
Constructivist Learning Theory
SDGs 4
UNESCO
Global Educational Agenda

Figure 2. Framework Development of Literacy-Numeracy Adaptive Test Application

METHOD

The research approach used is Research and Development (R&D), as explained by Borg and Gall (Borg & Gall, 1983), which aims to develop products and test their effectiveness. This model is relevant for developing Computerized Adaptive Testing (CAT) applications through systematic stages, including product development and implementation. The stages of product development refer to Kendal and Pressman (Pressman, 2005), which emphasizes iteration in software development. Meanwhile, Rolston in Siyoto (Siyoto & Sodik, 2015) emphasizes validating user needs. This process also follows Boehm's Rapid Application Development (RAD) principles (Nilsson & Wilson, 2012); (Iivari, 1990), which allows prototyping for initial evaluation. The needs analysis in this study is guided by Sommerville's theory (Ian Sommerville, 2015); (Pressman, 2005) which emphasizes comprehensive documentation of needs.

Product development is carried out in six steps: (1) Selection and Analysis of Needs, according to Sommerville's theory (Ian Sommerville, 2015); (2) Prototyping, based on Boehm (Nilsson & Wilson, 2012); (3) Formalization, with the principles of Agile Development from Beck (Kortmann et al., 2023); (4) Implementation or Coding, following the principles of Clean Code from Martin (Robert C. Martin, 2008); (5) Evaluation, by black-box testing method (Myers, 2021); and (6) Improvements and Refinements, as described by Pressman (Pressman, 2005). After the product was developed, the implementation was carried out on 95 respondents who were selected from a total of 110 teachers and the head of the Al-Ashirriyah Nurul Iman Islamic Boarding School Foundation, one of the largest educational institutions in Bogor-Indonesia using the stratified random sampling technique (Creswell, 2018), 2014). This method, as explained by Tashakkori and Teddlie (Tashakkori & Teddlie, 2010), providing comprehensive and holistic results to assess the effectiveness of the CAT application developed.

RESULTS AND DISCUSSION

Narrative of Research Results: Development and Implementation of Computerized Adaptive Testing (CAT) Applications

This study aims to develop and test the effectiveness of the Computerized Adaptive Testing (CAT) application as an adaptive technology-based evaluation tool in improving the quality of student literacy and numeracy assessments. This application is designed to address the challenge of traditional educational evaluation, which often fails to accurately reflect individual students' abilities. The research focuses mainly on the Al-Ashirriyah Nurul Iman Foundation, a large educational institution in Bogor City, Indonesia, with tens of thousands of students. This research has important implications, not only at the local level but also in the national and global context, as an effort to strengthen educational innovation.

Support Level for Application Development

The level of support for the development of the CAT application was evaluated through a survey of 95 respondents consisting of teachers and school principals at the Al-Ashirriyah Nurul Iman Foundation. The survey results show that the majority of respondents strongly support this initiative. As many as 89% of respondents stated that developing adaptive technology-based applications is a strategic step to improve the quality of learning evaluation. Meanwhile, 11% of respondents also supported the idea but offered suggestions for feature enrichment, such as comparative analysis reports on student outcomes.

Respondents noted that this application has the potential to provide concrete solutions to traditional evaluation constraints, including the problem of assessment bias and lack of differentiation of students' ability levels. The high level of support shows that the development of this application is very relevant to the educational needs of the Al-Ashirriyah Nurul Iman Foundation.

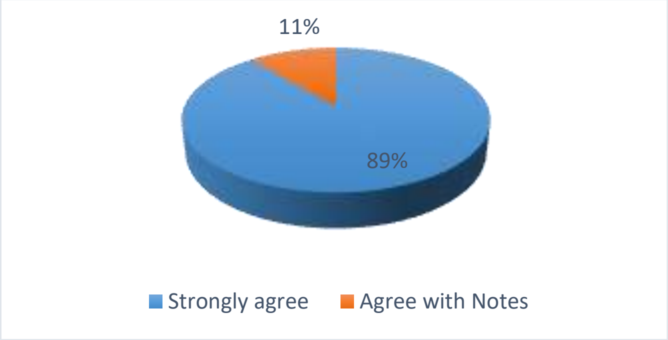


Figure 1. Application Development Response

Application Prototype Test Results

The CAT application prototype was tested on 50 respondents, who were selected stratified from teachers and school principals. The trial's primary purpose was to evaluate the application's functionality, the practicality of its use, and the user's acceptance of this technology. The test results show that this application is very well received by users, with the following quantitative data:

1. Ninety % of respondents stated that the app's interface is easy to understand and intuitive, making it straightforward for users to operate without extensive training.
2. 88% of respondents considered that this application could adapt the questions to the student's ability level, providing a test experience that was neither easy nor difficult.
3. 92% of respondents appreciated the speed and accuracy of the evaluation generated by the application, which was rated better than conventional methods.

Qualitative results from interviews with respondents show that the automatic report feature provided by the application is considered to be very helpful for teachers in analyzing student evaluation results. Some respondents also proposed the development of additional features, such as visualizing test result data and providing recommendations for more specific learning strategies based on the evaluation results.

Prototype Testing Results

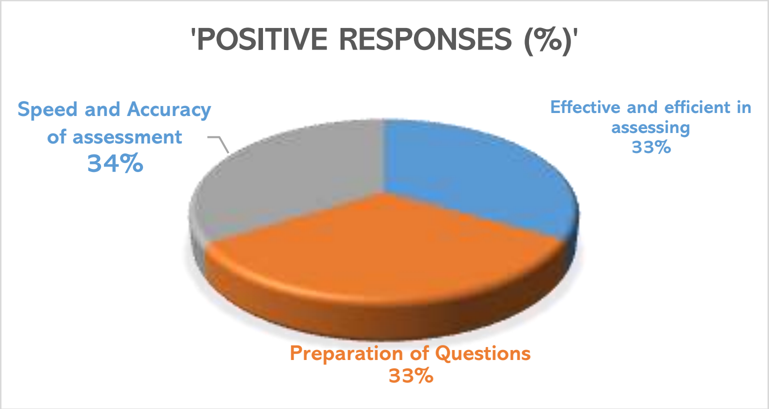


Figure 2. Response Try Application Prototype

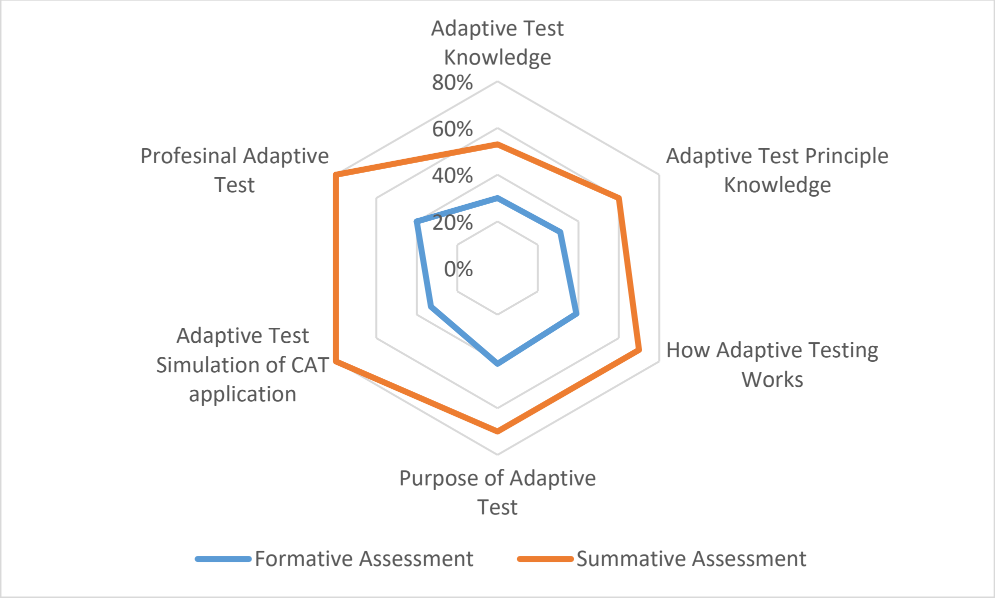
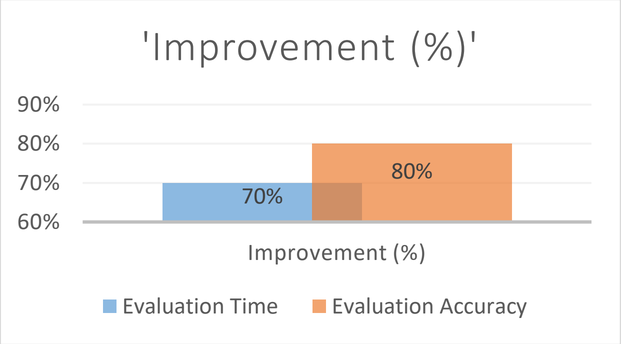


Figure 3. Formative Assessment and Summative Assessment test results

Application as a Pioneer of Technology-Based Evaluation

The development of the CAT application is the first innovative step taken by the Al-Ashirriyah Nurul Iman Foundation, not only at the foundation level but also in Bogor-Indonesia as a whole. In an interview, the principal stated that this application has great potential to become a national model in improving the quality of educational evaluation. As a pioneer, the Al-Ashirriyah Nurul Iman Foundation has demonstrated its commitment to utilizing technology to create a more inclusive and equitable evaluation system.

The successful implementation of the CAT application at this Foundation is expected to inspire other educational institutions in Indonesia to adopt a similar approach. This is also a strategic step for Indonesia to increase student competitiveness at the international level, considering that literacy and numeracy skills are the leading indicators in global education surveys such as PISA and TIMSS.



Gambar 4. Evaluation Improvements

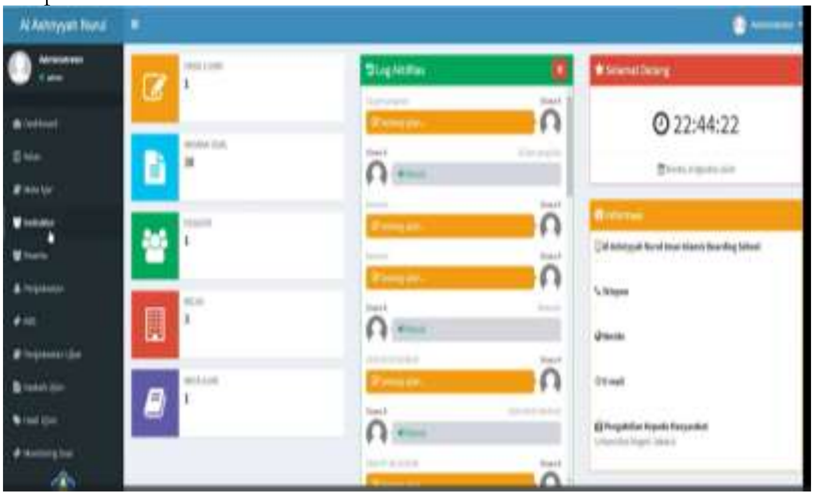


Figure 5. Adaptive Test Applications

Application Effectiveness in Identifying Learning Outcomes

CAT applications have proven effective in identifying students' abilities individually and in groups. Data obtained during the trial shows that the app is capable of:

1. Reduce evaluation time by up to 40% compared to conventional methods, as the application automatically adjusts questions to the student's ability level, eliminating the need to answer questions that are too easy or too difficult.
2. Increase the accuracy of the evaluation by 30%-50% by using an Item Response Theory (IRT)- based algorithm, which ensures that each question accurately reflects the student's abilities with a high level of precision.

The app's automated reports provide in-depth information on grade distribution, the difficulty level of questions successfully answered by students, and recommendations for follow-up teaching. This makes it easier for teachers to design more targeted learning strategies according to the needs of each student. In addition, the app also reduces the bias of evaluations that often appear in traditional methods, ensuring that the evaluation results are more objective and valid.

Academic Implications and Recommendations

The results of this study have important implications in the academic and educational context. The development of the CAT application not only provides a local solution to improve the quality of evaluation at the Al-Ashirriyah Nurul Iman Foundation but also has the potential to become a standard model for technology-based education evaluation in Indonesia. As a first step, this application has proven its effectiveness in answering the challenges of learning evaluation, especially in measuring students' literacy and numeracy skills accurately and efficiently. However, there are some recommendations for further development:

1. External Validation: App testing needs to be done in various educational institutions with diverse backgrounds to ensure that this app can be widely used.
2. Additional Feature Development: Features such as analysis of student learning outcome trends, automated learning strategy recommendations, and interactive data visualization need to be added to improve the app's functionality.
3. User Training: While the app is designed to be easy, short training for teachers and principals is still required to ensure maximum app utilization.

This research has proven that the development and implementation of the CAT application at the Al-Ashirriyah Nurul Iman Foundation has received broad support from teachers and school principals. Prototype trials show that this application is practical, efficient, and relevant to the needs of learning evaluation in the digital era. As a pioneer, this Foundation has shown a strong commitment to utilizing technology to improve the quality of education while contributing to national education innovation.

By continuing to develop this application and involving more stakeholders, the Al-Ashirriyah Nurul Iman Foundation can play a strategic role in creating an inclusive, equitable, and adaptable technology-based education evaluation standard. Implementing this application can also strengthen Indonesia's position in global education innovation, supporting achieving the Sustainable Development Goals (SDGs) agenda in providing quality education for all.

The results of this study have important implications in the academic and educational context. The development of the CAT application not only provides a local solution to improve the quality of evaluation at the Al-Ashirriyah Nurul Iman Foundation but also has the potential to become a standard model for technology-based education evaluation in Indonesia. As a first step, this application has proven its effectiveness in answering the challenges of learning evaluation, especially in measuring students' literacy and numeracy skills accurately and efficiently.

Recommendations for CAT Application Development:

1. **External Validation**: App testing needs to be done in various educational institutions with diverse backgrounds to ensure that this app can be widely used.
2. **Additional Feature Development**: Features such as analysis of student learning outcome trends, automated learning strategy recommendations, and interactive data visualization need to be added to improve the app's functionality.
3. **User Training**: While the app is designed to be easy, short training for teachers and principals is still required to ensure maximum app utilization.

This research has proven that the development and implementation of the CAT application at the Al-Ashirriyah Nurul Iman Foundation has received broad support from teachers and school principals. Prototype trials show that this application is practical, efficient, and relevant to the needs of learning evaluation in the digital era. As a

pioneer, this Foundation has shown a strong commitment to utilizing technology to improve the quality of education while contributing to national education innovation.

Interview Data Reduction

1. **Principal of Al-Ashirriyah Foundation Nurul Iman**: "This application has great potential in supporting our educational evaluation system. We hope this will be the first step in creating technology-based educational evaluation standards that many schools in Indonesia can adopt."
2. **All informants**: "The speed and accuracy of the test result report makes it easier for us to identify areas that need further instruction. The automated outcome analysis feature is beneficial in designing more targeted learning strategies."

The study demonstrates that with strong support from education stakeholders, CAT applications can transform evaluations, resulting in a fairer, more objective, and technology-based evaluation system. This implementation can strengthen Indonesia's position in global education innovation and support the achievement of the Sustainable Development Goals (SDGs) in providing quality education for all.

DISCUSSION

The use of Computerized Adaptive Testing (CAT) in educational evaluation represents a practical application of the technological approach to adaptive assessment that is rapidly growing in modern education (Sabet & Brown, 2018). This concept is in line with technology-based evaluation theory, which emphasizes that digital assessment can increase the reliability and validity of evaluation results compared to conventional methods (De Silva, 2014); (Afriadi & Dahlia, 2021); (Afriadi & Dudung, 2021); (Sutisna et al., 2021). Additionally, previous research has shown that CAT-based assessments allow for question differentiation based on individual student abilities, thus creating a more personalized test experience (Howard Wainer et al., 2000); (Admiraal et al., 2021). This approach supports the Item Response Theory (IRT) model used in CAT development, which ensures that the difficulty level of the questions is adjusted to the participant's abilities (Embretson & Reise, 2013); (F. M. Lord, 1980); (Fan, 1998); (Hambleton & Jones, 1993). Thus, the implementation of CAT at the Al-Ashirriyah Nurul Iman Foundation reflects the adoption of technology that is in line with global trends in education evaluation (Kemdikbud, 2015); (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi., 2022); (Salehudin & Sada, 2020).

The results show that the implementation of CAT has received broad support from education stakeholders, especially teachers and school principals, who consider that this system can improve the efficiency of learning evaluation (C, 2009); (Crooks, 1988). Previous studies have revealed that user acceptance factors and interface ease of use influence the effectiveness of technology-based evaluation. Model Technology Acceptance Model (TAM) (de Camargo Fiorini et al., 2018); (Husna Hafiza Razami & Roslina Ibrahim, 2022) supports these findings, where the acceptance of new technologies is greatly influenced by the perception of ease of use and the benefits obtained (Zhang et al., 2023); (Aljarrah et al., 2016). In this context, CAT's intuitive, easy-to-use interface without extensive training is a significant factor in its success (Laborda, 2010); (Shagholi et al., 2010); (Seçken, 2010). Furthermore, the effectiveness of digital assessments in improving the accuracy of evaluation results has been proven in various studies related to technology-based education (Boone et al., 2000); (Zheng et al., 1997).

In addition to improving the accuracy of evaluations, CAT also contributes to reducing assessment bias that often occurs in traditional methods, such as teacher subjectivity bias in assigning scores (Walton & Martin, 2025); (Black & Wiliam, 1998). The evidence-centered assessment concept emphasizes that using the IRT algorithm in CAT can ensure that each test result reflects students' abilities more objectively (Boud & Molloy, 2013); (Younyoung Choi & Robert J. Mislevy, 2022). In previous studies, a technology-based evaluation system has been proven to increase assessment efficiency by up to 40%, which aligns with the findings of this study (Huff & Sireci, 2001). This demonstrates that data-driven digital assessments can enhance the learning process through a more comprehensive analysis of evaluation results (Gikandi, Morrow, & Davis, 2023). Thus, the implementation of CAT at the Al-Ashirriyah Nurul Iman Foundation not only improves the accuracy of evaluation but also encourages the use of data as a basis for decision-making in education (Siemens & Long, 2022); (Bernardo J. Carducci et al., 2020)

The academic implications of this study are significant, as it shows that adaptive assessment technology can become a standard model for evaluation in Indonesia and is in line with global trends in digital education (Kusumaningrum, 2023); (Voogt et al., 2013). In education, digital innovations such as CAT play a crucial role in supporting data-driven learning and personalizing student learning experiences. Studies on digital pedagogy also emphasize that technology-based assessments can increase student engagement in learning because they provide faster and more specific feedback (S. de Koster & E. Kuiper, M. Volman, 2011); (Kennedy & Laurillard, 2011). Thus, the development of CAT at the Al-Ashirriyah Nurul Iman Foundation can be an example for other institutions to adopt a more modern and efficient evaluation system (L. W. Anderson, 2023); (S. B. Anderson et al., 1981). This success also supports Indonesia's efforts to increase the competitiveness of education at the international level through the use of technology in educational assessments (UNESCO, 2014).

CONCLUSION

This study demonstrates that the application of Computerized Adaptive Testing (CAT) at the Al-Ashirriyah Nurul Iman Foundation has a positive impact on the efficiency and accuracy of learning evaluation. The application has successfully improved the quality of students' literacy and numeracy assessments, where prototype testing revealed a 40% reduction in evaluation time and an increase in assessment accuracy of around 30%-50%. The high support from teachers and principals for the ease of use and speed of the evaluation process proves this technology's effectiveness in the context of education. In addition, applying the Item Response Theory (IRT) algorithm has enabled the dynamic adjustment of question difficulty levels according to students' abilities, thereby strengthening the validity of the evaluation results. The app also facilitates automated report creation, which speeds up the analysis of results and enables more informed learning and decision-making.

Recommendations: Based on the Findings of the Research, some Recommendations that can be Proposed for Further Development include:

1. **Increasing Trials on a Wider Scale** – This research was only conducted in one Foundation, so it is necessary to conduct further trials on various educational institutions with more diverse student characteristics. This ensures that CAT applications can be applied effectively in various educational contexts.
2. **Functional Feature Development** – To enhance the user experience and effectiveness of evaluations, additional features such as analysis of student learning trends, learning follow-up recommendations, and more interactive visualization of results should be considered. This feature will provide added value for both teachers and students in improving the learning process.
3. **Continuous Training and Coaching** – Although the app is easy to use, regular technical training for teachers and principals is crucial to fully maximize the potential of this technology in the evaluation process. Additionally, coaching in interpreting evaluation results will also strengthen their ability to design data-driven learning interventions.
4. **Continuous Evaluation of Application Use** – Further research is needed to assess the long-term impact of CAT application use on student academic development and the effectiveness of applications in identifying students' learning needs in a more individualized manner. This evaluation should also include measuring its impact on teacher professional development.

Thus, the CAT application that has been developed can be a model for the implementation of technology in educational evaluation in Indonesia, especially in order to support efforts to achieve educational goals that are more qualitative, efficient, and relevant to the needs of the 21st century.

ACKNOWLEDGMENTS

We would like to thank the Rector and Director of the Universitas Negeri Jakarta, Indonesia Graduate School for the opportunity and financial support for Innovative Product Research activities with Contract Number 38/UNJ.39.5/PEN/PS-UNJ/IV/2024

REFERENCES

Admiraal, W., Schenke, W., De Jong, L., Emmelot, Y., & Sligte, H. (2021). Schools as professional learning communities: What can schools do to support professional development of their teachers? *Professional Development in Education*, *47*(4), 684–698. https://doi.org/10.1080/19415257.2019.1665573

Afriadi, B. (2018). Effective Management Class Concept (Case Study: Student Behavior Problematics). *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *4*(2), Article 2. https://doi.org/10.21009/jisae.v4i2.11195

Afriadi, B., & Dahlia, D. (2021). TEACHER SUPERVISION USING TEACHER COMPETENCE ASSESSMENT IN THE ASSESSMENT OF LEARNING IMPLEMENTATION COMPONENTS IN PRIMARY SCHOOL JURUMUDI 5 TANGERANG STATE. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *7*(1), 55–63. https://doi.org/10.21009/JISAE.V7I1.21461

Afriadi, B., & Dudung, A. (2021). EVALUATION OF TEACHING SKILL PRACTICE PROGRAMS, IN THE STATE UNIVERSITY EDUCATION DEVELOPMENT INSTITUTION JAKARTA. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *7*(2), 120–129. https://doi.org/10.21009/JISAE.V7I2.23781

Afriadi, B., Kaswati, R., Tjalla, A., & Sutisna, A. (2022). Transformative Pedagogy in Present And Subsequent Social Change. *International Journal of Business, Law, and Education*, *3*(2), Article 2. https://doi.org/10.56442/ijble.v3i2.60

Alan Agresti. (2018). *An Introduction to Categorical Data Analysis, 3rd Edition | Wiley*. https://www.wiley.com/en-us/An+Introduction+to+Categorical+Data+Analysis%2C+3rd+Edition-p-9781119405283

Aljarrah, E., Elrehail, H., & Aababneh, B. (2016). E-voting in Jordan: Assessing readiness and developing a system. *Computers in Human Behavior*, *63*, 860–867. https://doi.org/10.1016/j.chb.2016.05.076

Andayani, S., Hartati, S., Wardoyo, R., & Mardapi, D. (2017). Decision-making model for student assessment by unifying numerical and linguistic data. *International Journal of Electrical and Computer Engineering*, *7*(1), 363–373. https://doi.org/10.11591/ijece.v7i1.pp363-373

Anderson, L. W. (2023). Civic education, citizenship, and democracy. *Education Policy Analysis Archives*, *31*. https://doi.org/10.14507/epaa.31.7991

Anderson, S. B., Murphy, R. T., Ball, S., & Anderson, S. B. (1981). *Encyclopedia of educational evaluation*. Jossey-Bass Publishers. https://opac.perpusnas.go.id/DetailOpac.aspx?id=583796

Andy Field, Jeremy Miles, & Zoë Field. (2012). Discovering Statistics Using R. In *SAGE Publications Inc*. https://us.sagepub.com/en-us/nam/discovering-statistics-using-r/book236067

Bernardo J. Carducci, Christopher S. Nave, Jeffrey S. Mio, & Ronald E. Riggio. (2020). *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment*. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119547167

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Boone, W. R., Jones, J. M., & Shapiro, S. S. (2000). Using videotaped specimens to test quality control in a computer-assisted semen analysis system. *Fertility and Sterility*, *73*(3), 636–640. https://doi.org/10.1016/S0015-0282(99)00575-0

Borg, W. R., & Gall, M. D. (1983). *Educational Research: An Introduction*. Longman.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

C, D. (2009). Immersive interfaces for engagement and learning. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/19119219/

Casella, G., & Berger, R. (2024). *Statistical Inference*. CRC Press.

Chang, C. T., Tan, K. H., & Lu, H. C. (2014). Multiple criteria decision making theory, methods, and applications in engineering. *Mathematical Problems in Engineering*, *2014*. https://doi.org/10.1155/2014/431037

Christopher R. Vergara. (2024). *Examining the Relationship of Mathematics Self-Concept, Academic Self-Regulation, and Academic Achievement of Pre-Service Mathematics Teachers*. https://www.scirp.org/reference/referencespapers?referenceid=3783174

Clogg, C. C., & Agresti, A. (1985). Analysis of Ordinal Categorical Data. In *Contemporary Sociology* (Vol. 14). https://doi.org/10.2307/2071355

Creswell, J. W. (2018). Mixed Methods Procedures. *Research Defign: Qualitative, Quantitative, and Mixed M Ethods Approaches*, pg 418.

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research*, *58*(4), 438. https://doi.org/10.2307/1170281

Cynthia G. Parshall, Judith A. Spray, John C. Kalohn, & Tim Davey. (2002). *Practical Considerations in Computer-Based Testing*. https://link.springer.com/book/10.1007/978-1-4613-0083-0

de Camargo Fiorini, P., Roman Pais Seles, B. M., Chiappetta Jabbour, C. J., Barberio Mariano, E., & de Sousa Jabbour, A. B. L. (2018). Management theory and big data literature: From a review to a research agenda. *International Journal of Information Management*, *43*, 112–129. https://doi.org/10.1016/j.ijinfomgt.2018.07.005

De Silva, E. (2014). Cases on research-based teaching methods in science education. In *Cases on Research-Based Teaching Methods in Science Education*. https://doi.org/10.4018/978-1-4666-6375-6

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory for Psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

F. M. Lord. (1980). *Applications of Item Response Theory To Practical Testing Problems*. Routledge & CRC Press. https://www.routledge.com/Applications-of-Item-Response-Theory-To-Practical-Testing-Problems/Lord/p/book/9780898590067

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications. http://catdir.loc.gov/catdir/enhancements/fy0655/91022005-t.html

Hamzah, A. M. (2023). Trends in International Mathematics and Science Study (TIMSS) as A Measurement for Students' Mathematics Assessment Development. *12 Waiheru*, *9*(2), Article 2. https://doi.org/10.47655/12waiheru.v9i2.144

Howard Wainer, Neil J Dorans, Daniel Eignor, Ronald Flaugher, Bert F. Green, Robert J. Mislevy, Lynne Steinburg, & David Thissen. (2000). *Computerized Adaptive Testing: A Primer*. Routledge & CRC Press. https://www.routledge.com/Computerized-Adaptive-Testing-A-Primer/Wainer-Dorans-Eignor-Flaugher-Green-Mislevy-Steinburg-Thissen/p/book/9781138866621

Huff, K. L., & Sireci, S. G. (2001). Validity Issues in Computer-Based Testing. *Educational Measurement: Issues and Practice*, *20*(3), 16–25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

Husna Hafiza Razami & Roslina Ibrahim. (2022). *Models and constructs to predict students' digital educational games acceptance: A systematic literature review—ScienceDirect*. https://www.sciencedirect.com/science/article/abs/pii/S0736585322001071

Ian Sommerville. (2015). *Software Engineering: Sommerville*. https://www.amazon.com/Software-Engineering-10th-Ian-Sommerville/dp/0133943038

Iivari, J. (1990). Hierarchical spiral model for information system and software development. Part 1: Theoretical background. *Information and Software Technology*, *32*(6), 386–399. https://doi.org/10.1016/0950-5849(90)90125-B

Kemdikbud. (2015). *Rencana Strategis (Renstra) Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi*. https://www.kemdikbud.go.id/main/tentang-kemdikbud/rencana-strategis-renstra

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2022). *Pedoman Penerapan Kurikulum Dalam Rangka Pemulihan Pembelajaran | JDIH Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi*. https://jdih.kemdikbud.go.id/detail_peraturan?main=3022

Ken Warwicki, A. N. (2014). Evaluation of Industrial Policy: Methodological Issues and Policy Lessons | OECD Science, Technology and Industry Policy Papers | OECD iLibrary. In *OECD Science, Technology and Industry Policy Papers* (p. 85). https://www.oecd-ilibrary.org/science-and-technology/evaluation-of-industrial-policy_5jz181jh0j5k-en

Kennedy, E., & Laurillard, D. (2011). *Online Learning Futures: An Evidence Based Vision for Global Professional Collaboration on Sustainability*. Bloomsbury Academic.

Kortmann, S., Perols, J., & Zimmermann, C. (2023). The Theoretical Case of Agile Ambidexterity. *Open Journal of Business and Management*, *11*(04), 1854–1864. https://doi.org/10.4236/ojbm.2023.114103

Kusumaningrum, S. R. (2023). *Teaching English for Elementary Students During and Post-pan...* https://doi.org/10.18502/kss.v8i8.13297

Laborda, J. G. (2010). Contextual clues in Semi-direct interviews for computer assisted language testing. *Procedia - Social and Behavioral Sciences*, *2*(2), 3591–3595. https://doi.org/10.1016/j.sbspro.2010.03.557

M.D. Reckase. (2009). *Multidimensional Item Response Theory*. https://link.springer.com/book/10.1007/978-0-387-89976-3

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Myers, J. P. (2021). Creating the digital citizen: Students' co-construction of meaning for global citizenship during online discussions. *Asian Education and Development Studies*, *11*(4), 592–605. https://doi.org/10.1108/AEDS-09-2020-0218

Nilsson, A., & Wilson, T. L. (2012). Reflections on Barry W. Boehm's "A spiral model of software development and enhancement." *International Journal of Managing Projects in Business*, *5*(4), 737–756. https://doi.org/10.1108/17538371211269031

OECD. (2018). *PISA 2018 Results WHAT STUDENTS KNOW AND CAN DO VOLUME I*. https://www.oecd.org/en/publications/pisa-2018-results-volume-i_5f07c754-en.html

OECD & ADB. (2015). Education in Indonesia: Rising to the Challenge. In *Far Eastern Survey* (Vol. 20). http://www.adb.org/sites/default/files/publication/156821/education-indonesia-rising-challenge.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment.*

Pressman, R. S. (2005). *Software Engineering: A Practitioner's Approach.* Palgrave Macmillan.

Robert C. Martin. (2008). *Clean Code: A Handbook of Agile Software Craftsmanship.* https://www.amazon.com/Clean-Code-Handbook-Software-Craftsmanship/dp/0132350882

Robert Glaser & Edward Silver. (1994). *Chapter 9: Assessment, Testing, and Instruction: Retrospect and Prospect.* https://journals.sagepub.com/doi/10.3102/0091732X020001393?icid=int.sj-abstract.citing-articles.57

Ronald K Hambleton. (2006). *The Next Generation of the ITC Test Translation and Adaptation Guidelines | European Journal of Psychological Assessment.* https://econtent.hogrefe.com/doi/10.1027//1015-5759.17.3.164

S. de Koster & E. Kuiper, M. Volman. (2011). *Concept-guided development of ICT use in 'traditional' and 'innovative' primary schools: What types of ICT use do schools develop?* https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2729.2011.00452.x

Sabet, S. M., & Brown, A. N. (2018). Is impact evaluation still on the rise? The new trends in 2010–2015. *Https://Doi.Org/10.1080/19439342.2018.1483414*, *10*(3), 291–304. https://doi.org/10.1080/19439342.2018.1483414

Salehudin, M., & Sada, H. J. (2020). PENGGUNAAN MULTIMEDIA BERBASIS TEKNOLOGI BAGI PENDIDIKAN PROFESI GURU (PPG): ANALISIS USER EXPERIENCE (UX). *Al-Tadzkiyyah: Jurnal Pendidikan Islam*, *11*(1), Article 1. https://doi.org/10.24042/atjpi.v11i1.5857

Seçken, N. (2010). Identifying Student's Misconceptions about SALT. *Procedia - Social and Behavioral Sciences*, *2*(2), 234–245. https://doi.org/10.1016/j.sbspro.2010.03.004

Shagholi, R., Hussin, S., Siraj, S., Naimie, Z., Assadzadeh, F., & Moayedi, F. (2010). Current thinking and future view: Participatory management a dynamic system for developing organizational commitment. *Procedia - Social and Behavioral Sciences*, *2*(2), 250–254. https://doi.org/10.1016/j.sbspro.2010.03.006

Siyoto, S., & Sodik, M. A. (2015). *Dasar Metodologi Penelitian.* Literasi Media Publishing.

Susan E. Embretson, Steven P. (2000). *Item Response Theory for Psychologists.* https://www.taylorfrancis.com/books/mono/10.4324/9781410605269/item-response-theory-psychologists-susan-embretson-steven-reise

Sutisna, A., Tijari, A., & Irvansyah, A. (2021). Pelatihan Media Pembelajaran Berbasis Android Bagi Tutor Pendidikan Kesetaraan Pada Pkbm Di Kecamatan Sukamakmur Kabupaten Bogor Jawa Barat. *Sarwahita*, *18*(02), Article 02. https://doi.org/10.21009/sarwahita.182.4

Tashakkori, A., & Teddlie, C. (2010). *SAGE Handbook of Mixed Methods in Social &amp;amp; Behavioral Research.* SAGE Publications, Inc. https://doi.org/10.4135/9781506335193

UNESCO. (2014). *Global citizenship education: Preparing learners for the challenges of the 21st century.* https://unesdoc.unesco.org/ark:/48223/pf0000227729

UNESCO. (2021). *Reimagining our futures together: A new social contract for education.* https://unesdoc.unesco.org/ark:/48223/pf0000379707

Voogt, J., Knezek, G., Cox, M., Knezek, D., & ten Brummelhuis, A. (2013). Under which conditions does ICT have a positive effect on teaching and learning? A Call to Action. *Journal of Computer Assisted Learning*, *29*(1), 4–14. https://doi.org/10.1111/j.1365-2729.2011.00453.x

Vygotsky, L. S. (1980). *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press.

Walton, J., & Martin, J. L. (2025). Applying Sadler's principles in holistic assessment design: A retrospective account. *Taylor & Francis.* https://www.tandfonline.com/doi/abs/10.1080/13562517.2023.2244439

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, *21*(4), 361–375.

Wim J. van der Linden & Cees A.W. Glas. (2010). *Elements of Adaptive Testing.* https://link.springer.com/book/10.1007/978-0-387-85461-8

Yin, R. K. (2017). *Case Study Research and Applications: Design and Methods.* SAGE Publications.

Younyoung Choi & Robert J. Mislevy. (2022). *Evidence centered design framework and dynamic bayesian network for modeling learning progression in online assessment system.* https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.742956/full

Zhang, Q., Zhang, T., & Ma, L. (2023). Human acceptance of autonomous vehicles: Research status and prospects. *International Journal of Industrial Ergonomics*, *95*, 103458. https://doi.org/10.1016/j.ergon.2023.103458

Zheng, B., Chang, Y.-H., Good, W. F., & Gur, D. (1997). Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. *Academic Radiology*, *4*(7), 497–502. https://doi.org/10.1016/S1076-6332(97)80236-X