การพัฒนาการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์เพื่อประเมินเมตาคอคนิชัน ของนักศึกษาในสภาพแวดล้อมการเรียนรู้แบบเปิดและทางไกล

Developing a Computerized Adaptive Test to Assess Students' Metacognition in Open and Distance Learning Environment

อนุสรณ์ เกิดศรี[1] และ ธัญสินี เล่าสัม[1*]
Anusorn Koedsri [1] and Thanyasinee Laosum [1*]

(Received: April 8, 2023; Revised: June 2, 2023; Accepted: June 2, 2023)

Abstract

This study assessed the adaptability of the MetaCAT system, which was developed on the Concerto platform for measuring metacognition, in an open and distance learning (ODL) context. In 2022, 70 undergraduate students, each with basic computer skills, from an open university participated in the study.

The findings, based on three indicators, provided evidence of the system's effectiveness in assessing metacognition among students of ODL. The first indicator assessed the relationship between the test taker's ability estimate and the difficulty of the items administered to them. A Pearson's correlation coefficient of 0.70 was obtained, denoting a significant correlation between ability estimates and test difficulty. This result was considered acceptable within the scope of the study. The second indicator evaluated the ratio of the standard deviation of item difficulty to the standard deviation of ability estimates, and was found to be 1.32. This suggests an imbalance in item difficulty parameters. The third indicator, which estimated the reduction in the variance of item difficulty parameters, yielded a value of 0.61. This value, albeit slightly lower than the results of Reckase et al.'s simulation study, still fell within the acceptable levels of fit for the MetaCAT system.

[1] ผู้ช่วยศาสตราจารย์ สำนักทะเบียนและวัดผล มหาวิทยาลัยสุโขทัยธรรมาธิราช
[1] Assistant professor, Office of Registration, Records and Evaluation, Sukhothai Thammathirat Open University
ได้รับทุนวิจัยสร้างองค์ความรู้เพื่อพัฒนาประเทศ
Research Grants to Create Knowledge for Country Development
* Corresponding Author E-mail: thanyasinee.lao@stou.ac.th

The study demonstrated that MetaCAT has satisfactory adaptability and reliability for measuring metacognition in ODL settings, as supported by indicators 1 and 3. Although minor deviations were observed in indicators 2 and 3, suggesting potential areas for refinement, future research could yield a more precise adaptability assessment by using a larger pool of questions and by increasing the sample size.

Keywords: computerized adaptive testing, Concerto platform, metacognition, open and distance learning environment

บทคัดย่อ

การศึกษานี้มุ่งประเมินความสามารถในการปรับเหมาะของระบบ MetaCAT ซึ่งพัฒนาขึ้นบนแพลตฟอร์มคอนแชร์โต้สำหรับการวัดเมตาคอคนิชันในบริบทการเรียนรู้แบบเปิดและทางไกล (open and distance learning: ODL) ในปีการศึกษา พ.ศ. 2565 นักศึกษาระดับปริญญาตรี จำนวน 70 คน จากมหาวิทยาลัยเปิด ซึ่งแต่ละคนมีทักษะในการใช้คอมพิวเตอร์ขั้นพื้นฐานได้เข้าร่วมในการศึกษาครั้งนี้

ข้อค้นพบจากตัวบ่งชี้ทั้ง 3 ตัว แสดงหลักฐานเกี่ยวกับประสิทธิภาพของระบบในการวัดเมตาคอคนิชันของนักศึกษาในบริบทการเรียนรู้แบบเปิดและทางไกล ตัวบ่งชี้แรกประเมินความสัมพันธ์ระหว่างค่าประมาณความสามารถของผู้สอบและค่าพารามิเตอร์ความยากของคำถามที่ผู้สอบแต่ละคนได้รับ ซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันเท่ากับ 0.70 แสดงถึงความสัมพันธ์ที่มีนัยสำคัญระหว่างค่าประมาณความสามารถและค่าความยากของข้อคำถาม ผลลัพธ์นี้ถือว่ายอมรับได้ในขอบเขตของการศึกษานี้ ตัวบ่งชี้ที่สองประเมินอัตราส่วนของส่วนเบี่ยงเบนมาตรฐานของความยากของคำถามกับส่วนเบี่ยงเบนมาตรฐานของค่าประมาณความสามารถของผู้สอบ พบว่าเท่ากับ 1.32 แสดงให้เห็นความไม่สมดุลของค่าพารามิเตอร์ความยากของคำถามกับความสามารถของผู้สอบ ตัวบ่งชี้ที่สามซึ่งประเมินการลดลงของความแปรปรวนของพารามิเตอร์ความยากของคำถาม มีค่าเท่ากับ 0.61 ค่าที่ได้นี้แม้ว่าจะต่ำกว่าผลการศึกษาในสถานการณ์จำลองของ Reckase และคณะเพียงเล็กน้อย แต่อยู่ในระดับที่ยอมรับได้สำหรับระบบ MetaCAT

ผลการศึกษาแสดงให้เห็นว่า MetaCAT มีความสามารถในการปรับเหมาะและความน่าเชื่อถือที่น่าพึงพอใจสำหรับการวัดเมตาคอคนิชันในบริบทการศึกษาแบบเปิดและทางไกล ซึ่งมีหลักฐานสนับสนุนจากตัวชี้วัดที่ 1 และ 3 แม้ว่าจะมีการเบี่ยงเบนเล็กน้อยที่สังเกตได้ในตัวชี้วัดที่ 2 และ 3 ที่บ่งชี้เป้าหมายที่เป็นไปได้สำหรับปรับปรุงด้วยการวิจัยในอนาคตซึ่งอาจทำให้ได้ผลการประเมินความสามารถในการปรับเหมาะมีความแม่นยำมากขึ้นโดยใช้คลังข้อสอบที่ใหญ่ขึ้นและการเพิ่มขนาดตัวอย่าง

คำสำคัญ: การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ แพลตฟอร์มคอนแชร์โต้ เมตาคอคนิชัน สภาพแวดล้อมการเรียนรู้แบบเปิดและทางไกล

Introduction

Metacognition, or the awareness and understanding of one's own thought processes, has been identified as an important aspect of learning. In the context of an open and distance learning (ODL) environments, where students may have different levels of prior knowledge, learning styles, and engagement with the material, assessing and improving metacognitive awareness becomes particularly important (Schraw & Dennison, 1994). However, assessing metacognition in an ODL environments presents unique challenges that require innovative assessment techniques. Traditional methods such as paper-and-pencil tests are often inadequate for measuring metacognition in an ODL environments. Therefore, developing a computerized adaptive test (CAT) system could provide a more accurate and efficient means of assessing students' metacognition in an ODL environments (Wainer & Mislevy, 2019).

Adaptive testing is a promising approach to measuring metacognitive awareness in an ODL environments. This method involves adjusting the difficulty of test items to match the skill level of the test-taker, allowing for a more efficient and accurate measurement of the construct being assessed (Chiu et al., 2021). Additionally, the use of item response theory (IRT) can further enhance the accuracy of the adaptive testing approach, as it allows for the estimation of an individual's metacognitive ability based on their responses to test items (Reckase, 2009). In developing a CAT system for measuring student metacognition in an ODL environments, it is important to ensure that the test is valid and reliable in measuring metacognitive awareness. According to the research conducted by Reckase et al. (2018) and Reckase et al. (2019), the use of a singular indicator cannot fully encapsulate all pertinent information pertaining to the degree of adaptation present in CAT. A comprehensive assessment of such adaptation must also consider multiple indicators, including but not limited to: the correlation between the mean item difficulty parameter and the final ability estimates for examinees, the ratio of the spread in item difficulty to that of examinee ability estimates, and the proportion of variance reduction in item difficulty parameters. Therefore, it is important to consider multiple indicators for ensure that the test is valid and reliable in measuring the targeted construct. Therefore, assessing the adaptability of a test is an important aspect of test development, particularly in the context of an ODL environments.

Assessing the adaptability of a test is important in test development, especially in the context of an ODL environments, because it allows the test to be tailored to the individual needs and abilities of each student. In an ODL, students may have different levels of prior knowledge, different learning styles, and different levels of engagement with the material,

making it important to have a test that can adapt to these differences. By measuring the adaptability of the test, developers can ensure that the test is providing accurate and reliable measurements of the construct being assessed for all students, regardless of their individual differences. This can lead to more effective and efficient learning outcomes, as students are more likely to engage with and benefit from a test that is tailored to their needs.

This study aims to contribute to the literature on adaptive testing and metacognition in an ODL environments by developing a CAT system that evaluates the adaptability of the test based on Reckase's framework (Reckase et al., 2018; Reckase et al., 2019). The potential of the CAT system for revolutionizing the measurement and improvement of metacognitive awareness is substantial. This innovative approach can lead to more accurate and efficient measurements of this important construct, which may have implications for educators and instructional designers seeking to develop more effective and efficient ways of assessing and improving metacognitive awareness in an ODL environments. Thus, the results of this study have the potential to positively impact the field of education, particularly in an ODL environments, by providing a more precise and adaptive measure of metacognitive awareness.

Open and distance learning environments

An ODL environments are becoming increasingly popular, offering learners the flexibility to study at their own pace and in their own environment. However, learning in such environments also presents unique challenges, such as the lack of face-to-face interaction with instructors and peers, which can lead to feelings of isolation and reduced motivation. To overcome these challenges, learners in an ODL environments must have strong metacognitive awareness. Metacognition is defined as the ability to think about one's own thinking, including knowledge about one's own cognitive processes, problem-solving strategies, and regulation of learning. In an ODL environments, learners must be able to self-regulate their learning, monitor their progress, and adjust their strategies as needed. Research has shown that learners who have strong metacognitive awareness are better equipped to succeed in an ODL environments (Arbaugh & Duray, 2002; Joo et al., 2011). Furthermore, the development of metacognitive awareness can be facilitated through instructional strategies such as self-reflection, self-evaluation, and peer feedback. Thus, metacognition is an essential aspect of learning in an ODL environments, and educators should focus on developing and assessing learners' metacognitive awareness in order to promote success and engagement in these contexts. Furthermore, metacognition is particularly important in an ODL environments because of the potential for isolation and lack of support. Without the benefit of regular face-to-face

interactions with instructors and peers, students in an ODL may be more likely to struggle with motivation and engagement (Moore & Kearsley, 2012). However, by developing metacognitive awareness, students can become more self-aware and self-regulated, which can help them to stay on track and remain engaged with their coursework.

Metacognition

Metacognition, defined as the ability to monitor and control one's own cognitive processes and regulate learning, is a crucial aspect of learning and cognitive development, particularly in the context of an ODL environments. An ODL requires students to be self-regulated and self-directed learners, making metacognitive awareness increasingly important (Pintrich, 2002). Research has shown that metacognitive awareness plays a significant role in academic achievement, problem-solving, decision-making, and self-regulation (Flavell, 1979; Zimmerman, 2000). Thus, understanding and assessing metacognition is essential for effective instruction and learning in an ODL environments.

In an ODL environment, students must take greater responsibility for their own learning, and metacognitive awareness can support them in doing so effectively (Schraw & Dennison, 1994). For instance, students who are aware of their own thinking processes can monitor their understanding of course material and identify areas where they need to focus their efforts, seek out additional resources, or request assistance when necessary (McCombs & Marzano, 1990). Metacognitive strategies such as planning and self-evaluation can also help students manage their time and resources more effectively, leading to increased productivity and better academic outcomes (Pintrich, 2002). Moreover, in an ODL environments, where students may experience isolation and lack of support due to limited face-to-face interactions with instructors and peers, developing metacognitive awareness can make students more self-aware and self-regulated, which can help them stay motivated and engaged with their coursework (Moore & Kearsley, 2012).

Assessing metacognition in an ODL environments can be achieved through various methods, such as self-report measures, think-aloud protocols, and performance-based tasks. Schraw & Dennison (1994) developed a self-report questionnaire, the Metacognitive Awareness Inventory (MAI), which assesses three components of metacognitive awareness: declarative knowledge, procedural knowledge, and conditional knowledge. Declarative knowledge refers to an individual's understanding of their own cognitive abilities, strategies, and resources for learning, while procedural knowledge involves the skills and strategies used to regulate thinking and learning, and conditional knowledge entails the ability to adapt learning strategies

to different contexts and tasks. The MAI consists of 52 items that measure these components, and respondents rate their agreement with statements on a Likert scale. The MAI has shown good reliability and validity (Schraw & Dennison, 1994; Schraw, 1997).

Computerized adaptive testing

Computerized Adaptive Testing (CAT) is an innovative approach to assessment that uses technology to dynamically adjust the difficulty of test items based on the individual test-taker's skill level. It has gained attention in education and measurement due to its potential to provide more accurate and efficient assessments compared to traditional fixed-item tests. CAT utilizes item response theory (IRT) to estimate the examinee's ability level based on their responses to previous questions (van der Linden, 2016). CAT has been successfully applied in various domains and continues to be an area of active research and development. In an ODL environments, CAT has significant benefits for assessing metacognition. It provides tailored assessments that are responsive to the unique learning needs of individual learners, reduces test administration time, minimizes test fatigue, and enhances test security. The data generated by CAT can also be used to inform instructional strategies, curriculum design, and learner support services, making it a valuable tool for educators and administrators in an ODL environments. Overall, CAT in an ODL environments has the potential to significantly enhance the assessment of metacognitive awareness and improve learning outcomes for learners in an ODL programs.

Polytomous item response theory

Polytomous Item Response Theory models, including the Graded Response Model (GRM), have become increasingly popular in educational and psychological research. The GRM is a type of IRT model that is used to analyze data from items with more than two response categories. The model assumes that the probability of responding in a particular category is a function of the underlying trait being measured and the item parameters, which describe the relationship between the trait and each response category. The GRM has several advantages over other polytomous IRT models, including the ability to model complex item structures and the ability to estimate item parameters with small sample sizes. Additionally, the GRM has been successfully implemented in CAT systems, which allow for the efficient and accurate measurement of latent traits. In a CAT system, items are selected based on the respondent's estimated trait level, resulting in a tailored and efficient testing experience. The GRM has been shown to perform well in CAT simulations and has been used in various CAT applications (Choi

& Swartz, 2011; Reckase, 2009). Overall, the GRM is a powerful tool for analyzing data from items with multiple response categories and has the potential to greatly enhance the efficiency and accuracy of measurement in educational and psychological research (Baker, 2001).

The GRM has several advantages over other polytomous IRT models, including the ability to model complex item structures and the ability to estimate item parameters with small sample sizes. Additionally, the GRM has been successfully implemented in CAT systems, which allow for the efficient and accurate measurement of latent traits. In a CAT system, items are selected based on the respondent's estimated trait level, which results in a tailored and efficient testing experience. The GRM has been shown to perform well in CAT simulations and has been used in various CAT applications (Choi & Swartz, 2011; Reckase, 2009). Overall, the GRM is a powerful tool for analyzing data from items with multiple response categories and has the potential to greatly enhance the efficiency and accuracy of measurement in educational and psychological research.

Concerto platform

Concerto, an open-source platform, was designed to create secure and user-friendly patient-reported assessments that utilize CAT and machine learning. It is accessible on various devices and requires no prior programming experience due to its point-and-click interface. The platform interacts with back-end functions, including scoring, CAT, and machine learning algorithms, using the R programming language, which has over 15,000 available packages for statistical computing tasks. Concerto allows for non-adaptive assessments or CATs using pre-written R code or item parameter tables uploaded by the user. Patient data is securely stored using the MySQL database management system, and the platform can be installed on cloud-based or local servers to comply with institutional security protocols. Concerto provides immediate results with feedback presented through graphs and text, enhancing the assessment experience. Results can be directly imported into electronic health records via application programming interfaces (Harrison et al., 2020).

Objectives

The study aimed to assess the adaptability of the computerized adaptive testing system in measuring metacognitive awareness in an Open and Distance Learning environments, based on Reckase's framework (Reckase et al., 2018; Reckase et al., 2019).

Method

1. Population and sample

The population of interest for this study consisted of undergraduate students who were enrolled in an ODL institution in the academic year 2022 and possessed basic computer skills. From this population, a sample of 70 voluntary research participants was selected to measure the degree of adaptation in the Computerized Adaptive Testing (CAT) system. These participants were chosen based on their willingness to participate in the study and provided informed consent. They also had the right to refuse or withdraw from the research at any point without facing any negative consequences.

2. Research instrument

The research instrument used in this study can be categorized into two parts: assessment tools and the MetaCAT system.

2.1 Assessment tools

The assessment tool utilized in this study was the Metacognition Computerized Adaptive Testing (MetaCAT) system. It was adapted from a previous study by Ngudgratoke et al. (2016), which focused on developing measurement and evaluation tools for metacognition in primary and secondary school students. The assessment tool consisted of a set of 52 items with five response categories, ranging from "1" (never) to "5" (always). It was grounded in the theoretical framework proposed by Schraw & Dennison (1994) and consisted of two factors: knowledge of cognition and regulation of cognition. To ensure the reliability of the assessment tool, a sample of 32 participants, representative of students enrolled in an ODL program, underwent a Cronbach's alpha reliability coefficient analysis. The resulting Cronbach's alpha coefficient showed a high level of internal consistency among the items (α = 0.94).

One of the main advantages of using CAT is that it allows for the administration of item banks with a relatively small number of items. According to CATs based on PROMIS item banks typically require respondents to complete only 3-7 items to obtain a reliable score, compared to 20-30 items with traditional questionnaires. This suggests that the 52-item bank used in the MetaCAT system is sufficient for assessing metacognitive awareness, as respondents can obtain reliable scores with a relatively small number of items. Additionally, a previous study by Koedsri et al. (2020) in the Preliminary Study of the Development of Computerized Adaptive Testing on Metacognition for Primary and Secondary School Students found that the estimate of metacognitive ability obtained from the full test (52 items) and simulated response using CAT (with an average of about 16 items) had a coefficient correlation

in the range of 0.90-0.94. This indicates that respondents can assess their metacognitive level through self-assessment using CAT and obtain similar results to the full test. These findings further support the sufficiency of the 52-item bank in measuring metacognitive awareness in the ODL context.

When determining the appropriate sample size for estimating item parameters using the Graded Response Model (GRM), Jiang et al. (2016) conducted a study titled "Performance of Polytomous IRT Models With Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data." Their findings concluded that a minimum sample size of 250 to 500 examinees is necessary to maintain reasonable root mean square errors (RMSEs). However, their research also suggested that even smaller sample sizes as low as 250 can still provide adequate recovery of item parameters for calibration samples in the GRM. Following these recommendations, the current study utilized a sample size of 455 students, carefully selected from a comprehensive list of students undergoing professional experience training at the university. The larger sample size was chosen to ensure the precision and reliability of the estimated item parameters in the GRM, aiming to enhance accuracy, facilitate robust analyses, and increase the generalizability of the study findings to a wider population of interest.

The analysis of the 52 item difficulties or thresholds (B) revealed a distribution spanning both negative and positive ranges (-7.09 to 3.23). The item with the lowest threshold of difficulty is "I ask myself periodically if I am meeting my goals" [B1:B4; -7.09 -4.18 -0.54 2.44]. The values of the threshold difficulty increase in ascending order, corresponding to the five response categories, ranging from "1" (never) to "5" (always). Lower scores indicated lower levels of metacognition, while higher scores indicated higher levels of metacognition. The analysis of the discrimination parameter (a) indicated varying degrees of discriminating power, ranging from 0.83 to 1.76, which represents moderate to high to very high discrimination. This finding highlights the effectiveness of the measurement instrument in evaluating metacognition among ODL students. Furthermore, the wide distribution of metacognitive information, ranging from -3.00 to 2.50 with associated information values, supports the appropriateness of the scale for assessing metacognitive abilities in the context of ODL, as illustrated in Figure 1.

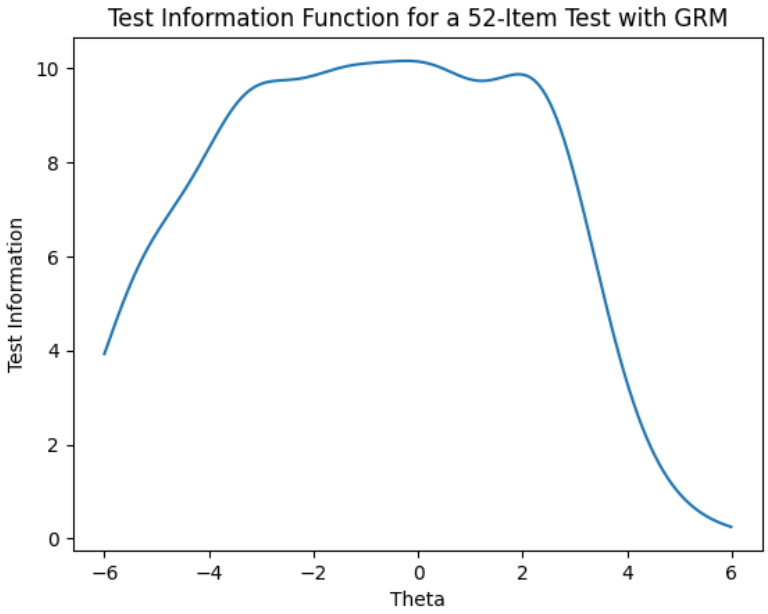

Figure 1 Test information function of metacognition scale

2.2 MetaCAT

The MetaCAT system was developed to assess students' metacognitive awareness in an ODL environment. It was designed using the Concerto platform, an open-source software that allows for the administration of computerized tests in a controlled online environment. The installation of the MetaCAT system followed the latest guidance available on the Concerto GitHub webpage (https://github.com/campsych/ concerto-platform/wiki). A CSV file named "CESDFlatItems.csv" was downloaded from the Open Science Framework (https://osf.io/m3wkc). This file was referred to as a "flat" item table because all the data was stored in one layer.

To create a new assessment and table for storing item responses, the researchers logged into the Concerto platform and selected "Add new" under the Tests tab. The researchers set the assessment name to "MetaCAT_adaptive" and the Type dropdown box to "flowchart." A new test was created under the Tests tab, and a new table to store responses was saved by editing the assessmentResponses table's name. The researchers uploaded the item bank and customized their CAT by selecting "Flat Table" from the Type dropdown menu under the Assessment node wizard's Items tab. They set stopping rules and psychometric model parameters for their CAT, such as GRM. The responses table was also selected, and the scores' uncertainty was calculated and reported.

Finally, the test start node was connected to the assessment node, and return ports were created by checking the boxes next to Theta and SEM, as shown in Figure 2 It is

worth noting that users who are not familiar with Amazon Web Services (AWS) should be careful when using the service, as it requires submitting credit or debit card details during the registration process.

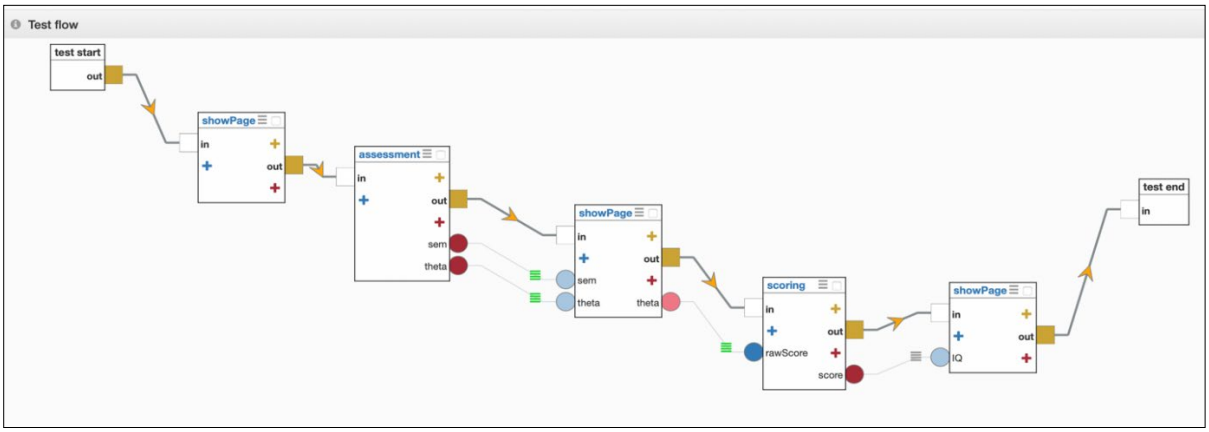

Figure 2 Connecting nodes in Concerto Platform

3. Data collection

The MetaCAT system was programmed and implemented using the Concerto platform, which allows for the administration of computerized tests in a controlled online environment. The adaptive testing algorithm dynamically selects and administers test items based on the examinee's ability level, maximizing measurement precision with minimal test length. The adaptive testing algorithm was calibrated using item parameters estimated from the metacognitive awareness assessment scale. The researchers ensured the confidentiality of the participants' information throughout the research period and disclosed only the results of the analysis without revealing their names. Participants had the right to decline or withdraw from the research at any time without facing any consequences.

4. Data analysis

Reckase et al. (2018) proposed three statistical indicators to measure the degree of adaptation in CAT, which is a testing approach that tailors test items for each examinee during the testing process. However, when the item bank is limited in size, content constraints are stringent, or strong exposure control is enforced, adaptation may be restricted. Therefore, the proposed indicators aim to quantify the extent to which adaptation is taking place in CAT.

The first indicator is the correlation between the mean parameter and the final estimates for the examinees, $r\left(\overline{b}_j, \hat{\theta}_j\right)$. The parameter is a measure of item difficulty (b), while the theta ($\hat{\theta}$) estimate is a measure of the examinee's ability. A high correlation between the two indicates that the item selection algorithm is working properly.

The second indicator used in the study is the ratio of the standard deviation of the mean difficulty parameters (b) for the administered items to the standard deviation of the ability estimates for the test-taker. This indicator measures the spread in difficulty of the administered items among a test-taker, $(s_{b_j} / s_{\hat{\theta}_j})$. This indicator accounts for the spread in difficulty of the items administered to a test-taker.

The third indicator measures the extent to which the MetaCAT reduces the variance (PRV) in item difficulty parameters for the selected items compared to the total variance in the item bank. This indicator gauges the degree to which the adaptive testing method effectively confines the spectrum of difficulty levels for the test-takers. The proportional calculation of this indicator is formulated as follows

$$\mathrm{PRV} = \frac{s_b^2 - \text{ pooled } s_{b_j}^2}{s_b^2}$$

where s_b^2 is the variance of the item difficulty parameters for all items in the bank and $pooled\ s_{b_j}^2$ is the variance of the item difficulty parameters for the items selected for examinees. The PRV measures the extent to which the test is adapting by selecting items that have different difficulty levels than the average difficulty level of all items in the bank. A larger PRV indicates that the test is adapting more by selecting items that are more different in difficulty level from the average difficulty level of all items in the bank.

Results

The study aimed to assess the adaptability of the MetaCAT system in measuring metacognitive awareness in an ODL environments, utilizing Reckase's framework (Reckase et al., 2018) The descriptive statistics presented in Table 1 demonstrate that the CAT system effectively measured the participants' metacognitive awareness in an ODL environment. The variables included in the table are TimeTaken, Theta, SEM, B1, B2, B3, and B4. ‘TimeTaken’ denotes the duration taken by participants to complete each item, ‘Theta’ represents the participants' estimated ability level, ‘SEM’ refers to the standard error of measurement, and ‘B1’, ‘B2’, ‘B3’, and ‘B4’ denote the thresholds of item difficulty in the GRM.

The mean time taken by participants to complete each item in the test was 11.76 seconds, with a standard deviation of 13.11. The mean theta was 1.12, with a standard deviation of 1.08. The mean standard error of measurement (SEM) was 0.50. The mean threshold values for item difficulty (B1, B2, B3, and B4) ranged from -5.25 to 2.55, indicating moderate overall difficulty of the items. These results suggest that the CAT system was able to effectively measure participants' metacognitive awareness in an ODL environments, as illustrated in Table 1.

Table 1 Descriptive statistics of variables in computerized adaptive testing system for measuring metacognitive awareness in an ODL environments

	Mean	SD	Median	Minimum	Maximum
TimeTaken	11.76	13.11	8.00	1.00	138.00
Theta	1.12	1.08	1.14	-3.52	3.42
SEM	0.50	0.15	0.45	0.30	0.95
B1	-5.25	0.47	-5.38	-7.09	-3.59
B2	-3.26	0.51	-3.36	-5.16	3.01
B3	0.20	0.53	0.24	-1.40	0.99
B4	2.55	0.38	2.47	1.59	3.17

Note. TimeTaken is measured in seconds, SD = standard deviation, SEM = standard error of measurement, B1, B2, B3, and B4 = thresholds of item difficulty in the GRM

The MetaCAT system is designed to evaluate students' metacognitive awareness in ODL environment. The item bank for this study consisted of 52 items. Table 2 presents the results obtained from the study, with variables such as session_id, N.Items, Theta, and SEM.

The data obtained from this study reveals that the respondents demonstrated a broad range of abilities, as evidenced by the Theta values, which varied from -3.77 to 3.42. On average, each respondent completed 14 items, with a standard deviation of 4.33. The standard error of measurement varied between 0.26 and 0.45, with an average of 0.34.

Overall, the MetaCAT system appears to be an effective tool for evaluating students' metacognitive awareness in ODL environments, as evidenced by the wide range of abilities and standard error of measurement values obtained. Further analysis and interpretation of the data may provide valuable insights into the effectiveness of the MetaCAT system and the metacognitive awareness of the respondents, as illustrated in Table 2.

Table 2 Descriptive statistics for metacognitive awareness measures in MetaCAT system

	N Items	Theta	SEM
M	13.63	0.82	0.34
SD	4.33	1.95	0.02
Min	4.00	-3.77	0.26
Max	28.00	3.42	0.45

Note. N Items = number of items on the test, Theta = the ability estimate, SEM = standard error of measurement, M = mean, SD = standard deviation, Min = minimum, Max = maximum

The first indicator, based on the analysis of adaptive measures in computer adaptive tests designed to assess metacognition in distance learning students of open universities, reveals a significant correlation (0.70) between the estimated ability of test takers and the difficulty of administered items for each examinee, as illustrated in Figure 3.

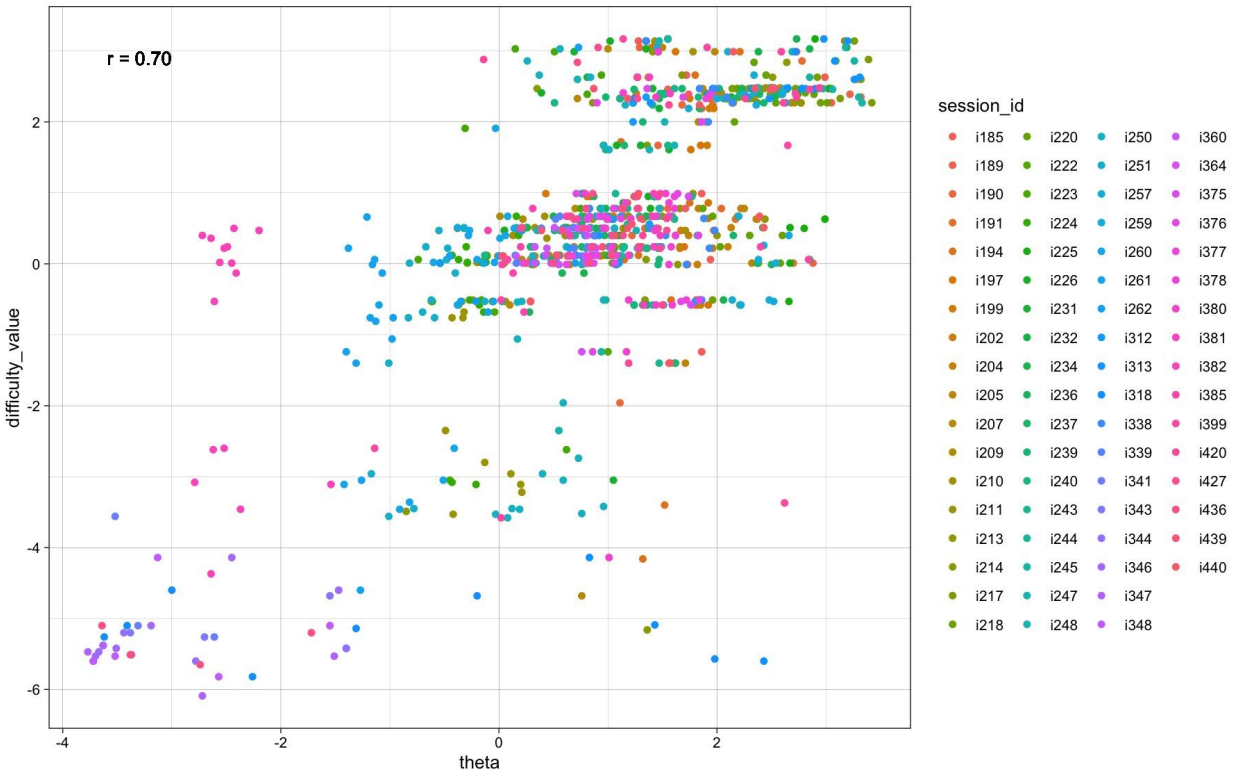

Figure 3 The correlation between test takers' ability estimates and item difficulty

Figure 3 employs different colors for dots to represent individual respondents (session_id), with the horizontal axis indicating the ability estimate (theta) of the respondents, and the vertical axis denoting the difficulty of the items (difficulty_value).

The second indicator, which is the ratio of the standard deviation of the difficulty level of the administered items to the standard deviation of the ability estimates for the examinees, was calculated to be 1.32. This value indicates that the difficulty level of the administered items is slightly higher than the variability in the ability estimates of the examinees. According to Reckase et al. (2018) and Reckase et al. (2019), an optimal value for this indicator is 1.0. However, values greater than 1.0 can still be obtained and are not necessarily problematic. It is the distance from 1.0 that is important when interpreting this statistic. In the case of the MetaCAT system, the ratio of 1.32 suggests that there may be a slight imbalance in the distribution of difficulty levels, with a slightly higher concentration of difficult items compared to the variability in the ability estimates of the examinees. This could

be due to the item bank having a few extremely easy and difficult items, but not enough middle-range items. Nonetheless, this slight deviation from 1.0 does not significantly impact the overall reliability and validity of the CAT system in measuring metacognitive awareness in an ODL environment.

The third indicator, which assesses the reduction in variance of difficulty parameters for the items administered to test-takers compared to the variance of difficulty parameters for all items in the item bank, yielded a value of 0.61 in the present research. This statistic provides valuable insight into the effectiveness of the item selection process in reducing the variability of the difficulty parameters. However, when comparing this value to the findings of Reckase et al. (2018) and Reckase et al. (2019) in their simulation study, it appears slightly lower. In their study, they observed that the proportion of within examinee difficulty variance (PRV) statistics were close to zero, indicating that the items selected for each examinee were almost as widely spread as the entire item pool. Moreover, they found that a PRV value of approximately 0.80 indicated a high level of adaptation. Notably, in the case of a 50-item pool, the PRV was reported to be 0.62.

Discussion

The discussion presents the findings of the study, highlighting the adaptability of the MetaCAT system in measuring metacognitive awareness in an ODL environment.

The first indicator, which shows a significant correlation (0.70) between the estimated ability of test takers and the difficulty of administered items for each examinee, supports the reliability of the computer adaptive tests in assessing metacognition in distance learning students of open universities (Reckase et al., 2018). This suggests that the adaptive measures employed in the computer adaptive tests effectively capture the relationship between the ability estimate of test takers and the difficulty of the administered items.

The second indicator, which measures the ratio of the standard deviation of the difficulty level of the administered items to the standard deviation of the ability estimates for the examinees, revealed a value of 1.32 in the present study. This value suggests a slight deviation from the optimal value of 1.0, as proposed by Reckase et al. (2018) and Reckase et al. (2019). It indicates that the difficulty level of the administered items is slightly higher than the variability in the ability estimates of the examinees. While higher values than 1.0 can be obtained, it is important to consider the distance from 1.0 when interpreting this statistic. The deviation from the optimal value in the MetaCAT system may indicate a slight imbalance in the distribution of difficulty levels, potentially stemming from the item bank containing a few

extremely easy and difficult items without sufficient representation of middle-range items. However, it is crucial to note that this slight deviation does not significantly impact the overall reliability and validity of the computer adaptive test (CAT) system in measuring metacognitive awareness in an open and distance learning (ODL) environment.

The third indicator, which assesses the reduction in variance of difficulty parameters for the items administered to test-takers compared to the variance of difficulty parameters for all items in the item bank, yielded a value of 0.61 in the present research. This statistic provides important insights into the effectiveness of the item selection process in reducing the variability of the difficulty parameters. However, when comparing this value to the findings of Reckase et al. (2018) and Reckase et al. (2019) in their simulation study, it appears slightly lower. In their study, they found that the proportion of within examinee difficulty variance (PRV) statistics approached zero, indicating that the items selected for each examinee were nearly as widely spread as the entire item pool. Moreover, they identified a PRV value of approximately 0.80 as indicative of a high level of adaptation. Notably, in the case of a 50-item pool, the PRV was reported to be 0.62. In the present study, the lower PRV value of 0.61 may be attributed to the relatively smaller item pool size of 52 items and the sample of 70 voluntary research participants. It is important to consider that the interpretation of the PRV statistic depends on the specific research context and design. While a higher PRV value suggests a greater level of adaptation, the observed PRV value in the MetaCAT system indicates its effective adaptation to the individual ability levels of the test-takers, resulting in a reduction in the variability of the difficulty parameters. Although the value is slightly lower compared to the simulation study, it still demonstrates an acceptable level of adaptation. However, further investigations utilizing larger item pools may yield higher PRV values and potentially enhance the level of adaptation even further.

Exposure control refers to the ability to control the number of times an item is presented to different examinees in a CAT system. The concern with exposure control is that it may impact the item pool, reducing the number of items available for future testing. However, in the case of a CAT system designed to assess metacognitive awareness, exposure control is not a problem because it is a self-assessment test and low-stakes testing. Self-report measures minimize the potential impact of exposure control on the item pool, as the examinee's responses are based on their personal experiences and perceptions. In addition, low-stakes testing reduces the examinee's motivation to cheat, further reducing the impact of exposure control. Thus, exposure control is not a significant issue in CAT systems designed to measure metacognitive awareness.

In conclusion, the slight deviation from the optimal ratio of difficulty level variability to ability estimate variability in the MetaCAT system may be attributed to the small voluntary participation in the study and the use of the Graded Response Model (GRM) as the item response theory (IRT) model. The smaller sample size and different IRT model used in the present study compared to the simulation study conducted by Reckase et al. (2018); Reckase et al. (2019) and Wyse & McBride (2021) may have influenced the standard deviation of difficulty levels in the administered items. It is possible that the SD of difficulty of items in the study is relatively larger, which could contribute to the observed deviation. However, despite this slight deviation, the overall performance of the MetaCAT system remains reliable and valid in measuring metacognitive awareness in an open and distance learning context.

Limitations

Despite the strengths and contributions of this study, it is important to acknowledge several limitations that may impact the interpretation and generalizability of the findings. First, the item bank size utilized in the study was relatively small, consisting of only 52 items. While extensive analysis was conducted to ensure the reliability and validity of these items, the limited number may restrict the comprehensive assessment of metacognitive awareness. A larger item bank size would offer broader coverage of metacognitive abilities, enabling a more precise measurement of the construct.

Furthermore, the study employed a relatively small sample size of 70 voluntary research participants to assess the degree of adaptation in the Computerized Adaptive Testing (CAT) system. Although efforts were made to recruit a diverse group of undergraduate students enrolled in an ODL institution, the small sample size may limit the generalizability of the findings and the accuracy of the adaptability assessment. A larger sample size would allow for a more robust analysis of the degree of adaptation in the CAT system, providing more reliable estimates of the correlations, ratios, and variances used to evaluate adaptability. Consequently, employing a larger sample size would enhance the accuracy and representativeness of the parameters used to determine the adaptability of the test.

Recommendation

1. Implication for practice

Based on the findings of this study, there are several suggestions for future research in the field of computer adaptive testing and metacognition in an ODL environment:

Ensure a sufficient item bank size: To enhance the comprehensive assessment of metacognitive awareness, it is recommended to increase the size of the item bank beyond the current 52 items. A larger item bank will provide a broader coverage of metacognitive abilities, allowing for a more precise measurement of the construct. This can be achieved by developing and incorporating additional items that assess different aspects of metacognitive awareness in an ODL environment.

Moreover, implement a diverse and representative sample: While efforts were made to recruit a diverse group of undergraduate students, it is important to increase the sample size to improve the generalizability of the findings. A larger sample size will provide more robust estimates of the adaptability of the Computerized Adaptive Testing (CAT) system and its ability to measure metacognitive awareness in an ODL environment. In practice, researchers and practitioners should aim to include a larger and more diverse sample of students from various academic disciplines enrolled in ODL programs.

2. Recommendation for future research

Based on the findings of this study, there are several suggestions for future research in the field of computer adaptive testing and metacognition in an ODL environment:

Firstly, investigate the impact of item pool size: Future research should explore the influence of item pool size on the adaptability of the CAT system in measuring metacognitive awareness. By utilizing a larger item pool, researchers can examine whether the slight deviation in the ratio of difficulty level variability to ability estimate variability can be minimized or eliminated. This will provide valuable insights into the scalability and effectiveness of the MetaCAT system when applied to larger populations.

Secondly, compare different IRT models: To further understand the performance of the CAT system in measuring metacognitive awareness, future research could compare the results obtained using different Item Response Theory (IRT) models. By using alternative models such as the 2-parameter logistic model or the generalized partial credit model, researchers can assess the impact of the chosen IRT model on the variability of difficulty levels in the administered items. This will contribute to a more comprehensive understanding of the measurement properties and accuracy of the MetaCAT system.

Finally, investigate the relationship between Item fifficulty and ability estimates: Future research could explore the relationship between item difficulty and ability estimates in more depth. By analyzing the specific items that contribute to the slight deviation from the optimal ratio, researchers can identify potential factors that may influence the difficulty levels

and refine the item bank accordingly. This will allow for the development of a more balanced and representative set of items that accurately reflect the metacognitive abilities of students.

References

Arbaugh, J.B., & Duray, R. (2002). Technological and structural characteristics, student learning and satisfaction with Web-based courses: An exploratory study of two on-line MBA programs. *Management Learning*, *33*(3), 331-347.

Baker, F.B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Chiu, Y.C., Douglas, J., & Liang, J.C. (2021). Investigating the Effectiveness of Computerized Adaptive Testing for Measuring Metacognition. *Journal of Educational Computing Research*, *59*(5), 1095-1117. https://doi.org/10.1177/0735633120932388

Choi, S.W., & Swartz, R.J. (2011). Comparison of CAT item selection criteria for the graded response model. Educational and Psychological *Measurement, 71*(1), 115-135. https://doi.org/10.1177/0013164410372102

Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906-911.

Harrison, C., Loe, B.S., Lis, P., & Sidey-Gibbons, C. (2020). Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning [Tutorial]. *Journal of Medical Internet Research*, *22*(10), e20950. https://doi.org/10.2196/20950

Jiang, S., Wang, C., & Weiss, D.J. (2016). Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model [Original Research]. Frontiers in Psychology, *7*. https://doi.org/10.3389/fpsyg.2016.00109

Joo, Y.J., Lim, K.Y., & Kim, C.J. (2011). Online university students' satisfaction and persistence: Examining perceived level of presence, usefulness and ease of use as predictors in a structural model. *Computers & Education*, *57*(2), 1654-1664.

Koedsri, A., Na Nakorn, N., & Watanasuntorn, K. (2020). *Preliminary Study of a Development of Computerized Adaptive Testing on Metacognitive for Primary and Secondary School Students.* Paper presented at the 28th Thailand Measurement Evaluation and Research Conference 2020, Faculty of Education Naresuan University.

McCombs, B.L., & Marzano, R.J. (1990). Putting the self in self-regulated learning: The self as agent in integrating will and skill. *Educational Psychologist*, *25*(1), 51-69.

Moore, M.G., & Kearsley, G. (2012). *Distance education: A systems view of online learning*. Cengage Learning.

Ngudgratoke, S., Na Nakorn, N., Chutinuntakul, S., Phonapichat, P., & Sittirit, P. (2016). *The development of a scale to measure and assess metacognition of primary and secondary school* [Research Report]. NIETS. https://www.niets.or.th/th/content/view/5869

Pintrich, P.R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*, *41*(4), 219-225.

Reckase, M.D. (2009). *The basics of item response theory* (2nd ed.). University of Michigan Press.

Reckase, M.D., Ju, U., & Kim, S. (2018). Some measures of the amount of adaptation for computerized adaptive tests. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology*: *The 82nd annual meeting of the Psychometric Society*. Springer International Publishing.

Reckase, M.D., Ju, U., & Kim, S. (2019). How Adaptive Is an Adaptive Test: Are All Adaptive Tests Adaptive?. *Journal of Computerized Adaptive Testing*. Springer International Publishing.

Samejima, F. (1997). Graded response model. In van der Linden, W.J. & Hambleton, R. K. (eds.), *Handbook of modern item response theory (pp.*85–100*)*. Springer.

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education, 65*, 135-146. http://dx.doi.org/10.1080/00220973.1997.9943788

Schraw, G., & Dennison, R.S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460-475. https://doi.org/10.1006/ceps.1994.1033

van der Linden, W.J. (2016). Computerized adaptive testing. *In International Encyclopedia of the Social & Behavioral Sciences.* Elsevier. 366-371.

Wainer, H., & Mislevy, R.J. (2019). *Computerized adaptive testing: A primer* (3rd ed.). Routledge.

Wyse, A.E., & McBride, J.R. (2021). A Framework for Measuring the Amount of Adaptation of Rasch-based Computerized Adaptive Tests. *Journal of Educational Measurement*, *58*(1), 83-103. https://doi.org/10.1111/jedm.12267

Zimmerman, B.J. (2000). Attaining self-regulation: A social cognitive perspective. In Monique Boekaerts, M, Pintrich, P. R. & Zeidner, M. (Eds), *Handbook of self-regulation (pp.13-39).* Elsevier.