IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 January 2023, accepted 18 March 2023, date of publication 6 April 2023, date of current version 12 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3265196

RESEARCH ARTICLE

Objectivity and Subjectivity in Variation of Multiple Choice Questions: Linking the Theoretical Concepts Using Motion in Mind

PUNYAWEE ANUNPATTANA[1], MOHD NOR AKMAL KHALID[1,2], (Member, IEEE), AND HIROYUKI IIDA[1]

[1]School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1211, Japan
[2]School of Computer Sciences, Universiti Sains Malaysia, Georgetown, Penang Island 11800, Malaysia

Corresponding author: Mohd Nor Akmal Khalid (akmal@jaist.ac.jp)

This work was supported by the Japan Society for the Promotion of Science through the Framework of the Grant-in-Aid for Challenging Exploratory Research under Grant 19K22893.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Associate Dean and coordinator staff of Thammasat Secondary School, Thailand.

ABSTRACT Multiple-choice questions (MCQs) have been considerably used for assessing the individual's performance in various contexts. The optimal number of options in MCQs is a debatable issue, followed by contradictions and discussions, and is needed for a firm conclusion from empirical and hypothetical findings. This study aims to link theoretical concepts, including challenge-based gamification, zone of proximal development, and prospect theory, and generate insight into educational assessment using motion-in-mind measures. Classical test theory was used to determine reliability and validity. Variations of MCQs experimented: the number of options, settings, and scoring methods. The experimental data was gathered from human and AI simulations and measured using motion-in-mind. It was found that increasing the number of options in the MCQ makes the test more challenging, explaining an increase of mass in mind m. The findings also revealed that time pressure provides competitiveness while scaffolding provides support. In addition, the hybrid system demonstrates the balance of education and entertainment. Finally, the results addressed the general discussion and analogical interpretation in the education context based on physics-in-mind values. These findings can be promising for analyzing the balance between competitiveness and entertainment while enabling the learning process in the practical assessment.

INDEX TERMS Educational assessment, motion-in-mind, multiple-choice questions, challenge-based gamification, number of options, zone of proximal development, prospect theory.

I. INTRODUCTION

Evaluation and assessment in education are critical markers of educators' performance and achievements. The scholar stated that a test is a measurement instrument aimed to numerically characterize the degree or amount of learning under uniform, standardized conditions [1]. However, it has been challenging to determine the ideal formative assessments for a legitimate goal because a test that is beneficial in one context may be entirely objectionable in another; in other words, there is no one-size-fits-all test format.

There are different types of multiple-choice questions based on the number of options provided in the test, ranging from 2 to, at most, five options. Depending on the outcomes, respondents will be confronted with a more significant number of plausible distractors and a more challenging situation. They can also test factual recall and more advanced cognitive functions such as diagnostic competence, appraisal, and

The associate editor coordinating the review of this manuscript and approving it for publication was Qingli Li.

reasoning [2]. They also encourage the evaluation of a wide range of content. This situation allows for use in various fields, such as proficiency tests, language testing, medical testing, and national examinations.

The MCQ format's major flaws were low validity, testing factual knowledge rather than high-level cognitive knowledge, increased guessing effect (especially as the number of options decreases), and requiring time in test development. As a result, most of this study focuses on challenges assessment, enhancing the reliability and validity of multiple-choice examinations (by removing non-functional distractors), reducing the guessing effect, and eliminating potential sources of confusion. The outcomes of these studies had synthesized best practices to provide practical recommendations for educators [3], [4].

The ongoing issue arises of whether there are optimal options for items in MCQs to utilize and how to apply them correctly to different educational levels and specific contexts. In practice, MCQ usually contains 4 or 5 options per item, particularly in standardized testing according to [3] and [5]. A large body of research investigated how the number of options impacts test quality by approaching the issue from several aspects. There are essential arguments among scholars about the increasing probability of answering an item correctly without sufficient knowledge (i.e., blind random guessing). Many scholars have stated the effect of adjusting the number of options which could be observed from the guessing rate [6], [7], [8]. It should be mentioned that reliability and validity will be improved when guessing is minimized due to the use of more options [1], [9], [10], [11]. As a result, determining the optimal number of options is critical in evaluating learners in the educational context.

Over decades, many practical trials have been identified and repeatedly conducted to consolidate long-term outcomes. One example is incorporating the game elements like gamification and practical techniques to facilitate short-term and long-term learning potential. When using MCQ to emerge potential learning, it is necessary to provide a challenging but manageable test. For this reason, the degree of challenge is vital to realizing this usefulness. Too simple questions can sabotage the positive effect, while too difficult questions can deliver an adverse effect. The zone of proximal development (ZPD) provided by Vygotsky, as cited in [12] and [13], represent the metaphorical gap between what a learner can do independently and what they can do dependently. This condition helps to understand and enable learning far beyond what the learners could achieve individually. Scaffolding is one technique introduced to mitigate frustration over the high challenging tasks by guiding learners to solve the tasks.

Moreover, there are different types of scoring methods in order to improve their psychometric properties and increase both reliability and validity as cited in [14], [15], and [16]. Tests should also be viewed as circumstances in which people generate various uncertain responses. As a result, the characteristics of risk-taking behavioral decision-making processes may shape individual behaviors during tests. This situation can provide perks or drawbacks for certain groups of test takers, thus lowering the validity of the assessments. Thus, the scoring method used in a test creates framing effects that influence the desire to take risks and guess in tests.

For these reasons, the characterization of MCQ is vital for increasing the understanding of objectivity and subjectivity in an educational context. Furthermore, this study will generate fresh insight into the link between MCQ and motion in mind to highlight the analogical interpretation in the education context described later in this study. Based on these premises, this research thus addresses the following questions:

1) How the number of options in the MCQ paradigm be represented with the motion in mind theory?
2) How the challenge-based gamification and scaffolding in the MCQ paradigm impacts the assessment results?
3) How the framing effect in the MCQ paradigm impacts the assessment results?

II. BACKGROUND

A. MULTIPLE CHOICE QUESTIONS

Multiple choice questions (MCQs) have been used in various educational and knowledge assessments, providing a general ground regarding standardization, equitability, and reliability. Moreover, they enable a comprehensive range of instructions and contents that examine factual recall and cognitive functions such as critical skills and reasoning. This situation signifies that the MCQs are suited to summative standardized and national examination [2], [17]. MCQs are composed of a stem/question/problem/task to be responded to or solved and a series of alternative responses or alternatives (possible solutions to the question). Among the options, there was a correct answer (the key), and the rest were wrong or less acceptable answers (the distractors). The essential weakness of this format is that high-quality questions are challenging to construct item stems, plausible and functional of correct options (keys) and (incorrect options (distractors) [8].

Empirical research attempts to gain reliability and validity when MCQs has incorporated to evaluate the test taker based on achievement [18]. This condition utilizes traditional item and sample-dependent statistics to analyze test item psychometric properties. Most scholars study the Classical Test Theory (CTT), which involves studying the characteristics of item discrimination, item difficulty (or item facility), and distractor plausibility [19]. The ability of an item to distinguish between test-takers with low and high abilities is known as item discrimination. It contrasts test-takers with greater mastery of the to-be-learned skills and knowledge from those with less mastery. The difficulty and plausibility of a test item's distractor determine how discriminatory it is [20]. The more significant difference interprets as a contrast between upper and lower groups of test takers, in which positive and negative values can be obtained depending on the performance of upper and lower groups. The percentage of students who correctly answer an item is how an item's difficulty is numerically expressed, reflecting how challenging or straightforward an item is. Similarly, item facility (IF, or item easiness) is the proportion of test-takers who correctly answered an item in a particular test [20]; ranges from 0.00, which signifies that the item in the test was challenging, to 1.00, which shows that the item was straightforward.

A functional discriminating item must have appropriate difficulty levels, and each distractor must have the feature of plausibility. The plausibility of the distractors implied that they were indisputably incorrect while being persuasive, rational, and logical enough to appear correct to those who do not have the relevant information assessment. According to [3], one of the most challenging and problematic aspects of constructing MCQs is writing distractors rather than the key option. The importance reflects the quality of question items in terms of item characteristics, including item facility and its discriminability. Many institutions use a specific type of MCQ known as the single-best answer (SBA), in which all of the incorrect distractors have some element of truth but are noticeably weaker than the one correct response to a competent candidate [21].

The fundamental goal of assessment is to determine the extent to which students have gained the skills and knowledge that characterize the educational experience's learning objectives. Validity is another term for solid and accurate outcomes, and reliability is required for consistent outcomes [1], [10]. This study compares the effects on validity and performance observed among the multiple-choice items in Thailand with various options and how the number of options in the multiple-choice items affected those changes [22]. The primary goal of using MCQs for learning is to produce durable knowledge and skills retained over long periods and generalizable so they can be flexibly used in different contexts. Therefore, the current study should focus on identifying the optimal number of options and varying MCQ characteristics.

B. NUMBER OF OPTIONS IN MCQ

According to empirical research, the optimal number of options has become vital in considering well-known standardized test settings. The previous study hypothesized that the reliability of a test would increase with the number of options per item where the number of items remains constant [23]. Following [24], they suggested that more options improve the reliability of the test. Thus, the general guideline is to materialize as many options as possible [3]. However, there is research literature determining the optimal number of multiple-choice questions using several mathematical approaches consistent with the 3-options item. The authors demonstrate that it is better to set the number of options as 2- or 3-options in terms of being full of information and power under certain conditions (without concerning external factors). In addition, they showed results in which value has been represented in terms of improved reliability [25], [26] and information efficiency [27], [28].

Most literature experimented with giving people different types of options at random. As a result, it is difficult to create test items that prevent guessing and distortion for ability evaluation. Nevertheless, solving this research was demanding because studies examined different numbers of options, generally some combination of 5 options to 2 option items. As the difficulty of the formats increases with the number of options, some earlier studies claimed that the 2-option version of the test was the most usable yet least reliable. They concluded that three options were the optimum number [29]. However, [10] argued that reducing the number of options caused a significant change in mean item difficulty (item easier), obviously in 2 options. The study revealed that in most cases, reducing the number of options resulted in a decrease in reliability due to the most significant item discrimination index.

Additionally, with more options available for each item, it would be less likely for someone to guess correctly, reducing the likelihood of incorrect answers. However, the problem persisted in a study by [30] that compared the effectiveness of 3- and 4-option items on the English university entrance exam. They concluded that the reliability of the test was not significantly affected by the deletion of one option and that 3-option tests are not significantly different from their 4-option equivalents in terms of their difficulty or capacity for discrimination. Furthermore, the difficulty of the items in the 3- and 4-option groups was also the same. In other words, the research showed that the 3-option MCQ performs almost as well as the 4-option MCQ.

Since there are few options, it is more important to consider the likelihood of successful random guessing and the magnitude of guessing effects, such as underestimating student proficiency. The significant arguments in favor of the 5-option format and against the three-option format are implied by this condition, which lowers the test results' psychometric quality, making them less accurate and reliable [6], [7], [8]. There were significant changes in item discrimination ability when there were only three options instead of five (both increases and decreases). The most notable decreases in discrimination were observed when only two options were available. In comparison, there was a notable increase in discrimination in one instance (four to three options) [10].

Based on the existing research, there are different results for determining the number of options in MCQ. Reference [31] stated that the 5-option format is suitable if the distractors are well-constructed. Some arguments found that the 4-options format is the optimal number. At the same time, another research argued that the 3-option format is the best number in the sense of time-consuming and easy-to-construct functional distractors. For these reasons, different levels of education have different capabilities for processing information. Each type of MCQ format has its strengths and weaknesses. It relates to item difficulty, item discrimination, and the guessing factor. There are other matters to regard as the number of options rises. Increasing the number of options is an increase in challenge to the students. 2- and 3-options are more appropriate for early levels of education, such as kindergartens and primary schools.

Meanwhile, students with high cognitive will handle 4- or 5-options due to the capability to process multidimensional information. The point of the statement is between support for 5- or at least 4-options endorsed because it is said to minimize the guessing effects. According to [32], [33], and [10] concluded from the meta-analysis that using three options provides the best balance between psychometric quality and efficacy over tests with 4- or 5- option items. (e.g., students can answer more 3-options than 4-options questions simultaneously). The researchers concluded that the 3-option MCQ may or may not be optimal. Stakeholders should consider several other factors in determining the optimal number of options. However, the results of different studies are contradictory. As evident in most studies, many have been provoked by needing a firm conclusion from empirical and theoretical findings. Therefore, it is necessary to review the practical discussions in their favor to support their worth.

C. THEORETICAL CONCEPTS IN EDUCATION

1) CHALLENGE-BASED GAMIFICATION

Gamification is a technique that applies game mechanics to non-game contexts in order to increase user engagement and satisfaction. In many different fields, many definitions are generally accepted as applying game-based thinking through intrinsic and extrinsic motivation to improve user engagement and performance [34], [35], [36], [37], [38]. Gamification is rapidly being used in educational contexts to boost student motivation and learning outcomes [35], [39]. However, the types of gamification that are advantageous concerning their educational contexts still need to be further researched in their usefulness in education. Reference [40] theory of gamified learning states that "game characteristics influence changes in behavior." In order to support learning engagement and potential relevance, this premise has been applied to various contexts, particularly in the educational context, for example, in a training program [41] and an educational environment [42].

The conceptual framework of gamification mainly comprises three main game design elements: the dynamics, mechanics, and components of the game (Table 1) [34]. In many educational sectors relating to statistics and stem education, challenge-based gamification (i.e., points, levels, challenges, and leaderboard) can be effectively mixed with traditional teaching techniques such as lectures and quizzing to increase learning outcomes [43]. Thus, gamification designers should consider students' profiles, as our findings reveal that advantages vary depending on the qualities of the students.

Numerous design principles have been grouped into different categories to find a practical way to support the suitability of gamification [36], [44]. For example, gamification based on challenges might motivate people to work harder, feel more accomplished, and learn more effectively. Researchers discovered that the challenge design principle, as opposed to other principles [45], can more readily engage various players. According to a different study by [46], the game's challenge influenced learning and engagement favorably. The challenge component is a moderately powerful motivator that yields various motivational effects depending on the usage context. Participants who attempted to answer challenging questions during the process enhanced their learning abilities under pressure and received feedback. In contrast, people reported feeling disengaged when the level of challenge surpassed a predetermined threshold. The authors attempt to summarise the ideal learning conditioning when game mechanics and elements were well-designed in this context.

TABLE 1. Game design elements.

Dynamics	Mechanics	Components
Constraints	Challenges	Achievements
Emotions	Chances	Avatars
Narratives	Competition	Badges
Progression	Cooperation	Collections
Relationship	Feedback	Unlockable Contents
	Resources	Leaderboards
	Rewards	Dashboard
	Turns	Levels/Tiers
	Win-Lose Status	Points/Scores
	Exchange	Virtual Goods

2) FLOW THEORY FOR LEARNING AND DEVELOPMENT: ZONE OF PROXIMAL DEVELOPMENT

Currently, most game designs provide the ability to adjust the difficulty level according to the player, and individuals may tackle increasingly tricky challenges with higher levels of skills [45], [47], [48]. Once reaching the flow, it often implies that the player intrinsically experiences a sense of enjoyment and satisfaction [49]. However, the existing studies needed comprehensive guidance for educational purposes. The claim that traditional testing needs to be extended by assisted performance has significantly impacted the assessment practice, including dynamic assessment and adaptive systems.

Reference [50] had provided a theoretical framework on the *Zone of Proximal Development* (ZPD) to assist in understanding and to learn [51]. The primary premise presented in this article is that it is too complex for learners to execute on their own but straightforward enough for them to do with guidance. With the appropriate guidance, learning will be more effective in this Zone of Proximal Development. Scaffolding is another critical component of the ZPD hypothesis to achieve the key objectives. Reference [52] adopted the word scaffolding to operationalize the concept of teaching in the ZPD [53]. Scaffolding is used in ZPD to illustrate the social and interactive aspect of instruction and learning that occurs in the ZPD. Several noteworthy articles have advanced theoretical understanding of the ZPD concerning instruction [51], [53], [54], [55], [56], [57].

As stated by the researchers, ZPD has significant implications for pedagogy and learning. Based on the approach, the goal is to set tasks that most students could solve with some aid. One approach to support this development is scaffolding, which involves structuring the concepts to be comprehended that encourage individuals to develop beyond and better themselves [12]. The main goal of education from a Vygotskian perspective is to keep learners in their ZPDs as much as possible [58]. This situation can be achieved by giving them exciting and culturally meaningful learning and problem-solving tasks that are slightly more difficult. They must work together to finish the task with another competent peer, teacher, or adult. ZPD is defined as the gap between the developmental level determined by individual critical reasoning and the possible developmental level determined by problem-solving under sufficient guidance or in collaboration with more capable peers [50]. ZPD concept turns the link between development and learning on its side, and the critical element of learning is that it generates the zone of proximal development, which is a necessary component of development and can result from learning.

3) PROSPECT THEORY FOR MAKING DECISION UNDER UNCERTAINTY FOR MCQ

Prospect theory (PT) is a theory of behavioral economics and behavioral finance developed by [59]. Based on psychophysical concepts of loss aversion where individuals felt losses asymmetrically more than gains. Its goal is to establish when the subject psychologically detects a change in the physical stimulation. This condition emphasizes the notion that humans are sensitive to wins and losses. However, most people avoid making such a decision when given an equal possibility of gaining or losing an equal amount. This situation implies that people are more sensitive to losses than gains; the fear of loss outweighs gaining satisfaction.

Prospect theory has some similarities to the concept of the expected utility function. Similarly, the value function is evolving based on Bernoulli's, von Neumann's, and Morgenstern's concepts derived from the expected value [60]. The normative decision model assumes that expected utility exists independently of probability. The utility is defined in a chance game as the point of indifference between an inevitable and uncertain outcome. In summary, utility is portrayed as a subjective way of determining the significance of any given objective outcome to a specific individual.

Prospect theory, under uncertain conditions, provides individuals with the likelihood of various outcomes and refers to cases in which the uncertainty is due to sources other than the decision maker [61]. For instance, in gambling, such as lotteries. However, the decision maker's uncertainty is rooted in internal sources in many circumstances. For example, the individual who must answer MCQs from a pool of all possible options means that people will encounter uncertainty as a result of insufficient or untrustworthy knowledge.

This study focuses on decision-making under conditions of internal uncertainty. It specifically investigates the effect of framing on MCQ activity. The sequence or manner a choice or option is presented to a decision-maker is referred to as *framing effects*. This situation is necessary to select the best option from all potential options. When the outcome of the two alternatives is the same, the framing effect occurs. When the options are framed differently, we prefer the one that is more positively framed. The goal is to simplify a choice maker's available options. As a result, determining and creating the available options is at the center of decision-making.

D. GAME REFINEMENT THEORY AND MOTION IN MIND

1) GAMIFIED EXPERIENCE AND GAME PROGRESS MODEL

A general model of game refinement, based on a logistic model of game uncertainty, was proposed by [62]. From the player's viewpoint, the information regarding the game result is an increasing function of time t (i.e., the number of moves in board games). Here, the information on the game result is defined as the amount of solved uncertainty (or information obtained) $x(t)$, as given by (1). The parameter n (where $1 \leq n \in \mathbb{N}$) is the number of possible options, $x(0) = 0$, and $x(T) = 1$.

$$x'(t) = \frac{n}{t}\,x(t) \tag{1}$$

where $x(T)$ is the normalized amount of solved uncertainty. Note that $0 \leq t \leq T$, $0 \leq x(t) \leq 1$. Equation (1) implies that the rate of increase of the solved information $x'(t)$ is proportional to $x(t)$ and inversely proportional to t. By solving (1), (2) is obtained. The solved information $x(t)$ is assumed to be twice derivable at $t \in [0, T]$. The second derivative of (2) indicates the accelerated velocity of the solved uncertainty along with the game progress described by (3). The acceleration of velocity implies the difference in the rate of acquired information during game progression. Then, a measure of game refinement GR is obtained as the root square of the second derivative, described by (4):

$$x(t) = \left(\frac{t}{T}\right)^n \tag{2}$$

$$x''(t)* = \frac{n\,(n-1)}{T^n}\,t^{n-2}\,|_{t=T} = \frac{n\,(n-1)}{T^2} \tag{3}$$

$$GR = \frac{\sqrt{n\,(n-1)}}{T} \tag{4}$$

Let p be the probability of selecting the best choice among n plausible options. As such, the definition of gamified experience is based on the notion of the risk frequency ratio or risk-taking probability, which is defined as $m = 1-p = \frac{n-1}{n}$ [63]. Then, the gamified experience is achieved if and only if the risk occurs with parameter $m \geq 0.5$, which implies $n \geq 2$. Knowing that the parameter n in (4) is the number of plausible moves, for a game with branching factor B and length D, $n \simeq \sqrt{B}$ is approximated where the GR is given as (5): The GR measure has been adopted and verified in various games, as demonstrated by previous studies [64], [65].

$$GR_{board} \approx \frac{\sqrt{B}}{D} \tag{5}$$

TABLE 2. Various GR measures of several board games and sports.

Game	$B \vee G$	$D \vee T$	GR	a
Western Chess	35	80	0.074	0.00547
Chinese Chess	38	95	0.065	0.00423
Japanese Chess	80	115	0.078	0.00608
Mah Jong	10.36	49.36	0.078	0.00608
Go	250	208	0.076	0.00578
Table Tennis	54.86	96.47	0.077	0.00593
Basketball	36.38	82.01	0.073	0.00533
Soccer	2.64	22	0.073	0.00533
DotA	68.6	106.20	0.078	0.00608

TABLE 3. An analogical link that relates physics in mind notations and its in-game counterparts [63].

Notation	Physics context	Game context
y	Displacement	Solved uncertainty
t	Time	Progress or length
v	Velocity	Solving rate
M	Mass	Solving hardness, m
g	Acceleration (gravity)	Acceleration, a (thrills)
F	Newtonian force	Force in mind (move ability)
$\vec{p}$	Momentum	Momentum (move intensity)
U	Potential energy	Potential energy, E_p (move potential)

For the scoring games, the GR measure is determined by 4. G stand for the average goals. Meanwhile, T is the total points or goals. GR is given as (6).

$$GR_{scoring} \approx \frac{\sqrt{G}}{T} \tag{6}$$

In this study, the solving rate is given as v, and the solved uncertainty of the game $y(t)$ is an increasing function of time t, which can be described by (7). A player may feel informational acceleration, which is formulated analogically to the physics formulation of motion $y(t) = ut + \frac{1}{2}at^2$. Since u is the initial velocity at $t = 0$, then (8) is obtained. The intersection can be calculated at $t = D$ or $t = T$, where (9) is identified as the informational acceleration that describes the gamified experience and comfortable thrills under consideration.

$$y(t) = vt \tag{7}$$

$$y(t) = \frac{1}{2}at^2 \tag{8}$$

$$a = \frac{2v}{D} = \frac{2v}{T} \tag{9}$$

Sophisticated games possess an appropriate game length to solve uncertainty while gaining the necessary information to identify the winner [63]. This condition can be found at the cross-point area between (7) and (8), where $a = \frac{B}{D^2}$ is identified as the noble uncertainty zone ($\in [0.07, 0.08]$) of the $GR = \sqrt{a}$ (or $a = GR^2$), as previously found (Table 2).

2) MOTION IN MIND PERSPECTIVES

In Newton's mechanics, an object's mass defines it. In a game, a stronger (weaker) player has a higher (lower) skill level for handling uncertainty. This scenario suggests that during a game, the stronger (weaker) player would encounter lower (higher) uncertainty. Therefore, the stronger (weaker) player is given a smaller (larger) value of mass (let's say m). This idea prompts the interpretation of mass (m) in game play, which relates to the degree of a player's challenge while playing. In the present context, v typically refers to the speed at which uncertainty is resolved. In contrast, m (where $m = 1 - v$) denotes the difficulty of solving such uncertainty. As such, both v and m are determined by the following [63].

Let B and D be the average number of possible moves and game length, respectively. Solving rate v is approximated as (10), by which v is equivalent with the slope v of the game progress model in (7). Let G and T be the total number of goals and attempts per game, respectively. Scoring rate v is given by (11), where the slope v of the game progress model is in (11). Then, interpretation in the education context can be inherently conducted. As such, v and m are defined as the correctness rate or the rate of solving the question correctly (capability) and winning hardness or difficulty (challenge), respectively. Intuitively, velocity v is defined as v_0, namely "objective velocity," which is the individual perceived game or activity experienced by an object.

$$v = B/2D \tag{10}$$

$$v = G/T \tag{11}$$

The fundamental element was measuring the mass and velocity, enabling the derivation of force (F), momentum ($\vec{p}$), and potential energy (E_p). Intuitively, Table 3 illustrates an analogical link that relates physics in mind notations and its in-game counterparts [63]. Based on such analogy, various motions in games can be determined, where the momentum, potential energy, and force were defined as (12), (13), and (14), respectively.

$$\vec{p} = m \cdot v \tag{12}$$

$$E_p = ma \cdot y(t) = ma\left(\frac{1}{2}at^2\right) = \frac{1}{2}ma^2t^2 = 2mv^2 \tag{13}$$

$$F = ma = (1 - v) \cdot a \text{ and } a = \frac{B}{D^2} \tag{14}$$

Previous works showed that $\vec{p}$ represents the player's growth rate determined by the different game depth and change over time [66]. The ability of the users to play the game is indicated in this paper by $\vec{p}$. In addition, the amount of movement potential (curiosity) the game may transfer to the player is reflected in the game's energy [63]. It was based on the concept of attractiveness toward players, like gravity based on move potential. For the challenge in a skill-based game to contribute to the anticipated risk (a higher m), there must be a certain amount of anticipation. The player sees this as motivation to move because it suggests a higher chance of progressing in the game (high E_p), and anticipation decreases if the player gathers enough information (i.e., results in a desirable outcome).

By converting the quizzing activities into two different types of quantities—game refinement (GR) and velocity

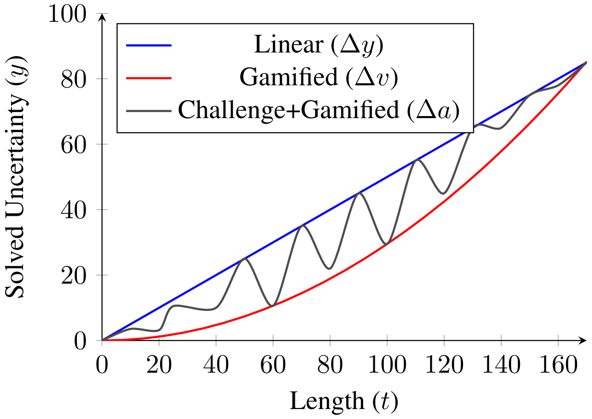


FIGURE 1. An illustration of extended game progress model based on solved uncertainty (y) over time t.

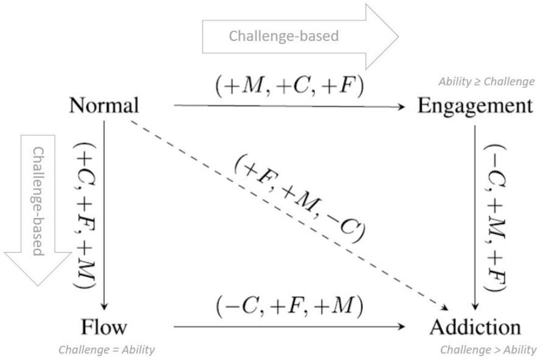


FIGURE 2. The conceptual model of engagement and addiction by [68] adopting challenge-based gamification.

(v)—this paper examines the challenge-based gamification in relation to the various analogies of motion and its conservation. The quizzing activities are viewed from the perspective of information science as a linear amount of solved uncertainty. The activity or the accelerated amount of solved uncertainty (i.e., Δv) can be thought of as being gamified as a result of adopting game refinement theory. Challenge-based gamification mimics the Δv and Δa signals that cause vibration or jerk (j) as an activity progresses. Jerk is a metaphor for abrupt changes in the experience of thrills and engagement in relation to the retention of motivation [63], [64]. This situation is illustrated in Figure 1.

Motivational engagement was defined by [67] as the conscious awareness of motivation for a specific action. A player's perceptions of their ability to influence the game's flow and their chances of success depend on how strong this awareness is. The study found that there are significant contextual engagement gaps between the gaming and educational domains related to pre-gaming decisions (such as game genres and enthusiasm for the activity) and "opposing interference force towards positive objectives" (such as mastering the game and overcoming obstacles).

According to motion in mind model by [68], motivation is related to the potential energy (E_p), which takes into account the importance of the play's development and the player's expectations. The velocity (v) and momentum ($\vec{p}$), which stand for the rate of an individual's development and activity, respectively, were linked to control and focus. The progression and performance of specific students were also investigated using challenge-based gamification. This circumstance may be regarded as a prerequisite for properly analysing transition gamification in relation to the motion-in-mind concept. Therefore, based on previous studies, a conceptual model linking motion in mind to such psychological attributes was constructed, as depicted in Figure 2.

3) VARIABLE RATIO (VR) SCHEDULE AND WINNING HARDNESS (m)

The highest response rate was discovered to be variable-ratio (VR) [69], indicating that regular and straightforward rewards for performing one thing are not the best strategy to elicit the intended behavior. The effectiveness of such a schedule can be increased by randomly changing the reward after various tasks. In VR schedules, the parameter N shows the average reward frequency, where $1 < N \in R$.

In this study, winning a game corresponds to obtaining a reward. It implies the game length, D in board games (total number of ply), and scoring games (total points or goals). Hence, $N = D$ or $N = T$, implying a general form of reward frequency of the game's winning rate. Based on this notion, the winning rate v_0 and winning hardness m is defined by 15. This study adopts an MCQ paradigm based on the VR schedule to establish the link between learning and playing. This approach utilizes motion-in-mind measures to propose the underlying relationship between the VR schedule and the number of options (N) in the MCQ.

$$v_0 = 1/N \tag{15}$$

E. OBJECTIVITY AND SUBJECTIVITY USING LAW OF CONSERVATION

It has recently focused mainly on a possible link between the motion in mind in several contexts [43], [70], [71], the goal is to find a link among all that the motion in mind may cover all theories as a grand unified theory. People's behaviors would tend to maximize the comfort in mind, establishing a behavioral pattern as culture [68]. It indicates the area of comfort in mind in various contexts based on the interpretation of solving hardness m to the player's ability k.

The objective velocity v_0 and mass m are the winning rate for the objective and hardness rate, respectively. Velocity v_0 holds on the function of mass m, given by the equation $v_0(m) = 1 - m$, which k equals 0 (i.e., perfect player). Based on the interpretation of solving hardness m to the player's ability k, the subjective velocity of the player (or individual) with ability level k is denoted mathematically as (16).

$$v_k(m) = (1 - km) \cdot v_0 \tag{16}$$

The subjective velocity v_k is varied by ability level k. The larger k means the player has less ability, while the smaller k

means the player has a higher ability. The subjective velocity will become zero when the player's ability holds at $k = \frac{1}{m}$ relative to mass m. As previously defined, potential energy E_p (or E_0) will be maximized at $m = 1/3$, so when considering typical board games like chess and Go, we reasonably assume $k = 3$. However, when considering sports games, $k = 4$ might be more reasonable. The illustration of motion in mind measurement has shown in Figure 3(a) and Figure 3(b). Based on the interpretation of m to the player's ability k, the subjective energy E_k of the player (or individual) with ability level k is denoted mathematically as (17), representing the people's engagement from the subjective point of view.

$$E_k(m) \ = 2mv_k^2(m) \tag{17}$$

The comfort in mind could be evaluated by the total energy in mind, namely E. Several comforts in mind in various contexts have been investigated while considering the mass m and new measurement Δ_k, which is given by the absolute difference between objective and subjective reinforcement. Masters and beginners should widely play the popular game at that time/era in its history. It postulates that the objective reinforcement meets the subjective reinforcement ($E_0 = E_k$) to be mass entertainment at $m = 2/k$, where m stands for the mass of the popular game under consideration. When assuming $k = 3$, the border is $m = 1/3$ and $m = 2/3$, respectively, which looks well-balanced (play comfort) in popular games such as Go, Chess, and Shogi. The reinforcement difference (Δ_k) can be depicted as a curiosity. The motion-in-mind concept depicts three areas based on the mass criterion m. The conjecture can be defined as follows.

Conjecture 1 (Interpretation of m With Respect to the Reinforcement Difference): The first is the known (or solved) are $0 \leq m \leq 1/k$, where implementing fairness/equality is essential. The subjective velocity v_k would become zero when $m = 1/k$, where, Δ_k ($= E_0 - E_k$) will be maximized. The second is the learning area $1/k \leq m \leq 2/k$, where people are roused to learn. The third is the play (or emotion/entertainment) area $2/k \leq m \leq 1$, where people enjoy playing activities with specific emotions.

One essential characteristic is that learning comfort holds the objective reinforcement E_0 dominating over the subjective one, so knowing the game-theoretical value or solving uncertainty is optimized. In the learning context, individuals would feel highly engaged at the peak of Δ_k ($= E_0 - E_k$). It thus implies incorporating challenge-based gamification. Then, motion in mind measures such as v_k will generalize δ_k, representing the competitiveness. People would feel comfortable at $m = 1/k$ due to full social equality (social comfort) [70]. Smaller δ_k (high competitiveness) corresponds to less motion of game score or stable/deterministic. δ_k is maximized at $m = 1/2$ in every k where people would feel comfort due to its fairness, namely $\delta_k = 0$. Larger δ_k (low competitiveness) corresponds to the greater motion of the game score or unstable/stochastic. It implies that a game depends on $\delta_k = kmv_0$ (momentum of play) or the balance between skill and chance.

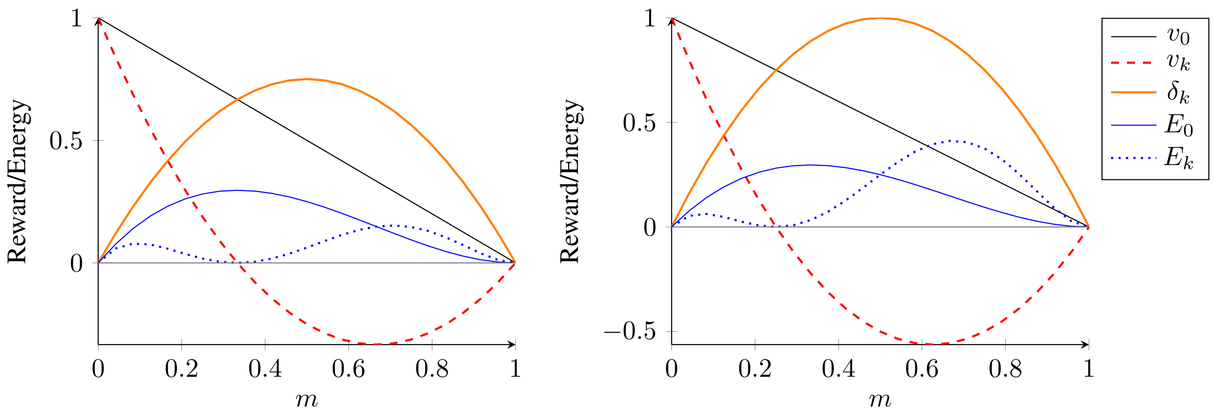
By the way, a certain degree of competition may be incorporated and varied in each society. Then, inequality or disparity issue happens as well. The critical point is to use the difference between objective and subjective ones as measurements. The objective aspect is essential in the ordinary context, like the physics of nature. At the same time, the subjective one is essential in the mind's perspective, but it cannot be measured in solitary. This condition implies why the difference would be employed as a reliable measurement. When focusing on competition, it is reasonable to assume that society is less competitive than competitive games and that educational games or gamification would be intermediate between society and competitive games. Go, and sports games would be typical examples, where mass m holds at $\frac{1}{3}$ to $\frac{1}{2}$ (approximately $k = 3$ and 4).

III. METHODOLOGICAL ASSESSMENT AND RESULTS

The participants of this study were 48 senior high school students within the age range of 15–18. They were voluntarily recruited with a declaration to find the participants whose qualifications matched the study, and written consent was obtained. The participants attended private schools in Thailand. They received two to five hours of English education every week as part of their compulsory education. Although under the current pandemic and the school's regulations on the ethical conduct of research, approval for the participants was sought from and granted by the Associate Dean and coordinator staff. In this study, we followed the directions that all participants must be informed and highlighted all the negative and positive effects during the consent procedure, including proclaiming the objectives and Nature of the research also, the workflow of the research, and expected results were explained to the participants [72], [73]. Although voluntary participation was voluntary, the overall sample was selected based on the participants' pre-performance, which achieved the standard score to apprehend the English structure and validate the results.

Two primary indices are used to systematically evaluate the effectiveness of individual items in a norm-referenced test: item difficulty and item discrimination. This study employs the classical test theory (CTT) model rather than other test theories (e.g., Item Response Theory) because this theory is viewed to be more flexible and trustworthy. It was employed in various situations [18], [74]. Participants took 300 MCQ items, including 3-, 4- and 5-options. MCQ in the item pool that was used in the study was gathered from TOEIC test examinations from Cambridge's, Oxford's, Barron's, and Longman's textbooks. In each test, the item stems were identical. Participants were encouraged to respond to all items, even unsure ones. Although they did not know the feedback of their responses and could not indicate correct responses at the time, only total points were given, to avoid a recall and rote memory that may burden the quality of results.

According to literature from [32], [75], and [76], standard MCQs have presented 4 or 5 options for high-stakes



(a) Ability level $k = 3$

(b) Ability level $k = 4$

FIGURE 3. Measures of motion in mind for $k = 3$ and $k = 4$.

assessment. At the same time, 3-option MCQs are advised for classroom-based assessments and achievement tests requiring more knowledge to be taught in a short period. Our proposed experiment was thus that of a standardized external examination. A time limit of 30 minutes was provided as a baseline in the variation of MCQ sections. An example item is illustrated as Example 1.

Example 1: This is the example of the cloze test question used in this study.

Stem: Being a wise politician, Mr. Brown tends to reserve his .... till he knows all the facts.

Key & Distractors

(A) benefits
(B) bookings
(C) appearances
(D) judgements

A. ITEM ANALYSIS

Item analysis is essential in improving items and eliminating ambiguous or misleading items in a single test administration. CTT provides a simple and intuitive approach to item analysis [77]. It utilizes nothing more complicated than proportions, averages, counts, and correlations. For this reason, it is helpful for small-scale exams or uses with groups that do not have psychometric expertise [78]. As mentioned, CTT quantifies item difficulty for dichotomous items as the proportion (P value) of test takers correctly answering it. Another psychometrics is item discrimination, which we use to discriminate among test takers. Suppose the item is substantial and measures the topic well. As part of our experiment, a purposeful mix of 3-options to 5-options MCQ was selected from the item pools. The classical test theory arbitrarily categorized criteria for determining easy, moderate, and difficult items. The theory quantifies item difficulty and discriminability by the item's scores to the test's total score.

1) ITEM DIFFICULTY AND DISCRIMINATION

Question items in the pool combined 77 items of 3 options, 170 items of 4 options, and 53 items of 5. Participants are asked to complete all question items. The item difficulty is merely the percentage of participants who answer an item correctly. In this case, it is also equal to the item mean, which can be calculated by dividing the number of participants by the total number. The item difficulty index ranges from 0.00 to 1.00 (or 0 to 100%); the higher (lower) the value, the easier (harder) the question. In this study, we define that all the question items are marked dichotomously. Participants will achieve one point for the best single correct answer and no point for choosing distractors (incorrect answers).

Table 4 visualizes the scores achieved by the 3-option group were higher than those of the 4-option group, which were more significant than those of the 5-option group. We arbitrarily classify item difficulty as easy if the index is 75% or above, moderate between 51 and 74%, and hard if it is 50% or below. Moreover, it displays the difficulty index values based on the proportion of the number of correct answers in each session against the total number of participants. From the analysis, the test consisted of 33 easy items, 160 moderate items, and 107 difficult items. In deeper detail, all 33 easy items are accumulated from the 3-options group; 160 moderate items are accumulated from 44 items from the 3-options group, 106 items from the 4-options group, and ten items from the 5-options group. Also, 107 moderate items are accumulated from 43 items from the 4-options group and 64 from the 5-options. This condition implies evidence that the difficulty of test items relies on the number of options, a tendency for the 3-option items to be easier than their 5-option.

The procedures have traditionally been used to compare item responses to total test scores using high and low-scoring groups of students. To be clarified, participants were divided into higher and lower ability groups of 27% each for various options formats, in which upper and lower groups are counted as 12 and 12, respectively. A discrimination index (DI) was established for each item based on the number of correct answers in the upper group less the number of correct answers in the lower group, with the product divided by the participants in each of the higher and lower groups. This study

TABLE 4. Difficulty indices and discrimination indices on our proposed question items.

	3-options	4-options	5-options	Total
Easy	33	0	0	33
Moderate	44	106	10	160
Hard	0	43	64	107
			Total	300
Strong discrimination	70	150	34	254
Fair discrimination	7	10	10	27
Poor discrimination	0	10	9	19
			Total	300

TABLE 5. Concurrent validity based on correlation between TOEIC and scores in our proposed MCQ, and *KR* − 20 reliability coefficients of the experimental tests.

	3-options	4-options	5-options
# participants (n)	48	48	48
# questions	30	30	30
TOEIC scores (%)		53.33	
# questions	77	170	53
mean scores	56.25	89.65	17.92
mean scores (%)	73.05	52.73	33.80
Correlation coefficient (r)	0.964	0.986	0.981
significance ($p <$)	0.0001	0.0001	0.0001
$KR-20$	0.95	0.97	0.79

classifies item discrimination as good if the index is above 0.30, fair between 0.10 and 0.30, and poor if it is below 0.10. Items with negative indices should be examined to determine why a negative value was obtained. For instance, the item was a mistake, or participants may need clarification, but they respond correctly. Table 4 also shows that 254 items hold strong discriminability, 27 hold fair discriminability, and 19 hold poor discriminability. This condition implies the evidence that the discriminability of test items relies on the number of options, a tendency to guess or misunderstand the item's key increases, as well as the number of options and difficulty.

2) VALIDITY AND RELIABILITY

The MCQs validity consideration of this study was primarily aimed at showing that the construct being measured by the experimental tests was broadly similar to that measured by the Reading Part 5 of the TOEIC test. It consisted of 30 items and was therefore confined to comparing the correlations between the overall scores of the 48 participants, the three types of multiple choice format in the TOEIC test, and their experimental multiple choice test scores. Initially, the TOEIC test consists of 4 options, and test takers are asked to choose the single best answer among all options. The test result will be used as criterion reference data against which comparison of the experimental tests.

Based on the number of question items in the experiment, the score of each session is generalized into a percentage to calculate the correlation between the two tests. First, Pearson's correlation is introduced to determine the correlation coefficient between practice and experimental tests. Then, the correlation coefficient (r score) and significance p-value were calculated. From Table 5, the correlation coefficients between the 4-options MCQs TOEIC in the reading part and their respective scores in the three experimental tests were all significant at the 0.00001 level. The 3-option format had the lowest value, while the 4-option group had the highest because the criterion reference test was in the 4-option format.

Moreover, Cronbach's alpha is an essential tool for measuring the strength of internal consistency. This tool refers to how closely a set of items are as a collective, typically associated with measuring scale reliability. In other words, it is defined as the purpose of the number of items in a test, the average covariance between pairs, and the total score variance. They are widely used to quantify the reliability of reporting research scales and survey questionnaires. An analysis known as the Kuder-Richardson 20 formula ($KR-20$) can be conducted to determine the reliability of the scale, computed in a binary form (between 0 and 1) [79].

This index checks the internal consistency of measurements with dichotomous choices. Cronbach's alpha is usually reported in scales ranging from 0-1, with the larger values representing more reliability. Inherently, $\alpha \geq 0.7$ is usually considered acceptable, while too high a value ($\alpha > 0.9$) indicates a homogeneous test. The coefficients $KR-20$ for the three formats were tabulated in Table 5. The coefficient for the 4-options format is the highest, while the 5-options format is the lowest. However, all coefficients are situated above 0.7, which is acceptable. The coefficients for the 3-option and 4-option formats were very close, lying between 0.87 and 0.89. The coefficient for the 5-option format was lower, mainly due to the small number of items and more considerable variance among the scores.

B. VARIATION OF NUMBER OF OPTIONS IN MCQs

On this basis, the chance rate in MCQ is aligned with the terminology of guessing, which calls the guessing rate equal to $1/N$, where N is the average ratio of the number of options. This condition means that the average of N plausible options in the test would be the correct answer. For example, the probability of choosing the desired response by random guessing decreases from 0.5 for 2-options items (i.e., true or false) to 0.33 for 3-options MCQ to 0.25, 0.20, and 0.17 for 4-,5-, and 6-options items, respectively. The chance to answer correctly depends on the number of options. The deeper detail claims that only two possible correct or not scenarios exist. This condition implies that the success rate is defined as the number of correct answers over the total question items.

The analysis of the number of options in MCQ focuses on the optimal number of options required for each context. This study commenced with two game progress models introduced

TABLE 6. Motion in mind measures of three experimental multiple-choice question formats based on the partial knowledge.

Options	#Qs	Mean Scores	v_0	m	N	E_0	$\vec{p}$
3	77	56.25	0.73	0.27	1.4	0.2876	0.1968
4	170	89.65	0.53	0.47	1.9	0.2629	0.2492
5	53	17.92	0.34	0.66	3.0	0.1513	0.2237

#Qs: number of questions;

by [63] and [65], where the number of options in the test can be reduced by gaining knowledge. The intuition is based on the transition of the chance-based scenario to a skill-based scenario. This condition corresponds to the move selection model in board games. However, examination, test, or quiz are the assessment method that evaluates performance based on the score. This situation corresponds to the score progress model, usually used in sports games.

Assuming that participants have partial knowledge of English subjects, the data collection was conducted through raw data of $n = 48$ participants. The measure based on the motion-in-mind model was conducted to compute all data. First, it was found that the least successful was the 5-options format ($v_0 = 0.33$). Then, the most successful one was the 3-options format ($v_0 = 0.73$). Finally, the 4-options format is considered the midpoint between the two with the success rate of $v_0 = 0.53$. Table 6 illustrates the experimental data on the motion in mind values of v_0, m, $\vec{p}$, E_0, and N.

For the 3-options format, the objective energy, momentum, and average options were $E_0 = 0.2876$, $\vec{p} = 0.1969$, and $N = 1.4$, respectively. This result indicates that participants will encounter the MCQ event with similar options. This condition means that participants could achieve more than 70% of correct responses. For the 4-options format, the $E_0 = 0.2629$, $\vec{p} = 0.2492$, and $N = 1.9$. This situation means that the event is close to having a chance to select one out of two plausible options, where ability and chance are balanced. For 5-options format, the $E_0 = 0.1513$, $\vec{p} = 0.2238$, and $N = 3$. This condition implies that three out of five options must be successful. It indicates that the event is equivalent to guessing among three options, where participants can achieve more than 33% of correct responses. Figure 4 illustrates the motion-in-mind measure of mass (m) when $k = 3$.

The experiment was validated by running it using AI simulation written in Algorithm 1 to perform the test from our proposed question items. The motivation for experimenting with several approaches to the multiple-choice tests is because its characteristics develop and originate from guessing terminology. It interprets the success rate as the probability for an individual answer the test correctly. The outcome relies on the individual ability and branching factor, which is the number of options. The results of the velocity (v_0) for each multiple-choice question format using our proposed agent model are given in Table 7. The agent's algorithm followed the guessing terminology, in which the agent will be assumed to have partial knowledge of the questions. Then, the agent will counter the questions by trying to guess based on the probability of each option. To simplify the calculation and given that we deal with hypothetical questions, we bound the ability level of the agent to not knowing at all. For this case, the probability of being correct equals to $\frac{1}{N}$, where $N =$ the number of options. The probability of being correct with partial knowledge is greater than $\frac{1}{N} \rightarrow 1$. The results were collected from analysis and 2400 simulation rounds, where the v_0 was computed based on different ability levels of the agent.

Algorithm 1 MCQ AI With Various Level of Knowledge

```
Data: Assign key option with 1 points, question items
$v_0$:= success rate;
initialization;
sample ← 2400;
$N$ ← 300;
totalPoints ← 0;
for $x \in Sample$ do
    $Sum[x] \longleftarrow 0$;
    for $y \in N$ do
        $p1$ ← randomProbability() ∈ [0.33, 1];      /* 3-options */
        $p2$ ← randomProbability() ∈ [0.25, 1];      /* 4-options */
        $p3$ ← randomProbability() ∈ [0.2, 1] ;      /* 5-options */
        keyOption() ← {$p1, p2, p3$ };
        $A[y]$ ← selectedOption();
        if $1 \in A[y]$ then
            $Sum[x] \leftarrow Sum[x] + 1$;
        $Sum[x] \leftarrow Sum[x] + 0$;
    $totalPoints \leftarrow totalPoints + Sum[x]$;
$avg \leftarrow \frac{totalPoints}{sample}$;
$v_0 \leftarrow \frac{avg}{N}$;
return $v_0$;
```

Table 7 delivers the experimental results of 3-, 4-, and 5-options formats incorporated in MCQ. The procedure of the experiment was similar to the investigation in pre-measurement. It was conducted in a test set used before, which identical question pools and the number of questions. In the case of partial knowledge, a random participant might take the outcome beyond the pure guessing strategy. This situation corresponds to how the options might be reduced to a small number while applying the knowledge. Based on the guessing terminology, the chances of successful random guessing with a 5-options format, the degree of likelihood of success is 20%; with 4- options, it is 25%; and with 3-options, it is 33.3%. By random participants, the chance of being correct in the 3-, 4-, and 5-options format increases from 0.33 to 0.61, 0.25 to 0.49, and 0.2 to 0.41, respectively. With the reduction of the number of options N, the success rate (v_0) also improved. Such results determine the optimal number of

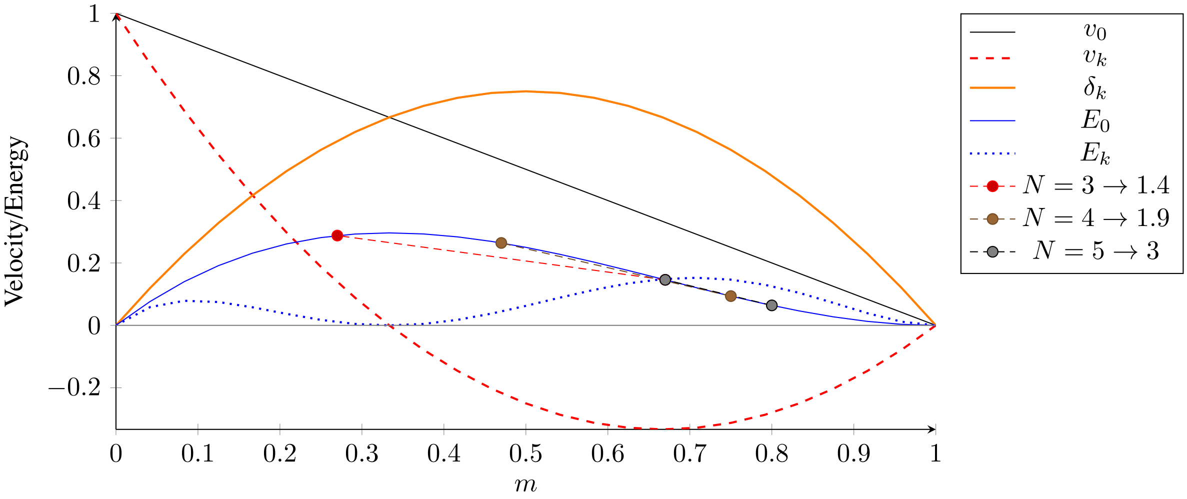


FIGURE 4. Measures of motion in mind with indicator of N for $k = 3$.

TABLE 7. Motion in mind measures three multiple-choice question formats between human data and simulation data.

Types of Experiments		v_0	m	N	E_0	$\vec{p}$
3-options	Human	0.73	0.27	1.4	0.2878	0.1971
	Simulation	0.61	0.39	1.6	0.2902	0.2379
4-options	Human	0.53	0.47	1.9	0.2640	0.2491
	Simulation	0.49	0.51	2	0.2449	0.2499
5-options	Human	0.33	0.67	3	0.1459	0.2211
	Simulation	0.41	0.59	2.4	0.1984	0.2419

TABLE 8. Motion in mind measures of three multiple-choice question formats with time pressure from t = 30 to t = 15 minutes.

Options	Time*	v_0	m	N	E_0	$\vec{p}$	GR
3	30	0.65	0.35	1.5	0.296	0.2264	0.147
	15	0.37	0.63	2.7	0.17	0.232	0.11
4	30	0.55	0.45	1.8	0.2736	0.2471	0.136
	15	0.26	0.74	3.9	0.0994	0.1919	0.093
5	30	0.36	0.64	2.8	0.1634	0.2293	0.109
	15	0.22	0.78	4.6	0.0735	0.1697	0.085

*: in minutes;

options for participants to receive the desired outcomes. The human and simulation data results have framed a margin of the average number of options from $N = \{3, 4, 5\}$ (without knowledge) to $N = [1.5, 3]$ (partial knowledge) for 3-,4-, and 5-options, respectively.

C. VARIATION OF SETTINGS IN MCQs

1) CHALLENGE-BASED GAMIFICATION VIA TIME PRESSURE

This experiment allowed the participants to be reinforced by challenges and changed individual actions toward risk. Time pressure can change individual behavior and positively and negatively impact students' performance, which discerns potential learning and engagement. It is hypothesized that the impact of time pressure would shrink the difference between students in the upper and lower group. In Table 8, the quantitative measure by using the motion-in-mind model was provided comparatively for two groups; 30 minutes group (pre-measure) and 15 minutes group (post-measure)—the experiment was conducted for 3-options, a 4-options, and 5-options formats. The number of question items was fixed at 30 items. The flow of experiments begins with the pre-measure experiment (30 minutes group), followed by the post-measure experiment (15 minutes group). The difficulty level of question items for both experiments is identical by positioning at a moderate level, but the question stems are different. When computing the measurement, it can be observed that both experimental groups have disparity. Participants perform better in ample time, while time pressure reduces overall performance. However, the game refinement value of the time pressure group nearly converges to the sophisticated zone of ($GR \in [0.07, 0.08]$), which implies the magnitude of the thrilling sense.

For the 3-options format, the results depict the objective velocity (v_0) of 0.654 and 0.366 for $t = 30$ and $t = 15$ minutes, respectively. The objective energy (E_0), momentum ($\vec{p}$), and the average number of options (N) were 0.296 and 0.169, 0.226 and 0.232, and 1.5 and 2.7, at $t = 30$ and $t = 15$ minutes, respectively. Meanwhile, the v_0 reduces to 0.5535 and 0.259 at $t = 30$ and $t = 15$ minutes, respectively, for the 4-options format. Its E_0, $\vec{p}$, and N also change to 0.2736 and 0.0994, 0.2471 and 0.1919, 1.8 and 3.8, respectively. For the 5-options format, the results showed that v_0 at 0.356 and 0.644 when $t = 30$ and $t = 15$ minutes, respectively. At the same time, E_0 was 0.163 and 0.073, $\vec{p}$ was 0.229 and 0.170, and N was 2.8 and 4.6, respectively. Curiously, the changes between the option format showed that when $t = 30$, v_0 tends to reduce with increasing options while reciprocating when $t = 15$—however, other quantities responded as expected.

Additionally, the GR values were calculated to investigate the attractiveness of the gamified test. Figure 5 depicts score

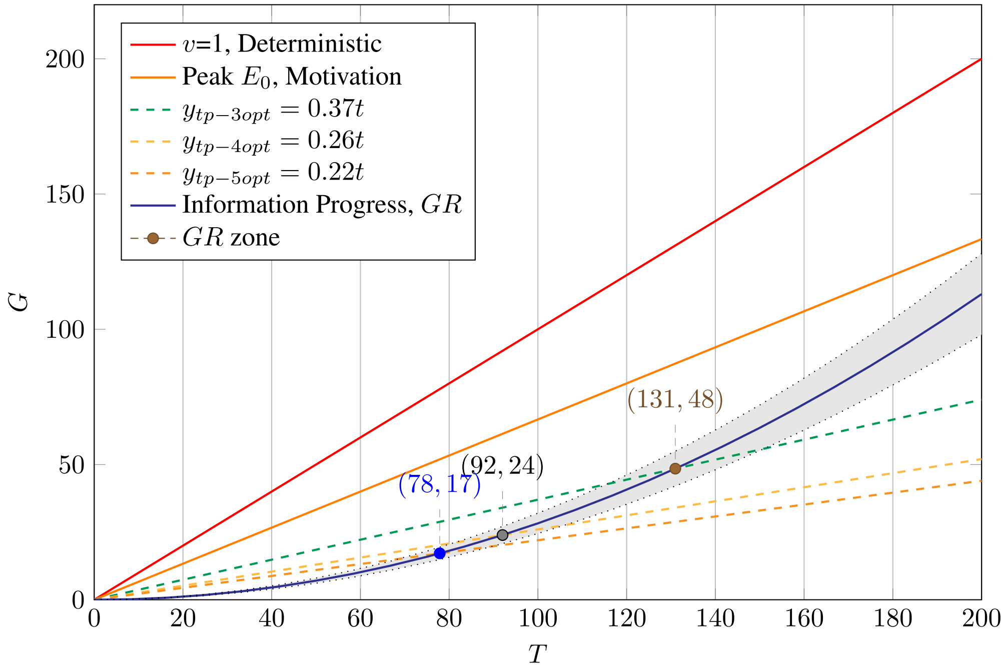


FIGURE 5. An illustration of guessing progression with with changes in score (*G*) and its difficulty (Δ*v*) over total score (*T*).

progression with changes in the score (G) and its difficulty (Δv) over the total score (T) in each test format applying time pressure. The results from the illustration present a suitable configuration for the gamified test, which, assuming the correct response, rewards 1 point. The total scores would be approximately 131, 92, and 78 for 3-options, 4-options, and 5-options, respectively. Because the number of questions was fixed, *GR* values are determined by the total time spent on the test, which explains the variation in mass (m). Reduced time or input time pressure affects the increase of m, and participants can evaluate themselves in the proper test length, explaining the drop in *GR*. As shown in Table 8, the results indicate that time pressure lowers the success rate (v_0) while *GR* emerges gamified experience. Such a situation is similar to the game design principle that time pressures characterized the test to become more stochastic, similar to gameplay. Together these results provide important insights into the goal of gamification, which suggests an impact on student engagement.

2) SCAFFOLDING IN MCQ

Effective learning will engage students in a productive struggle that challenges but does not frustrate them. One approach by [80] advocates keeping students in a Zone of Proximal Development (ZPD) by pushing beyond individuals' growth mindset. Another scholar allows the participants to request and ask for support by hints and keep individual effective learning to struggle with novel difficulties [81]. The scaffolding technique would stretch the potential of learning beyond the level the participants can do on their own. One feature of assessing the scaffolding technique is that it allows participants to learn to execute these tasks independently. It immediately follows that giving students the most challenging tasks they can accomplish with scaffolding resulted in substantial learning gains.

TABLE 9. Motion in mind measures three multiple-choice question formats with exclusion and inclusion of scaffolding.

		Time*	#Qs	v_0	m	N	E_0	$\vec{p}$
Exc.	Pre	15	15	0.48	0.53	2.1	0.2414	0.2497
	Post	15	15	0.66	0.34	1.5	0.2963	0.2231
	Total	30	30	0.57	0.43	1.7	0.2806	0.2446
Inc.	Pre	15	15	0.59	0.41	1.7	0.2859	0.2416
	Post	15	15	0.75	0.25	1.3	0.2802	0.1861
	Total	30	30	0.67	0.33	1.5	0.2962	0.2203

#Qs: number of questions; Exc.: exclusion; Inc.: inclusion;

In Table 9, we provide the quantitative measurement by using motion in mind comparatively for two groups of adaptive difficulty; scaffolding inclusion (Inc) and exclusion (Exc) — the number of question items was fixed at 30. Participants are asked to complete the test separately in 15 minutes for the first part and 15 minutes for the second part ($t_1 = 15$, $t_2 = 15$, $t = t_1 + t_2 = 30$). The experiment will distribute five items for the 3-options format, five for the 4-options format, and five for the 5-options format in the first half of the adaptive test; the rest will be distributed depending on the performance of the first part. The experiment flow has been inspired by the Computerized Adaptive Test (CAT), which is more suited to execution at the initial stage of studying the topic to diagnose the student's initial level of knowledge, and further improvement by presenting tasks of optimal difficulty. There are now three possibilities for creating an algorithm for an adaptive system [82]. First, the pyramidal approach was used as a criterion, where each student was given a medium-difficulty task. Then, based on the response, the following assignment is generated, the scale of which is lower or higher by two times [83].

This study generalized the adaptive algorithm to be a simpler version by operating moderate-level question items to participants. Evaluation criteria depend on the performance of the first part; the second part is formed by shifting difficulty, which is easier or harder. For example, the adaptive system would distribute more accessible question items if the performance was below a lower bound. Likewise, the adaptive system will distribute more challenging question items if the performance is above an upper bound. This study's upper and lower bounds were specified at 75% and 50%, respectively. When computing the measurement, we see that the inclusion of scaffolding significantly improves knowledge even in challenging tasks, which explains a decrease in N. The adaptive system will provide a practical challenge to the participants to enable the zone of proximal development process. This condition implies that the result from the first part certainly improves in the second part. The objective velocity v_0 will increase in the second part; also, the velocity will be improved if the participants gain knowledge, which observes in the inclusion of scaffolding. The results depict $v_0 = 0.573$ and $N = 1.7$ (Exclusion), while $v_0 = 0.0.672$ and $N = 1.5$ (Inclusion). These effects indicate that inclusion is better for novice-level participants when we assume ability level $k = 3$.

3) HYBRID SYSTEM

Per our hypotheses, learners can maximize their learning potential if we assign a task at the further limits of their ZPD. It might depend on the challenge and support, which affect learning potential and engagement [55]. The properties of challenge-based gamification and scaffolding can sharpen mass m in either increasing or decreasing based on the motion in mind concept. This experiment allows the participants to request and ask for support through hints and keep individual effective learning to struggle with novel difficulties. However, participants were reinforced by time pressure to observe the concurrent situation's impact. The scaffolding would extend potential learning beyond the level, whereas time pressure will maintain the balance between skill and chance as the goal of gamification. The finding allows us to see the impact of both learning and engagement aspects. Table 10 provides the experimental data of variation of MCQ by time pressure, scaffolding, and the hybrid system. The results show a practical cycle of emotion and performance state, in which the mass m will shift between arousal and control zone following the flow theory. This situation enables the final approximate above the flow zone, which implies optimal arousal. These views surfaced mainly concerning ZPD, where the challenge level corresponds to the individual ability. As a result, the hybrid system maintains the scoring rate v_0, which is related to the fairness perspective, which explains a balance between skill and chance.

TABLE 10. Motion in mind measures three multiple-choice question formats with all variations.

Types of Experiments		v_0	m	N	E_0	$\vec{p}$
Time Pressure	3-options	0.37	0.63	2.7	0.1700	0.232
	4-options	0.26	0.74	3.9	0.0994	0.1919
	5-options	0.22	0.78	4.6	0.0735	0.1697
Scaffolding	Exclusion	0.57	0.43	1.7	0.2806	0.2446
	Inclusion	0.67	0.33	1.5	0.2962	0.2203
Hybrid		0.49	0.51	2.0	0.2447	0.2499

The results depict the refinements of velocity v_0 to 0.49. The objective velocity v_0 aligns at the point where a test taker is required for assistance to tackle more challenging problems. In addition, reward frequency N became 2, which means the participants have to choose between incorrect and correct options. This condition shows that momentum would be 0.249, which implies fairness in the participant performance. When $k = 3$, m is located in the boundary between social comfort ($m = \frac{1}{3}$) and play comfort ($m = \frac{2}{3}$), which defines the hybrid system's learning potential. If novice participants joined, they would gain knowledge from this system. The value of mass m satisfies that participants with $k \in [2, 4]$ are reasonable for the hybrid test due to its simplicity in prolonging learning and achieving mastery.

D. VARIATION OF SCORING METHODS IN MCQs

The traditional scoring rule for multiple-choice questions is the Number of Right (NR) rule, in which the test score is simply the number of correct responses multiplied by some constant. Initially, a considerable concern with this rule is the guessing of responses. A guessed answer may be valid. Therefore, test takers can achieve points for questions even if they have yet to learn the answer. In addition, guessing adds random falsehood to the variability of the test score, which downsizes the reliability of the test. Many possible scoring rules can eliminate and reduce the guessing effect in MCQ.

Generally, Scholastic Aptitude Test (SAT) has applied the scoring rule, subtracting points for incorrect answers. Furthermore, no points are accounted for omitting a question. This condition allows participants to guess than omit an answer if they have partial knowledge. Under uncertain situations, some factors affect decision-making and the response to answers. This situation relates to the general terminology of decision-making based on prospect theory [59]. The basis originated from a descriptive model of risky choice, which integrates expected value and utility theory. This aspect leads to preferences that depend on the 'framing effect.' If an outcome is considered as a gain, decision-makers will be risk-averse to maintaining their outcome. Otherwise, decision-makers will be risk-seeking to solidify their results if an outcome is regarded as a loss.

The proposed experiment aimed to demonstrate that the ‘framing effect’ increases participants’ tendency to guess answers in MCQ. We conducted two scoring methods: the *positive* and *mixed* rules. In all experiments, the dominant strategy was to answer, even without knowledge. This situation implies that the experiments were conducted assuming that the expected guessing score equals the expected score of omitting. Individuals have some knowledge, and then the probability of choosing the correct answer is $> \frac{1}{N}$, and the expected score for guessing will necessarily be greater than 0. They might omit or guess according to their current conditions if faced with challenging questions.

For the *mixed* rule, the scoring rule is that 1 point is given for a correct response, $\frac{1}{N-1}$ points are deducted for each incorrect answer (where N denotes the number of options), and no points are gained or lost for omitting a question. On the other hand, *positive* rules prevent the deduction of points for incorrect responses; 1 point is given for a correct response, $\frac{1}{N}$ points are provided for each omission, and no points are given for an incorrect answer. Nonetheless, the two scoring methods result in distinct score framing since the *mixed* rule includes both gains and losses, whereas the *positive* rule only incorporates gains.

Based on the prospect theory, the tendency to guess is affected by the framing effect, which can be manipulated through how test takers achieve the score. The simple strategy is to try to achieve higher scores as much as they can until it reaches the particular point where the framing effect is active. This experiment compared the tendency to answer question items for three relevant rules; the number of rights (NR), Positive, and Mixed. The results from the NR were obtained from the previous experiment, both human and simulation data. Table 11 shows the positive and mixed rules from the experiment with human and simulation data results (by running Algorithm 2 and Algorithm 3). We found that the positive rule provides the best success rate, while the mixed rule provides the least success rate v_0. According to Table 11, there was a significant difference among scoring methods. Regarding reward frequency $N=\frac{1}{v_0}$, it was found that participants faced a comfortable way in the positive rule since there was no deduction, then they would consider less choice with the minimal risk. This situation affects the number of options during test taking. The guessing effect depends on the individual’s risk aversion or seeking requirement since framing effect results are potentially vital and may be challenging to eliminate.

IV. DISCUSSION

This study suggests that the number of options in MCQ suits the individual at different levels of ability, the impact of various styles of MCQ by incorporating gamification and scaffolding, and optimal strategy in the variation of scoring methods. Furthermore, since the motion-in-mind concept also contributed subjective and objective numerical results, the methodology could be strengthened to incorporate human and simulation data to overcome some of the method’s limitations. First, we observed and computed the psychometric properties of our question items. Then, item analysis was conducted by classical test theory to determine the difficulty index of each item and test and compute the discrimination index between high-cognitive and low-cognitive participants. Lastly, this study verifies the validity and reliability using Pearson’s correlation coefficient and Kuder-Richardson 20 Formula ($KR-20$), respectively.

Algorithm 2 MCQ AI With Various Levels of Knowledge in Positive Scoring

```
Data: Assign Key Option with 1 point, Distractor
      with 0 point, question items $I$
$v_0$:= success rate;
initialization;
$Sample \leftarrow 2400$;
$I \leftarrow 300$;
$N \leftarrow \{3, 4, 5\}$ ;        /* 3, 4, 5 options */
$TotalPoints \leftarrow 0$;
for $x \in Sample$ do
    $Sum[x] \longleftarrow 0$;
    for $y \in I$ do
        $p_1 \leftarrow$ randomProbability() $\in [0.33, 1]$
        ;                        /* 3-options */
        $p_2 \leftarrow$ randomProbability() $\in [0.25, 1]$
        ;                        /* 4-options */
        $p_3 \leftarrow$ randomProbability() $\in [0.20, 1]$
        ;                        /* 5-options */
        keyOption() $\leftarrow \{p1, p2, p3\}$;
        if $p_1$ or $p_2$ or $p_3 \geq 0.33$ or $0.25$ or $0.20$ then
            $A[y] \leftarrow A$ Selected Option;
            if $1 \in A[y]$ then
                $Sum[x] \leftarrow Sum[x] + 1$;
            else
                $Sum[x] \leftarrow Sum[x] + 1/N$;
        else
            $Sum[x] \leftarrow Sum[x] + 0$;
    $TotalPoints \leftarrow TotalPoints + Sum[x]$;
$Avg \leftarrow \frac{TotalPoints}{Sample}$;
$v_0 \leftarrow \frac{Avg}{N}$;
return $v_0$;
```

Our proposed questions are considered functional items, which explains the difficulty index and discriminability. Comparisons of the difficulty index in three formats delivered significant differences with a general tendency for the difficulty of these formats to increase with the change from 3-options to 5-options observable. Results indicated that more than 50% of question items are moderate, while most 5-options items are difficult. This condition supports that the larger number of options items were consistently more difficult than the fewer options. The change in the number

TABLE 11. Motion in mind measures the variation of scoring methods between human and simulation data.

Scoring	3-options				4-options				5-options			
Methods	Human		Simulation		Human		Simulation		Human		Simulation	
	v_0	m	v_0	m	v_0	m	v_0	m	v_0	m	v_0	m
NR	0.73	0.27	0.61	0.39	0.53	0.47	0.49	0.51	0.33	0.67	0.41	0.59
Positive	0.78	0.22	0.65	0.35	0.58	0.42	0.53	0.47	0.4	0.6	0.51	0.49
Mixed	0.66	0.34	0.33	0.67	0.44	0.56	0.37	0.63	0.25	0.75	0.40	0.60

Algorithm 3 MCQ AI With Various Levels of Knowledge in Mixed Scoring

```
Data: Assign Key Option with 1 point, Distractors
      with -$\frac{1}{N-1}$ point, question items $I$
$v_0$:= success rate;
initialization;
$Sample \leftarrow 2400$;
$I \leftarrow 300$;
$N \leftarrow \{3, 4, 5\}$ ;        /* 3, 4, 5 options */
$TotalPoints \leftarrow 0$;
for $x \in Sample$ do
    $Sum[x] \longleftarrow 0$;
    for $y \in I$ do
        $p_1 \leftarrow$ randomProbability() $\in [0.33, 1]$
         ;                      /* 3-options */
        $p_2 \leftarrow$ randomProbability() $\in [0.25, 1]$
         ;                      /* 4-options */
        $p_3 \leftarrow$ randomProbability() $\in [0.20, 1]$
         ;                      /* 5-options */
        keyOption() $\leftarrow \{p1, p2, p3\}$;
        if $p_1$ or $p_2$ or $p_3 \geq 0.33$ or $0.25$ or $0.20$ then
            $A[y] \leftarrow$ selectedOption();
            if $1 \in A[y]$ then
                $Sum[x] \leftarrow Sum[x] + 1$;
            else
                $Sum[x] \leftarrow Sum[x] + 0$;
        else
            $Sum[x] \leftarrow Sum[x] - \frac{1}{N-1}$;
    $TotalPoints \leftarrow TotalPoints + Sum[x]$;
$Avg \leftarrow \frac{TotalPoints}{Sample}$;
$v_0 \leftarrow \frac{Avg}{N}$;
return $v_0$;
```

of options affects the overall item performance, suggesting that the number of options in MCQ is a significant factor in influencing test construction. Likewise, the discrimination analysis found no significant difference in overall items; only 19 indicated poor discriminability. The number of options affects the guessing rate whenever the difficulty level of each question is higher than individual ability. These results verify the practicality of our question items, which could be proven by item analysis.

The comparison of concurrent validities based on the TOEIC score found no significant differences between the correlations with the experimental test scores derived from the MCQ. However, a high correlation was found for the 4-options test compared to the 3-options and 5-options formats. Also, internal consistency was ensured, determined by $KR - 20$. There were some significant differences in the $KR-20$ reliability coefficients of the three formats. Primarily, the 5-options format provides the least $KR - 20$ reliability, with the coefficients for the format relatively lower and more widely spread due to the lower number of items on which the calculation of the coefficients was founded. These findings align with the previous studies [10] and others, that number of options reflects the impact in terms of validity, reliability, and difficulty.

A. THE OPTIMAL NUMBER OF OPTIONS IN MCQ

This study is to determine the optimal number of options in MCQ and the characteristics of test formats by comparing human and simulation data. We conducted 3-options, 4-options, and 5-options formats starting from the traditional way to variation of MCQ. The finding indicates that the optimal number of options depends on learners' knowledge. Increasing the number of options generally decreases the success rate, which implies difficulty in the learner aspect. Due to random choice options in simulation, the algorithm tends to produce varying results based on the individual's ability level in each stem. There is a difference between human and simulation data since the ability level of participants might be higher than the standard. In practice, if a skillful participant encounters a challenging question, a participant will seek to solidify the optimal strategy by avoiding guessing in the risk situation with unclear information to enlarge their score margin. This condition implies that the process generates the outcomes depending on the initial individual condition, ability, and risk assessment. It relates to the nature of the motion mind concept, which increases in m value. At the same time, it simplifies understanding the subjective matter more deeply to strengthen the motion-in-mind concept.

The motion in mind measures indicated significant shifts between the success rate v_0 of 3-options, 4-option, and 5-options formats when applying knowledge. According to the guessing terminology, the small number of options would be a good choice for the beginner level. Two options are typically stochastic events where people have a half chance to achieve a correct response. The results indicated that the 5-options format is the most challenging MCQ format. In other words, the study's results suggested that increasing the number of options in MCQ to 4 and 5 made the test more complex, like expanding the reward frequency. With a decrease in the number of options, the test became easier (fewer options and greater chance rate). Fewer options result in a relatively more straightforward test because there is a higher chance of selecting the correct responses [84], [85]. Significantly, this would be much easier if individuals avoided blind guessing and adopted educated or informed guessing due to their partial knowledge. Using more options decreases item and test scores in plausible distractors, similar to our question pools. Participants' scores significantly reduced when the number of options increased. This condition relates to the motion-in-mind concept where the success rate v_0 decreased due to the difficulty increase m. However, it will be effective in the case of a well-prepared MCQ in terms of discriminability and reliability; then, 3-options would be good enough and perform as the 4-options format.

The results of this study verify the findings regarding the number of options in MCQ that test takers preferred 3-options for assessment in their entrance examination and performed poorly in 5-option and 6-options due to extraneous factors such as test anxiety [85]. They generally feel anxious, which increases when encountering such a complex situation. This condition relates to [86] that 3-options were optimal for test takers, especially at the medium level of ability, which underlies the ability level is inversely proportional to the risk rate. Consequently, the 3-options format provides evidence for the highest motivation, where E_0 is maximized. This situation implies that individuals begin in such tests, which are generally easy to try and do not move the individuals' minds; thus, they tend to be highly attractive and provide curiosity, which is used in most simple quizzes. However, the 3-options format may be optimal depending on the point of view taken – from the test score users or the test stakeholders supported by the relevant work in CSAT analysis [85]. More options most likely increase the possibility that the test can evaluate additional or actual skills from individuals. The conservative position would be that the optimal number of options depends on the specific testing context, which can be beneficial for assessing individuals or establishing the test. This condition emphasizes that test developers or stakeholders should consider several factors by linking theoretical concepts to consider evaluation impacts when deciding the optimal number of options.

In addition, tests taken by students meet the definition of a standardized test, in which everyone takes an identical test, time, under the same circumstances. Such tests are often considered fairer and more objective than a system in which some students get a more straightforward test and others get a more difficult one. However, in general, the tests are criticized by some people. The aim is to construct a well-standardized MCQ that measures the abilities of practical applications to solve a particular problem and remove flaws. The findings align with the previous study by [33] that a more considerable number of options per question would increase the content of variability of test results. This condition is required in high-stakes testing in which they have to be assessed a broad range of subjects and evaluate the competencies of a larger group of candidates. Most tests are found in the language proficiency test, license test, or university examination, with the majority still using 4-options and 5-options formats, as seen in Table 12 [87]. It is apparent from this table that very few MCQs constructed on fewer options are much less prevalent in the academic environment. With these premises, the 4-options and 5-options are popular in standardized tests, reducing the guessing chance (improves test quality) and reliability by the number of options.

From the relevant results, the number of options in MCQ characterizes competitiveness, a high number of options regarding difficulty, and an ability level. The 3-options format provides a greater chance of random guessing v_0 and such a learning comfort over 4-options and 5-options. Such a debate suggests that beginners taking MCQ with more options are better for improving themselves in the sense of competitiveness. More options tests are suitable for the learner whose ability level is high, requiring more comprehensive skills to reduce the plausible distractor. Lowering options may be optimal in such a test with less time, such as a quiz, to better the skill being measured; otherwise, they will lose engagement. Such information also clarifies what results from implementing changes suggested by quantitative investigations into the optimal number of multiple-choice item options in a given context.

This analysis provided empirical evidence to advance understanding of the number of options in proficiency tests and national examinations. Motion in mind measure was used to calculate the statistics in popular tests and examinations. Table 13 compares summary statistics for the test by using motion in mind. A closer inspection of the illustrations shows the values $m \in [0.33, 0.5]$, except for GAT tests. There are changes in the number of options for SAT in 2017 and GAT in 2015. SAT lowers the number of options from 5 to 4, while GAT increases from 4 to 5. Only no change was revealed in SAT since the test is much more predictable and provides the exact amount of time and number of questions compared with the old version; thus, individuals can prepare themselves and estimate the strategy in the test. Whereas GAT was opened for applicants only one time after 2017, and there are many tests accounted for in the admission system. Therefore, they want to assess students more accurately to effectively screen and not burden them since they would take several examinations in one year. They adjusted to the 5-options format to assess the students and reformat, which reflects stability in v_0. However, the quality of the test should be considered since the scoring rate v_0 is very low, and it is a skill-driven test. It could be stated that the GAT test still needs improvements. The most exciting aspect is that GR values were under 0.07, except for GAT. This situation implies that these tests are suitable

TABLE 12. List of standardized tests with number of options in MCQ and types of scoring methods.

Notable Multiple Choice Tests	Number of Options	Scoring Method
TOEIC Listening	3,4	NR
TOEIC Reading	4	NR
TOEFL iBT	4	NR
Scholastic Aptitude Test (SAT) – Before 2017	5	Mixed
Scholastic Aptitude Test (SAT) – After 2017	4	NR
Graduate Record Education (GRE)	5	Negative
Common Law Admission Test (CLAT)	5	Mixed
American College Testing (ACT)	4, 5	NR
Indonesian National Exam	3, 4, 5	NR
Law School Admission Test	5	NR
O-NET (National Examination Test)	4	NR
GAT-PAT (Thailand) – Before 2017	4	NR
GAT-PAT (Thailand) – 2017-Current	5	NR
Japan Language Proficiency Test (JLPT)	4	NR
Test of Proficiency in Korean (TOPIK)	4	NR
Medical College Admission Test	4	NR
National Center Test for University Admissions	5	NR
Driver's License Tests (Japan)	4	NR
Driver's License Tests (Thailand)	4	NR
College Scholastic Ability Test (CSAT)	5	Mixed
Graduate Management Ability Test(GMAT)	5	Mixed

TABLE 13. Motion-in-mind measures of the notable multiple-choice tests [87].

Notable MCQs	G	T	Options	v_0	m	E_0	GR
GAT 1/2013	52.43	150	4	0.35	0.65	0.1589	0.0483
GAT 2/2013	59.26	150	4	0.40	0.60	0.1888	0.0513
GAT 1/2014	49.91	150	4	0.33	0.67	0.1477	0.0471
GAT 2/2014	51.78	150	4	0.35	0.65	0.1561	0.0480
GAT 1/2015	40.39	150	5	0.27	0.73	0.1060	0.0424
GAT 2/2015	45.14	150	5	0.30	0.70	0.1266	0.0448
GAT 1/2016	40.36	150	5	0.27	0.73	0.1058	0.0424
GAT 2/2016	45.34	150	5	0.30	0.70	0.1275	0.0449
GAT 2017	46.35	150	5	0.31	0.69	0.1320	0.0454
GAT 2018	53.63	150	5	0.36	0.64	0.1643	0.0488
GAT 2019	55.09	150	5	0.37	0.63	0.1707	0.0495
GAT 2020	52.43	150	5	0.35	0.65	0.1589	0.0483
GAT 2021	43.64	150	5	0.29	0.71	0.1200	0.0440
Reading TOEIC 2016	262	495	4	0.53	0.47	0.2637	0.0327
Reading TOEIC 2017	261	495	4	0.53	0.47	0.2629	0.0326
Reading TOEIC 2018	259	495	4	0.52	0.48	0.2611	0.0325
Reading TOEIC 2019	265	495	4	0.54	0.46	0.2663	0.0329
Reading TOEIC 2020	323	495	4	0.65	0.35	0.2959	0.0363
Reading TOEIC 2021	279	495	4	0.56	0.44	0.2773	0.0337
Listening TOEIC2016	317	495	3,4	0.64	0.36	0.2950	0.0360
Listening TOEIC2017	320	495	3,4	0.65	0.35	0.2955	0.0361
Listening TOEIC2018	321	495	3,4	0.65	0.35	0.2956	0.0362
Listening TOEIC2019	323	495	3,4	0.65	0.35	0.2959	0.0363
Listening TOEIC2020	337	495	3,4	0.68	0.32	0.2959	0.0371
Listening TOEIC2021	331	495	3,4	0.67	0.33	0.2963	0.0368
SAT 2012	514	800	5	0.64	0.36	0.2952	0.0283
SAT 2013	514	800	5	0.64	0.36	0.2952	0.0283
SAT 2014	513	800	5	0.64	0.36	0.2950	0.0283
SAT 2015	511	800	5	0.64	0.36	0.2948	0.0283
SAT 2016	508	800	5	0.64	0.37	0.2944	0.0282
SAT 2017	527	800	4	0.66	0.34	0.2962	0.0287
SAT 2018	531	800	4	0.66	0.34	0.2963	0.0288
SAT 2019	528	800	4	0.66	0.34	0.2962	0.0287
SAT 2020	523	800	4	0.65	0.35	0.2960	0.0286
SAT 2021	528	800	4	0.66	0.34	0.2962	0.0287

for educational purposes, where curiosity is reasonable for education.

B. THE IMPACT OF CHALLENGES AND SUPPORT ON THE MOTIVATION AND PERFORMANCE

The results of this study show that gamification of challenges in testing and quizzing engages students in both learning and entertainment. The results also imply that a challenge is a crucial component and mechanic in game design, offering a foundation for further application in the context of education. According to [49], one of the essential works in this field was motivated by the flow concept, which holds that a player's skill level increases when they are more engaged and have a better understanding of the challenge in a game. To bridge the gap between physics and psychology, we investigated the activity's ideal degree of challenge-based gamification. We outlined the individual's position using the ideas of motion in mind and flow theory. The study's findings highlight the variety in challenge-based gamification, including time constraints and difficulty aspects. The outcomes also demonstrate that challenge-based gamification significantly changed the value of m. This condition implies that subjectivity can increase while complexity or uncertainty can increase competition. A balance between uncertainty and ability is necessary to emphasize their importance in education, especially in a conceptual learning environment.

Furthermore, this study involves the concept of the zone of proximal development [50], which is a part of the sociocultural theory of learning, explaining how deliberate intellectual activity develops from social and cultural influences. In particular, it is considered a critical contribution to the field of education. This study investigated how to integrate the scaffolding-based concept into the test. The result showed that scaffolding decreased the changes in value m. This situation could be described by the motion in mind and zone of proximal development to bridge the gap between physics and learning theory. When $m \leq 0.5$, which is a more intuition-driven way, these results show that the ability is shifted from arousal zone to flow zone that supports the conjecture of gamified learning theory by [40]. Its transition from the challenging zone to the control zone indicates learning potential in which a combination of two approaches can enhance learner ability. In an assumption, Vygotsky suggested that it would be best to give students the most challenging tasks they can do with scaffolding as this will lead to tremendous learning gains. This situation means the challenge level must be greater than the ability level, where risk ratio (m) and velocity (v) indicate challenge and ability, respectively. The feasible outcome can assist the potential significance of gamification and scaffolding in learning and engagement.

The effects of time pressure on decision-making under uncertainty were investigated, and the finding depicted the affective state, information process, and task structure in decision-making. The findings showed that time pressure produces anxiety [88]. In contrast, the effort was reinforced by integrating different decision-making behaviors to cope with the task conditions. References [89] and [90] find that time pressure changes individual attitudes toward risk situations. However, time pressures benefit students by requiring explicit decision-making, but they also affect decision performance and suboptimal decisions, which restrict the acquisition of new knowledge or strategies [91]. This condition extends from the finding that individuals respond differently to time pressure when required to handle the situation, as this essential may change the optimal decision strategy. Another study by [92] showed that time pressure induces a student to perform an activity more efficiently. In contrast, another study claimed that time pressure changes how people explore and respond to uncertainty [93].

According to motion in mind, the potential energy value E_0 is primarily located around $m = 0.33$, which denotes the least resistance to informational acquisition. Once time constraints are in place, students face challenging circumstances that encourage them to exert more effort and alter their behavior. A sense of curiosity and uncertainty is indicated by this circumstance, which influences engagement and paves the way for learning-related outcomes. Promoting engagement and the student's emerging potential is the fundamental concept behind increasing the impact of learning (i.e., acquiring information). The sense of engagement contributing to behavior change emerges because gamification does not directly improve learning outcomes [94]. So, to contribute to the learning outcome, learning-related behavior is introduced as an analogous bridge. In contrast to focusing solely on the activity result, [40] emphasizes the development of engagement, which helps to change behaviors and generally translates into better learning behavior. A strategy for maximizing and determining learning-related outcomes is scaffolding. Students may exhibit their best performance with the most outstanding competence to achieve learning-related outcomes once a moderate challenge occurs. The results of the motion-in-mind value ($v0$, m, and $E0$) reveal such a discussion.

From the experiment, it was found that mass m significantly increased with time pressure. This condition implies that a larger m can foster greater motivation, which verifies the previous study by [43] that conjectured that time pressure would provide the intensity of competitive aspects, forcing them to dominate discomfort in the activity. On the other hand, the inclusion of scaffolding by giving hints will improve the success rate v_0. Our findings express the decrease in the number of options participants will encounter, showing the elimination process when gaining enough information. Since the average number of options in the practical test was around four, the inclusion of scaffolding verifies that it aids participants in solving the challenging questions. Furthermore, scaffolding improves the value of m by nearly 0.33, which is optimal for novice learners. This situation implies more chance to progress in the test with high curiosity (high objective energy E_0).

The combination of the two approaches allows learners to experience that the intensity was reasonable because of the perceived fairness. These consequences show that time pressure and scaffolding shifted to the flow zone in which individual ability and activity challenge is equal. Learners will be situated as gameplay, which implies $N = 2$ (where $m = 0.5$). The exploring area with $\frac{1}{k} \leq m \leq \frac{1}{2}$ implies the occurrence of curiosity (encouraged to learn); hence people would explore it. This situation means Δ_k is maximized, corresponding to the greater scoring rate and stochastic event motion, implying that the momentum of the test provides the comfort of balance between skill and chance. This region enlarges learner ability so that it is conjectured as a learning comfort at $k = 4$. However, the score may rely on the learner's difficulty and complexity during different occasions, which affects the value m. Learners will gain knowledge by focusing on reinforcers (gamified elements) and avoiding mistakes (using hints). This situation drives value m to reach the balance between chance and skill, denoted as low competitiveness with fairness and maintaining learning comfort. This condition was aligned with the support that the educational context can achieve learning and engagement, represented by $1.5 \leq N \leq 2$. Figure 6(a) and Figure 6(b) outlined the interpretation of challenge-based gamification and scaffolding based on the motion-in-mind concept. A general discussion could be the possible presence of challenge-based gamification to support this explanation using the motion-in-mind concept, which affirms the balance of education and entertainment proposed by [40]. This condition also verifies the theory by Vygotsky of the zone of proximal development. The optimal arousal region is denoted as $m \in [0.5, 0.67]$ for $k = 3$. Their potential mechanism can draw the alternative paradigm in the test, directing us to the education context's novel assessment and evaluation method.

C. DECISION MAKING PROCESS UNDER THE FRAMING EFFECT

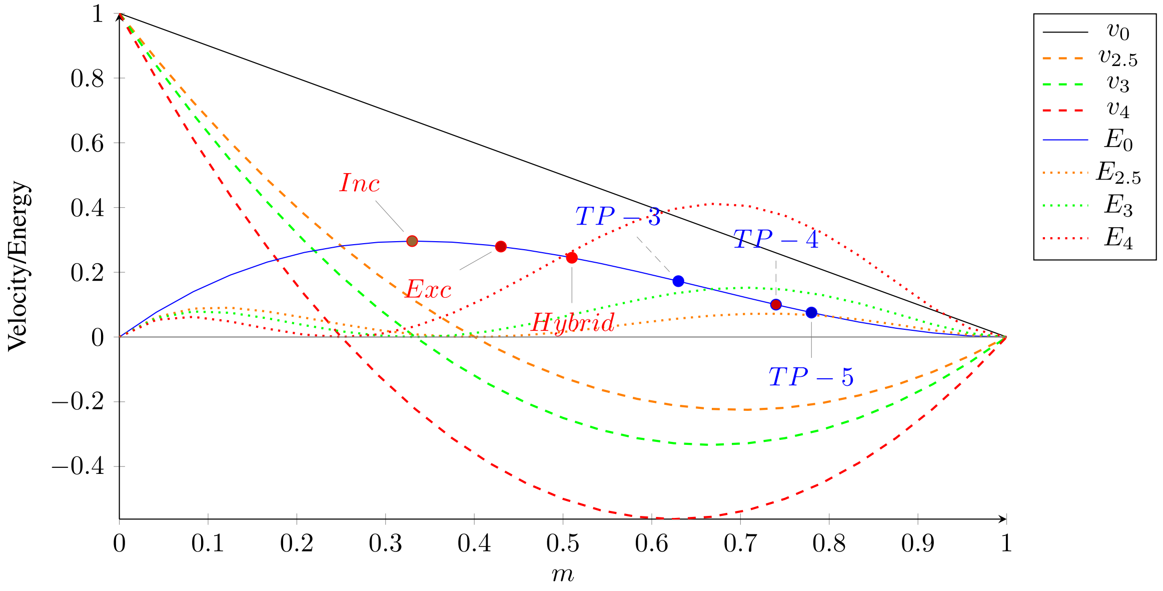
The current study focuses on decision-making under subjective uncertainty, which focuses on precisely the effect of framing on MCQ, referring to the order or manner in which a decision-maker presents a choice or option. This situation is required to choose the best option among all possible options based on the expected outcome, which is explained by gain or loss. While there are numerous alternative instances in which uncertainty is internal, little is revealed about the decision process that underpins these decisions. The two scoring methods focused on responses to MCQ as an instance of this decision dilemma. It can be detected that there is the highest variance between human and simulation data in the mixed rule comparing the two results. This condition suggests that the framing effect may sensitively cause decision-making due to including gain and loss characteristics. Our proposed experiment is based on an implicit cognitive model of participant behavior in MCQ scenarios under prospect theory. This study underlies the framing effect of the scoring method on an individual's behavior, and it was hypothesized that variation in the scoring methods would add the probabilities to represent various behavior patterns. Then, the expected value would be increased due to the change in the decision.

Prospect theory aims to provide a thorough explanation for selecting between objective probability in monetary gambles [95]. Theoretically, the common perception is that prospect theory indicates risk-seeking, where individuals perceive loss, while risk-averse, where they perceive gain. This condition implies that the gain scenarios will persuade individuals to be cautious in their decisions if there is a loss on any mistakes. It illustrates that prospect theory can account for choice behavior even when probabilities are not presented. Because practical instances in which the option is explicitly stated are scarce, this theory is vital for interpreting these scenarios. Scoring rules were initially oriented to assess the individuals' actual scores better. It was hypothesized that participants to do differently with tests that employ scoring rules. The experiment revealed that participants did satisfactorily in the positive rule and did more flawed in the mixed rule. The results indicate that the scoring rule strongly affects the decision in which the success rate v_0 is varied based on the benefits. As predicted from the theory, the success rate v_0 was more significant in the positive rule than in the mixed rule. This situation suggests that it allows individuals to learn and engage in the learning comfort, where $m \leq \frac{1}{k}$.

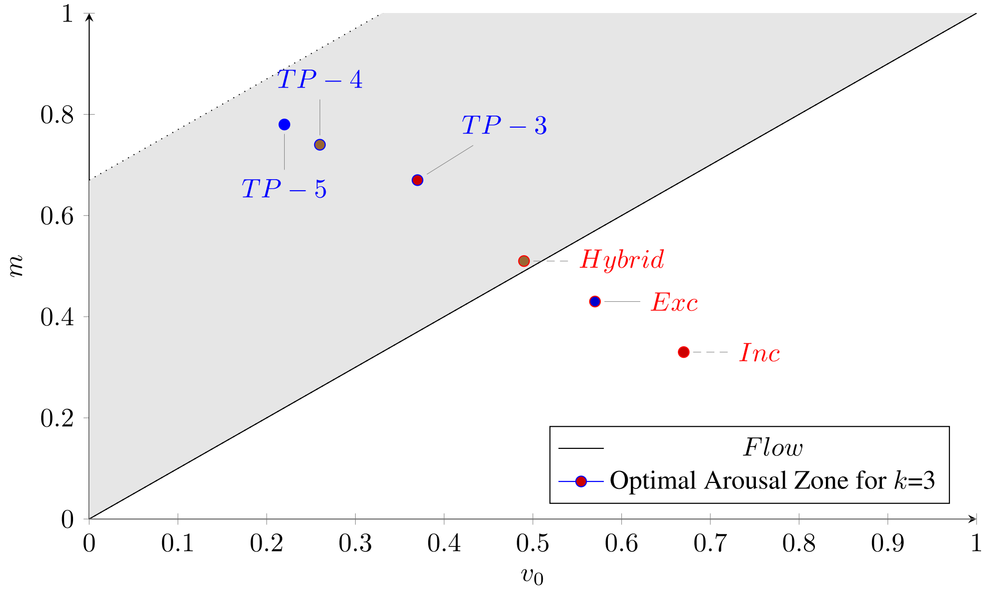
The characteristics of decision-making processes may shape how they perform during the test, and different scoring methods are likely to alter the patterns and strategies of answering. This situation lowers the test's validity and provides some advantages for the particular group with a specific ability. Participants tend to be risk-averse in the gambling situation, whereas they tend to be risk-seeking if the outcome is worth gaining. Participants will reveal their actual performance in the mixed rule and engage in a gamble in which a loss is a possible outcome. Participants must choose between remaining in the same position with no gain or at least gambling by guessing. It drives participants to guess rather than omit since the remaining position provides less expected value than gambling.

The results showed that the mixed rule is the largest of m, which implies high-stake testing, which requires skill-driven, implying that this rule is required sufficient skills to overcome the risk. Therefore, most ability tests choose the mixed rule to assess the test takers that satisfy the minimum requirements and desired ability. The tendency to omit the positive rule is greater than the tendency to guess if the participants do not know. The tendency to guess in the positive rule will be greater than in the mixed rule since there is no loss scenario. Specifically, it was found that no standardized tests use the positive rule since the scores may be greater than usual. This situation reflects that velocity v_0 in the positive rule is greater than in the mixed rule. Therefore, participants will at least omit the question, and assessing their ability from this test takes work.

Their behavior in test events will be similar to decision makers' responses in other situations where people make decisions under ambiguity. Every decision has a consequence, in which the scoring rule varies, yielding results and a tendency to answer. However, a study on behavioral decisions may introduce biases since it is in line with subjective measurement. Penalty reduces motivation and inputs more stress, which explains an increase of value m and a reduction in E_0. This condition would be hard to indicate a good performance if the framing effect is becoming robust, then the participant will omit it to avoid penalization. The findings support the work from [14] that it is worth abandoning penalties in conventional tests. This condition is why the number of rights (NR) rule is still popular nowadays. This analysis will likely provide valuable contributions to



(a) Measures of motion in mind of variation in MCQs by applying challenge-based gamification and scaffolding for $k = \{2.5, 3, 4\}$



(b) An illustration between v_0 and m of challenge-based gamification and scaffolding related to the flow concept and the zone of proximal development

FIGURE 6. Interpretation of variation MCQs based on the motion-in-mind measures.

decision-making in an educational context by focusing on the input-output affinity of consequences of decision when taking a test.

D. INTERPRETATION OF m WITH RESPECT TO LEARNING COMFORT From MOTION IN MIND PERSPECTIVE

Table 14 shows various motion-in-mind measures of popular board games and MCQ with 3-options, 4-options, and 5-options, as well as sports games. The table shows a clear trend of *GR* values in the standardized tests below 0.07. However, the experimental results indicated *GR* values in the sophisticated zone. There was significant evidence that 4-options MCQ provided fairness while conserving its *GR* value of nearly 0.05. This condition implies the feature as same as in board games where skill is an essential element of play. 4-options MCQs primarily depend on skills that can be made known (or controlled) to influence the test outcome (i.e., level of knowledge of an individual). Therefore, 4-options and 5-options MCQs are best practices in the standardized test, while these formats are consistent when the number of questions or score range is getting bigger. To identify the optimal number of options and learning comfort, understanding the reward frequency N, risk frequency ratio m, and motivation E_p are essential. Since the value of N can be regarded as the measure of the number of options, it serves as an indicator to determine the optimal number of options that border between learning and competitive comfort. Hence, analysis of the changes of N provides insights into MCQ terminology and captures its subjective interpretation concerning m value. This study aligns the interpretation based on adoption from [68].

From our results, the authors conjecture on four parameters: E_k, Δ_k, v_k, and δ_k. Subjective energy E_k was applied to demonstrate motivation for various individuals' ability levels to determine the learning comfort from mass m. As for the objective measures, it was found that participants struggle with test potential that quantifies by objective energy E_0. Curiosity can be defined in this context by calculating the difference between objectivity and subjectivity. Δ_k becomes the distance between objectivity and subjectivity, which is the reinforcement difference (E_0-E_k). It would be optimized by specifying two peaks when the success rate v_0 is ranged between the maximum threshold of objective reinforcement E_0 at $v = 0.67$ and the subjective reinforcement E_k at $v_0 = \frac{k-2}{k}$ where the objective reinforcement E_0 or reinforcement difference Δ_k will be maximized with the peak at the mass $m = 1/3$ and $m = 2/k$ respectively.

In option analysis, simulation data was computed from various ability levels of individuals by indicating $N = \{3, 4, 5\}$. It was found that the risk chance m was situated at 0.39 (3-options), 0.51 (4-options), and 0.59 (5-options), respectively. This situation reflects that three test formats are promising for learning since the reinforcement differences are optimized in the range. The analogical studies of popular board games and the evolution of multiple-choice questions analyzed these results. 4-options and 5-options require skillful participants to overcome, which explains changes in ability level k. This result states the potential of an optimal number of options in the popular tests. The effect of the number of options is related to the variable ratio (VR) described in the motion-in-mind concept. Sophisticated board games such as Go, Shogi, and Chess have distinct values of N and represent different aspects [68]. Go is aligned $N = \frac{5}{3}$ at $m = 0.4$, in which the reward frequency is high and requires a low ability to move the game. Shogi would relate to a situation that induces high curiosity (motivated effort) to a situation related to mass entertainment since $N = 3$. The region beyond this area, like Chess ($N = 5$), requires the ability and effort to push the game. Such a region implies competitiveness, which is often motivated. However, sports games provide N in the competitive region, which implies a chance game.

Table 15 and Table 16 depict the analogical interpretation of value m compared to game context and implication in the educational context based on the motion in mind concept. The deterministic region implies mastery in the education aspect, where $v_0 = 1$. The learning comfort holds the objective reinforcement E_0 dominating over the subjective one E_k, where Δ_k is maximized at peak E_0. This condition verifies the region where learning comfort was identified, with a strong reinforcement and curiosity at $\frac{1}{k} \leq m \leq 0.5$, while the momentum of test δ_k becomes bigger. A minor degree of uncertainty assures repeating the action, making it appealing but stochastic. This scenario is analogous to the game design principle of access to learning but challenging to master [96]. Fairness is situated at $m = 0.5$, where δ_k is maximized. This situation implies the balance between ability and chance, where test takers encounter correct and incorrect answers. Then, solving uncertainty is optimized by challenge-based gamification. This situation endorses that time pressure accumulates risk ratio m, which affects the required skills to maintain learning comfort. Participants would feel pressure depending on the amount of time; therefore, the performance will reduce when the time is decreased. Furthermore, potential energy E_0 reduces, which reflects inhibiting reinforcement. Participants noticed that the test would be challenging to accomplish desired outcomes, which explains a decrease of v_k. Reinforcement difference Δ_k nearly becomes 0, which satisfies the harmonic balance between objectivity and subjectivity at $E_0 = E_k$. This point is denoted as learning engagement. While the momentum of test δ_k becomes small. It would be referred to as the competitive zone where such tests move an individual's mind. This region possesses a challenge equivalent to the player's ability. In a sense, this underlines the level of knowledge that an individual must attain to overcome challenging tasks. This condition can be seen by negative peak v_k and peak E_k, denoted as optimal arousal point and competitive zone, respectively.

Figure 7 compares various motion-in-mind measures of variation MCQs. The 4-options tests highlight an ideal measurement that balances chance and skill. Then the number of rights is famous for involving in the tests. The least competitiveness was determined in the number of right (NR) rule where δ_k is nearly maximized. This condition verifies that most standardized tests technically employ the 4-options format, including the English proficiency tests and national tests at the high school level. On the other aspects, potential energy E_0 is maximized in the scaffolding-based test. It would be better if beginners attempted to learn the test contents. Individuals may learn from what they know nothing about until they achieve the potential learning. With this conjecture, this study reassures that scaffolding is a good option for learning and developing their ability while playing, as shown in [97] and [98]. This condition affirms that scaffolding can yield potential development based on the notion of ZPD. A hybrid system (based on [55]) that simultaneously included time pressure and scaffolding indicated that m value was improved until relatively fair $m = 0.5$, and the reward frequency between individuals and options is stochastic. The ability difference evolves larger than in other systems. This framework finely highlights the aspects that individuals would feel learning comfort due to its fair properties and versatility compared to other tests.

TABLE 14. Comparison of various motion-in-mind measures of popular games and sports (adopted from [63]) including MCQs.

Games/Sports/MCQs	G, B	T, D	m	E_0	GR	$\vec{p}$	N_{avg}
Chess	35	80	0.78	0.0748	0.0740	0.1709	5
Shogi	80	115	0.65	0.1578	0.0778	0.2268	3
Go	208	250	0.58	0.2021	0.0577	0.2429	2.5
Basketball	36.38	82.01	0.56	0.2190	0.0735	0.2468	2.5
Soccer	2.64	22	0.88	0.0253	0.0739	0.1056	9
3-Options	56.25	77	0.27	0.2878	0.0974	0.1971	1.5
4-Options	89.65	170	0.47	0.2640	0.0557	0.2491	2
5-Options	17.92	53	0.67	0.1459	0.0799	0.2211	3
3-Options*	46.97	77	0.39	0.2902	0.0890	0.2379	1.5
4-Options*	83.3	170	0.51	0.2449	0.0537	0.2499	2
5-Options*	21.73	53	0.59	0.1984	0.0880	0.2419	2.5
GAT 2020	52.43	150	0.65	0.1589	0.0483	0.2274	3
GAT 2021	43.64	150	0.71	0.1200	0.0440	0.2063	3
Listening TOEIC2020	337	495	0.32	0.2959	0.0371	0.2173	1.5
Listening TOEIC2021	331	495	0.33	0.2963	0.0368	0.2215	1.5
Reading TOEIC2020	282	495	0.43	0.2793	0.0339	0.2451	2
Reading TOEIC2021	279	495	0.44	0.2773	0.0337	0.2460	2
SAT 2020	523	800	0.35	0.2960	0.0286	0.2264	1.5
SAT 2021	528	800	0.34	0.2962	0.0287	0.2244	1.5

TABLE 15. The analogical interpretation of value *m* compare to game context.

m	Indication	Game context	Educational context (MCQ)
0	v_0=1	Deterministic	Mastery
0.33	Peak E_0	Objective Motivation	Learning Comfort
$\frac{1}{k}$	v_k=0, E_k=0	High Reinforcement	Curiosity (Encourage to learn)
0.5	Peak δ_k	Fairness, Low Competitiveness	Balance between ability and chance
$\frac{2}{k}$	E_0=E_k	Mass Entertainment, Play Comfort	Learning Potential
$\frac{1+k}{2k}$	Negative peak v_k	Perceptive Turnover, Gamified	Flow to Optimal Arousal
$\frac{N-1}{N}$	$v = \frac{1}{N}$	Equiprobability	Uncertainty among N options (Guessing)

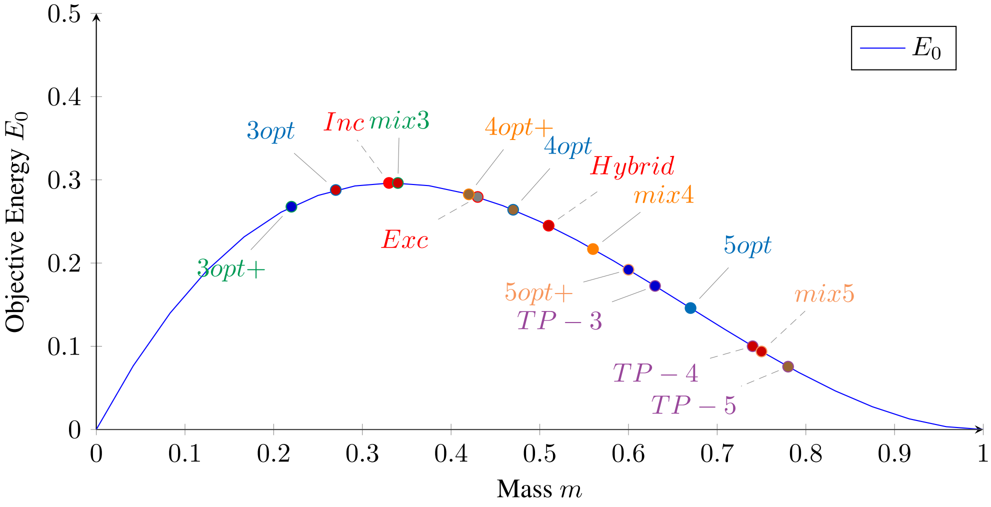


FIGURE 7. Comparison of variation in MCQs based on motion in mind measures, E_0.

According to the scoring methods, this study piloted the experiment to determine the relationship between decision-making and reinforcement. A decision maker tends to be risk-averse or risk-seeking depending on how the minimum expected outcome is represented in terms of risk chance and knowledge. In general, decision-makers are risk averse when faced with positively framed difficulties [16] and risk-seeking when faced with negatively framed challenges. Meanwhile, the results depicted that the mixed rule provided the most unsatisfactory outcome compared to each other. The 5-options format with mixed rule contains the highest competitiveness (low δ_k), while the least competitiveness δ_k

TABLE 16. Implication of scoring difficulty (m^*) in the educational context based on motion in mind concept (aligned with [68]).

Range	m^*	Implication
E_k=0, max Δ_k	$\frac{1}{k}$	Learning Comfort
max δ_k	0.5	Fairness
E_0=E_k	$\frac{2}{k}$	Learning Potential
max E_k	$\frac{\sqrt{9(k^2+1)-2k}+3(k+1)}{10k}$	Competitive

in the 3-options. This situation implies that the mixed rule might be suitable to apply in the lower number of options to satisfy the learning comfort. The loss exploits larger when the number of options N increases since the risk chance m will become stronger. In this case, 3-options mixed rule is a reasonable alternative for producing learning comfort: beginners can take a test and learn at their own pace while maintaining engagement and high motivation. The 5-options mixed rule requires high skill to stay motivated since the test will be challenging. It implies the competitive zone where an individual's performance level should be robust.

E. LIMITATIONS AND FUTURE WORKS

There are two limitations found in the current study. First, the proposed experiment aimed to link between motion in mind and theoretical approach corresponding to several domains, such as psychological analysis and learning theories, to serve the interpretation in the educational context. The purpose of the experiment is to observe the impact of multiple-choice question characteristics and their variations (i.e., gamification, scaffolding, and scoring method). As such, the approach was confined. The experiments were designed to compare the number of options from 3-options to 5-options formats. In addition, the time pressure and scaffolding were loosely employed, which may affect different groups of participants differently. Our findings may require a more profound analysis that may work in different circumstances. The proposed algorithm might be improved to cope with participants with partial knowledge. This experiment focused on the input time pressure to observe the impact of time pressure in a test. However, one essential thing must be considered to determine the number of question items. This point could be mentioned in the test length, which may be essential to determine optimal time pressure in the practical test.

Secondly, although the internal validity and reliability were verified by classical test theory, external validity could not be achieved due to the small sample size of question items and participants for comparison with other practical cases. Likewise, the participants might have been affected by taking 300 items in pre-test measurement, so they would have behaved differently after being exposed to our approaches. Hence, these results may need to be consistent with other conceptual schemes. As such, this limitation suggests caution when interpreting the results since such findings were not externally validated, requiring further investigation. Item response theory is a better choice to analyze the test since this model includes parameters of difficulty index, discriminability, and a guessing parameter. These are necessary for estimating a valid relationship between the chance of a correct response to an item and the individual's ability. In addition, this condition would clarify the guessing terminology in the individuals encountered under unexpected circumstances. Future studies should also consider complex details when designing the test in educational settings.

V. CONCLUSION

This study investigated the effect of the number of options in multiple-choice questions using analysis of motion in mind theory to link theoretical approaches, including gamification, scaffolding, and prospect theory, by variation of test elements and scoring methods. The current study's findings suggest that increasing the number of options in MCQ makes the test more challenging, followed by increasing the m value. As with most research studies, the current study is originally subject to the guessing terminology, which assumes that individuals have no knowledge. Once applying partial knowledge, the number of plausible options would be specified by lowering the options and m value. The results from the current study indicate the changes in the success rate, which reflects suitable ability level k on taking our question items.

Following motion in mind theory, the interpretation of m concerning learning comfort was addressed. The evidence showed that the 4-options format characterizes fairness properties among others. In contrast, the 3-options format implies simplicity as owning non-competitiveness, which is suitable for an individual with a low ability to drive their effort at the maximum level of motivation. The 5-options format was the most difficult due to competitiveness, which is required for individuals with high abilities. This study concludes that 4-options and 5-options are the assessment through which challenge-based gamification can moderate gamified and competitive experiences and produce a learning-related behavior [99]. In light of the statement, the results reveal that it is entirely rational for most standardized tests usually attempted on the 4-options and 5-options format.

Our findings revealed that time pressure provides competitiveness since the challenge shifts individuals' states into an arousal zone, which explains an increase in m. This finding is consistent with the essence of flow theory, which develops if the challenge and students' abilities are fostered. This condition also arouses individuals to push their effort and allows different strategies to cope with challenging tasks. On the other hand, this study incorporated scaffolding-based hints to deliver knowledge to individuals, implying the presence of a learning process under a control state where the challenge was adequate. A lower m value allows individuals to improve their learning without extrinsic motivation, as described by the increases in velocity v. The proposed system improved the learning environment to reduce frustration and develop a genuine learning process, corresponding to the zone of proximal development, where individuals are scaffolded to do tasks beyond their ability. They will learn to do tasks independently. Our experiment in the scoring method also yields the generality of prospect theory in which individuals tend to answer under uncertainty when the gain is obtained. Our findings found that the number of rights and the mixed rule is worthwhile to consider applying in educational assessment. The framing effect is anticipated to contribute to our understanding and insight into educational and standardized tests. Regardless of the number of options and context, our findings can be an extension study for analyzing the balance between competitiveness and entertainment while seeking the learning process in the educational context. These findings enable scholars to improve educational assessment to increase stakeholders' motivation and engagement while also challenging and entertaining them.

REFERENCES

[1] T. M. Haladyna, *Developing and Validating Multiple-Choice Test Items*, 3rd ed. New York, NY, USA: Taylor & Francis, 2012, doi: 10.4324/9780203825945.

[2] R. M. Epstein, "Medical education assessment in medical education," *New England J. Med.*, vol. 356, pp. 387–396, Jan. 2007.

[3] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied Meas. Educ.*, vol. 15, no. 3, pp. 309–333, 2002.

[4] R. Moreno, R. J. Martínez, and J. Muñiz, "Directrices para la construcción de ítems de elección múltiple," *Psicothema*, vol. 16, no. 3, pp. 490–497, 2004.

[5] A. Hughes, *Testing for Language Teachers*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[6] F. Abad, J. Olea, and V. Ponsoda, "Analysis of the optimum number alternatives from the item response theory," *Psicothema*, vol. 13, no. 1, pp. 152–158, 2001.

[7] K. Woodford and P. Bancroft, "Multiple choice questions not considered harmful," in *Proc. Conf. Res. Pract. Inf. Technol.*, vol. 42, 2005, pp. 109–116.

[8] P. I. Nwadinigwe and L. Naibi, "The number of options in a multiple-choice test item and the psychometric characteristics," *J. Educ. Pract.*, vol. 4, no. 28, pp. 1–9, 2013.

[9] M. S. Trevisan, G. Sax, and W. B. Michael, "The effects of the number of options per item and student ability on test validity and reliability," *Educ. Psychol. Meas.*, vol. 51, no. 4, pp. 829–837, Dec. 1991, doi: 10.1177/0013164491051004O4.

[10] M. C. Rodriguez, "Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research," *Educ. Measurement: Issues Pract.*, vol. 24, no. 2, pp. 3–13, Jun. 2005, doi: 10.1111/j.1745-3992.2005.00006.x.

[11] J. Charan, B. Vegada, A. Shukla, A. Khilnani, and C. Desai, "Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study," *Indian J. Pharmacol.*, vol. 48, no. 5, p. 571, 2016, doi: 10.4103/0253-7613.190757.

[12] A. A. Margolis, "Zone of proximal development, scaffolding and teaching practice," *Cultural-Historical Psychol.*, vol. 16, no. 3, pp. 15–26, 2020, doi: 10.17759/chp.2020160303.

[13] A. Nardo, "Exploring a vygotskian theory of education and its evolutionary foundations," *Educ. Theory*, vol. 71, no. 3, pp. 331–352, Jun. 2021, doi: 10.1111/edth.12485.

[14] Y. Bereby-Meyer, J. Meyer, and D. V. Budescu, "Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules," *Acta Psychologica*, vol. 112, no. 2, pp. 207–220, 2003, doi: 10.1016/S0001-6918(02)00085-9.

[15] A. Carvalho, S. Dimitrov, and K. Larson, "On proper scoring rules and cumulative prospect theory," *EURO J. Decis. Processes*, vol. 6, nos. 3–4, pp. 343–376, Nov. 2018, doi: 10.1007/s40070-018-0081-8.

[16] Q. Wu, M. Vanerum, A. Agten, A. Christiansen, F. Vandenabeele, J.-M. Rigo, and R. Janssen, "Certainty-based marking on multiple-choice items: Psychometrics meets decision theory," *Psychometrika*, vol. 86, no. 2, pp. 518–543, Jun. 2021, doi: 10.1007/s11336-021-09759-0.

[17] L. W. Schuwirth and C. P. V. D. Vleuten, "ABC of learning and teaching in medicine," *Brit. Med. J.*, vol. 326, no. 7390, pp. 643–645, 2003.

[18] R. Jabrayilov, W. H. M. Emons, and K. Sijtsma, "Comparison of classical test theory and item response theory in individual change assessment," *Appl. Psychol. Meas.*, vol. 40, no. 8, pp. 559–572, Nov. 2016, doi: 10.1177/0146621616664046.

[19] R. C. Foster, "A generalized framework for classical test theory," *J. Math. Psychol.*, vol. 96, Jun. 2020, Art. no. 102330, doi: 10.1016/j.jmp.2020.102330.

[20] J. D. Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*. New York, NY, USA: McGraw-Hill, 2005.

[21] J. M. Kilgour and S. Tayyaba, "An investigation into the optimal number of distractors in single-best answer exams," *Adv. Health Sci. Educ.*, vol. 21, no. 3, pp. 571–585, Aug. 2016, doi: 10.1007/s10459-015-9652-7.

[22] I. Thanyapa and M. Currie, "The number of options in multiple choice items in language tests: does it make any difference? Evidence from Thailand," *Lang. Test. Asia*, vol. 4, no. 1, p. 8, Dec. 2014, doi: 10.1186/s40468-014-0008-7.

[23] A. D. Garvin and R. L. Ebel, "Essentials of educational measurement," *Educ. Researcher*, vol. 9, no. 9, p. 21, Oct. 1980, doi: 10.2307/1175572.

[24] R. L. Thorndike and E. P. Hagen, *Measurement and evaluation in psychology and Education*, vol. 60, 4th ed. Boston, MA, USA: Educational Leadership, 2002.

[25] J. B. Grier, "The number of alternatives for optimum test reliability," *J. Educ. Meas.*, vol. 12, no. 2, pp. 109–112, Jun. 1975, doi: 10.1111/j.1745-3984.1975.tb01013.x.

[26] A. M. Andrés and J. D. L. Castillo, "Multiple choice tests: Power, length and optimal number of choices per item," *Brit. J. Math. Stat. Psychol.*, vol. 43, no. 1, pp. 57–71, May 1990, doi: 10.1111/j.2044-8317.1990.tb00926.x.

[27] A. Tversky, "On the optimal number of alternatives at a choice point," *J. Math. Psychol.*, vol. 1, no. 1, pp. 386–391, 1964, doi: 10.1016/0022-2496(64)90010-0.

[28] J. E. Bruno and A. Dirkzwager, "Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective," *Educ. Psychol. Meas.*, vol. 55, no. 6, pp. 959–966, Dec. 1995, doi: 10.1177/0013164495055006004.

[29] D. V. Budescu and B. Nevo, "Optimal number of options: An investigation of the assumption of proportionality," *J. Educ. Meas.*, vol. 22, no. 3, pp. 183–196, Sep. 1985, doi: 10.1111/j.1745-3984.1985.tb01057.x.

[30] T. Shizuka, O. Takeuchi, T. Yashima, and K. Yoshizawa, "A comparison of three- and four-option English tests for university entrance selection purposes in Japan," *Lang. Test.*, vol. 23, no. 1, pp. 35–57, Jan. 2006, doi: 10.1191/0265532206lt319oa.

[31] P. Baghaei and N. Amrahi, "The effects of the number of options on the psychometric characteristics of multiple choice items," *Psychol. Test Assessment Model.*, vol. 53, no. 2, pp. 192–211, 2011.

[32] A. Dehnad, H. Nasser, and A. F. Hosseini, "A comparison between three- and four-option multiple choice questions," *Procedia Social Behav. Sci.*, vol. 98, pp. 398–403, May 2014, doi: 10.1016/j.sbspro.2014.03.432.

[33] M. Panczyk, H. Rebandel, and J. Gotlib, "Comparison of four- and five-option multiple-choice questions in nursing entrance tests," in *Proc. 7th Int. Conf. Educ., Res. Innov. (ICERI)*, 2014, pp. 4131–4138.

[34] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining 'gamification,'" in *Proc. 15th Int. Academic MindTrek Conf., Envisioning Future Media Environments*, Sep. 2011, pp. 9–15, doi: 10.1145/2181037.2181040.

[35] C. Dichev and D. Dicheva, "Gamifying education: What is known, what is believed and what remains uncertain: A critical review," *Int. J. Educ. Technol. Higher Educ.*, vol. 14, no. 1, pp. 1–36, Dec. 2017, doi: 10.1186/s41239-017-0042-5.

[36] T. H. Laine and R. S. N. Lindberg, "Designing engaging games for education: A systematic literature review on game motivators and design principles," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 804–821, Oct. 2020, doi: 10.1109/TLT.2020.3018503.

[37] S. Nicholson, "Exploring gamification techniques for classroom management," *Games Learn. Soc.*, vol. 9, pp. 21–27, Jul. 2013.

[38] G. Surendeleg, V. Murwa, H.-K. Yun, and Y. S. Kim, "The role of gamification in education—A literature review," *Contemp. Eng. Sci.*, vol. 7, pp. 1609–1616, 2014, doi: 10.12988/ces.2014.411217.

[39] D. R. Sanchez, M. Langer, and R. Kaur, "Gamification in the classroom: Examining the impact of gamified quizzes on Student learning," *Comput. Educ.*, vol. 144, Jan. 2020, Art. no. 103666, doi: 10.1016/j.compedu.2019.103666.

[40] R. N. Landers, ''Developing a theory of gamified learning,'' *Simul. Gaming*, vol. 45, no. 6, pp. 752–768, Dec. 2014, doi: 10.1177/1046878114563660.

[41] S. Farcas and I. Szamosközi, ''The effects of working memory trainings with game elements for children with ADHD. A meta-analytic review,'' *Transylvanian J. Psychol.*, vol. 15, no. 2, pp. 21–44, 2014.

[42] A. I. Canhoto and J. Murphy, ''Learning from simulation design to develop better experiential learning initiatives: An integrative approach,'' *J. Marketing Educ.*, vol. 38, no. 2, pp. 98–106, Aug. 2016, doi: 10.1177/0273475316643746.

[43] P. Anunpattana, M. N. A. Khalid, H. Iida, and W. Inchamnan, ''Capturing potential impact of challenge-based gamification on gamified quizzing in the classroom,'' *Heliyon*, vol. 7, no. 12, Dec. 2021, Art. no. e08637, doi: 10.1016/j.heliyon.2021.e08637.

[44] A. C. T. Klock, I. Gasparini, M. S. Pimenta, and J. Hamari, ''Tailored gamification: A review of literature,'' *Int. J. Hum.-Comput. Stud.*, vol. 144, Dec. 2020, Art. no. 102495, doi: 10.1016/j.ijhcs.2020.102495.

[45] N.-Z. Legaki, N. Xi, J. Hamari, K. Karpouzis, and V. Assimakopoulos, ''The effect of challenge-based gamification on learning: An experiment in the context of statistics education,'' *Int. J. Hum.-Comput. Stud.*, vol. 144, Dec. 2020, Art. no. 102496, doi: 10.1016/j.ijhcs.2020.102496.

[46] A. Denisova and P. Cairns, ''Adaptation in digital games: The effect of challenge adjustment on player performance and experience,'' in *Proc. Annu. Symp. Comput.-Hum. Interact. Play*, Oct. 2015, pp. 97–101, doi: 10.1145/2793107.2793141.

[47] C.-C. Chang, C. A. Warden, C. Liang, and G.-Y. Lin, ''Effects of digital game-based learning on achievement, flow and overall cognitive load,'' *Australas. J. Educ. Technol.*, vol. 34, no. 4, pp. 1–13, Sep. 2018, doi: 10.14742/ajet.2961.

[48] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, and T. Edwards, ''Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning,'' *Comput. Hum. Behav.*, vol. 54, pp. 170–179, Jan. 2016, doi: 10.1016/j.chb.2015.07.045.

[49] J. Nakamura and M. Csikszentmihalyi, ''The concept of flow,'' in *Flow and the Foundations of Positive Psychology*. Cham, Switzerland: Springer, 2014, pp. 239–263, doi: 10.1007/978-94-017-9088-8_16.

[50] L. S. Vygotsky, *Mind in Society: Development of Higher Psychological Processes*. Cambridge, MA, USA: Harvard Univ. Press, 1978, [Online]. Available: http://www.jstor.org/stable/j.ctvjf9vz4, doi: 10.2307/j.ctvjf9vz4.

[51] A. Kozulin, ''The zone of proximal development in Vygotsky's analysis of learning and instruction,'' in *Vygotsky's Educational Theory in Cultural Context*, A. Kozulin, B. Gindis, V. S. Ageyev, and S. M. Miller, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2003, doi: 10.1017/CBO9780511840975.004.

[52] I. Verenikina, ''Understanding scaffolding and the ZPD in educational research,'' Fac. Arts, Social Sci. Humanities, Univ. Wollongong, Wollongong, NSW, Australia, Tech. Rep., 2003.

[53] G. Wells. (1999). *Dialogic Inquiry*. [Online]. Available: https://doi.org/10.1017/cbo9780511605895

[54] M. Cole, ''Cultural psychology: a once and future discipline?'' in *Proc. Nebraska Symp. Motivat.*, 1989, p. 37, doi: 10.1097/00005053-199809000-00012.

[55] R. Wass and C. Golding, ''Sharpening a tool for teaching: The zone of proximal development,'' *Teach. Higher Educ.*, vol. 19, no. 6, pp. 671–684, 2014, doi: 10.1080/13562517.2014.901958.

[56] R. M. Silalahi, ''Understanding vygotsky's zone of proximal development for learning,'' *Polyglot, Jurnal Ilmiah*, vol. 15, no. 2, p. 169, Aug. 2019, doi: 10.19166/pji.v15i2.1544.

[57] G. G. Kravtsov and E. E. Kravtsova, ''Relationship between learning and development: Problems and perspectives,'' *Cultural-Historical Psychol.*, vol. 16, no. 1, pp. 4–12, 2020, doi: 10.17759/chp.2020160101.

[58] E. B. Braaten, ''Vygotsky's zone of proximal development,'' in *The SAGE Encyclopedia of Intellectual and Developmental Disorders*. Newbury Park, CA, USA: Sage, Feb. 2018, doi: 10.4135/9781483392271.n533.

[59] D. Kahneman and A. Tversky, ''Prospect theory: An analysis of decision under risk,'' *Exp. Environ. Econ.*, vol. 47, no. 2, p. 263, Mar. 1979, doi: 10.2307/1914185.

[60] A. Chiu, G. Wu, J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, and J. C. Smith, ''Risk-taking in international politics, chapter: Prospect theory,'' in *Wiley Encyclopedia of Operations Research and Management Science*. Ann Arbor, MI, USA: Univ. of Michigan Press, 2010.

[61] M. O. Rieger, M. Wang, and T. Hens, ''Estimating cumulative prospect theory parameters from an international survey,'' *Theory Decis.*, vol. 82, no. 4, pp. 567–596, Apr. 2017, doi: 10.1007/s11238-016-9582-8.

[62] H. Iida, K. Takahara, J. Nagashima, Y. Kajihara, and T. Hashimoto, ''An application of game-refinement theory to Mah Jong,'' in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3166, 2004, doi: 10.1007/978-3-540-28643-1_41.

[63] H. Iida and M. N. A. Khalid, ''Using games to study law of motions in mind,'' *IEEE Access*, vol. 8, pp. 138701–138709, 2020, doi: 10.1109/ACCESS.2020.3012597.

[64] H. Iida, ''Fairness, judges and thrill in games,'' SIG, Neuhausen am Rheinfall, Switzerland, Tech. Rep. 28, 2008, pp. 61–68.

[65] A. P. Sutiono, A. Purwarianti, and H. Iida, ''A mathematical model of game refinement,'' in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering LNICST*, vol. 136, 2014, pp. 148–151, doi: 10.1007/978-3-319-08189-2_22.

[66] S. Agarwal, M. N. A. Khalid, and H. Iida, ''Game refinement theory: Paradigm shift from performance optimization to comfort in mind,'' *Entertainment Comput.*, vol. 32, Dec. 2019, Art. no. 100314, doi: 10.1016/j.entcom.2019.100314.

[67] B. Hoffman and L. Nadelson, ''Motivational engagement and video gaming: A mixed methods study,'' *Educ. Technol. Res. Develop.*, vol. 58, no. 3, pp. 245–270, 2010. [Online]. Available: http://www.jstor.org/stable/40603176

[68] M. N. A. Khalid and H. Iida, ''Objectivity and subjectivity in games: Understanding engagement and addiction mechanism,'' *IEEE Access*, vol. 9, pp. 65187–65205, 2021, doi: 10.1109/ACCESS.2021.3075954.

[69] B. F. Skinner, ''Selection by consequences,'' *Behav. Brain Sci.*, vol. 7, no. 4, pp. 477–481, Dec. 1984, doi: 10.1017/S0140525X0002673X.

[70] H. P. P. Aung, M. N. A. Khalid, and H. Iida, ''What constitutes fairness in games? A case study with scrabble,'' *Information*, vol. 12, no. 9, p. 352, Aug. 2021, doi: 10.3390/info12090352.

[71] Z. Zhang, K. Xiaohan, M. N. A. Khalid, and H. Iida, ''Bridging ride and play comfort,'' *Information*, vol. 12, no. 3, p. 119, Mar. 2021, doi: 10.3390/info12030119.

[72] H. Davis and J. Waycott, ''Ethical encounters: HCI research in sensitive and complex settings,'' in *Proc. Annual Meeting Austral. Special Interest Group Comput. Human Interact.*, 2015, pp. 667–669.

[73] H. Parveen and N. Showkat, ''Research ethics,'' Aligarh Muslim Univ., Aligarh, India, Tech. Rep., 2017, pp. 1–12.

[74] N. Salkind. (2012). *Encyclopedia of Measurement and Statistics*. [Online]. Available: https://doi.org/10.4135/9781412952644

[75] R. Vyas and A. Supe, ''Multiple choice questions: A literature review on the optimal number of options,'' *Nat. Med. J. India*, vol. 21, pp. 130–133, Jun. 2008.

[76] S. Yaman, ''The optimal number of choices in multiple-choice tests: Some evidence for science and technology education,'' *New Educ. Rev.*, vol. 23, no. 1, pp. 227–241, 2011.

[77] R. L. Brennan, ''Generalizability theory and classical test theory,'' *Appl. Meas. Educ.*, vol. 24, no. 1, pp. 1–21, Dec. 2010, doi: 10.1080/08957347.2011.532417.

[78] I. Himelfarb, ''A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating,'' *J. Chiropractic Educ.*, vol. 33, no. 2, pp. 151–163, 2019.

[79] *Cronbach's Alpha—A Measure of the Consistency Strength*. Accessed: Dec. 19, 2022. [Online]. Available: https://www.bachelorprint.eu/statistics/cronbachs-alpha/

[80] B. A. Ermeling, ''Pivotal moments in teaching: Zoom in on specific points to create meaningful learning,'' *Learning Prof.*, vol. 39, no. 3, pp 28–32, Jun. 2018.

[81] N. Delson, L. Van Den Einde, E. Cowan, and B. Mihelich, ''Mini-hints for improved spatial visualization training,'' in *Proc. ASEE Annu. Conf. Exposit.*, 2019, pp. 1–12, doi: 10.18260/1-2–33112.

[82] O. S. Diahyleva, I. V. Gritsuk, O. Y. Kononova, and A. Y. Yurzhenko, ''Computerized adaptive testing in educational electronic environment of maritime higher education institutions,'' in*Proc. CEUR Workshop*, vol. 2879, 2020, pp. 411–422.

[83] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy, *Computerized Adaptive Testing*. Evanston, IL, USA: Routledge, 2000, doi: 10.4324/9781410605931.

[84] M. Currie and T. Chiramanee, ''The effect of the multiple-choice item format on the measurement of knowledge of language structure,'' *Lang. Test.*, vol. 27, no. 4, pp. 471–491, Oct. 2010, doi: 10.1177/0265532209356790.

[85] H. Lee and P. Winke, "The differences among three-, four-, and five-option-item formats in the context of a high-stakes english-language listening test," *Lang. Test.*, vol. 30, no. 1, pp. 99–123, Jan. 2013, doi: 10.1177/0265532212451235.

[86] F. M. Lord, "Optimal number of choices per item- a comparison of four approaches," *J. Educ. Meas.*, vol. 14, no. 1, pp. 33–38, Mar. 1977, doi: 10.1111/j.1745-3984.1977.tb00026.x.

[87] J. Park, "Constructive multiple-choice testing system," *Brit. J. Educ. Technol.*, vol. 41, no. 6, pp. 1054–1064, 2010.

[88] A. J. Maule, G. R. J. Hockey, and L. Bdzola, "Effects of time-pressure on decision-making under uncertainty: Changes in affective state and information processing strategy," *Acta Psychologica*, vol. 104, no. 3, pp. 283–301, 2000, doi: 10.1016/S0001-6918(00)00033-0.

[89] M. G. Kocher, J. Pahlke, and S. T. Trautmann, "Tempus fugit: Time pressure in risky decisions," *Manage. Sci.*, vol. 59, no. 10, pp. 2380–2391, Oct. 2013, doi: 10.1287/mnsc.2013.1711.

[90] A. Bollard, R. Liu, A. Nursimulu, A. Rangel, and P. Bossaerts, *Neurophysiological Evidence on Perception of Reward and Risk: Implications for Trading Under Time Pressure*. Zürich, Switzerland: University of Zurich, 2007.

[91] A. Conte, M. Scarsini, and O. Ssrrcc, "Does time pressure impair performance? An experiment on queueing behavior," *SSRN Electron. J.*, vol. 11, pp. 260–274, Mar. 2015, doi: 10.2139/ssrn.2579053.

[92] S. Leroy, "Why is it so hard to do my work? The challenge of attention residue when switching between work tasks," *Organizational Behav. Human Decis. Processes*, vol. 109, no. 2, pp. 168–181, Jul. 2009, doi: 10.1016/j.obhdp.2009.04.002.

[93] C. M. Wu, E. Schulz, T. J. Pleskac, and M. Speekenbrink, "Time pressure changes how people explore and respond to uncertainty," *Sci. Rep.*, vol. 12, no. 1, Mar. 2022, doi: 10.1038/s41598-022-07901-1.

[94] C. Hursen and C. Bas, "Use of gamification applications in science education," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 14, no. 1, pp. 4–23, Jan. 2019, doi: 10.3991/ijet.v14i01.8894.

[95] P. E. Tetlock and B. A. Mellers, "The great rationality debate," *Psychol. Sci.*, vol. 13, no. 1, pp. 94–99, Jan. 2002, doi: 10.1111/1467-9280.00418.

[96] T. W. Malone, "Heuristics for designing enjoyable user interfaces: Lessons from computer games," in *Proc. Conf. Human factors Comput. Syst.*, 1982, pp. 63–68, doi: 10.1145/800049.801756.

[97] P. Hakkarainen and M. Bredikyte, "The zone of proximal development in play and learning," *Cultural-Historical Psychol.*, vol. 4, pp. 2–11, Feb. 2008.

[98] S. Siyepu, "The zone of proximal development in the learning of mathematics," *South Afr. J. Educ.*, vol. 33, no. 2, pp. 1–13, May 2013, doi: 10.15700/saje.v33n2a714.

[99] P. Garone and S. Nesteriuk, "Gamification and learning: A comparative study of design frameworks," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2019, pp. 473–487, doi: 10.1007/978-3-030-22219-2_35.

PUNYAWEE ANUNPATTANA received the B.Sc. degree from the Department of Electronics and Communication Engineering, Sirindhorn International Institute of Science and Technology, Thammasat University, in 2017, and the M.Sc. degree from the Japan Advanced Institute of Science and Technology (JAIST), in 2020, where he is currently pursuing the Ph.D. degree with the School of Information Science. He is also a member of the Research Center for Entertainment Science of Professor Hiroyuki Iida Laboratory. His specializations are game informatics, information theory, and human–computer interaction. His work focuses specifically on the development and improvement of gamified experience and platform transformation using game informatics and gamification techniques in the fields of education and entertainment technology.

MOHD NOR AKMAL KHALID (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the University of Science Malaysia, in 2013, 2015, and 2018, respectively. He is currently an Assistant Professor with the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), a member of the Research Center for Entertainment Science, from 2019 to 2020, and has been a member of the International Research Center for Artificial Intelligence and Entertainment Science, since 2022. His specializations are artificial intelligence techniques, game informatics, evolutionary computing and algorithms, and decision support systems. His work focuses specifically on the methods for optimization and game informatics in operation research and entertainment technology. His research interests include, but are not limited to, manufacturing systems, artificial intelligence techniques, game analysis and informatics, search algorithms, optimization techniques, advancement in scheduling and planning, and machine-learning methods.

HIROYUKI IIDA received the Ph.D. degree in heuristic theories on game-tree search from the Tokyo University of Agriculture and Technology, Tokyo, in 1994. He was with Shizuoka University, Hamamatsu, and a Guest Researcher with Maastricht University. He is currently a Japanese Computer Scientist and a Computer Games Researcher with a focus on game refinement theory, opponent model search, and computer Shogi. He is also the Trustee and the Vice President of Educational and Student Affairs with the Japan Advanced Institute of Science and Technology (JAIST), the Director of the Global Communication Center, the Director of the Research Center for Entertainment Science, from 2019 to 2020, has been a member of the International Research Center for Artificial Intelligence and Entertainment Science, since 2022, and the Head of the Iida Laboratory. His research interests include artificial intelligence, game informatics, game theory, mathematical model, search algorithm, game-refinement theory, game tree search, and entertainment science. He is also a professional 7-dan Shogi player and the coauthor of the Shogi Program Tacos, the four times Gold Medal Winner at Computer Olympiads. He is a member of the Board of the ICGA as a Secretary-Treasurer and a Section Editor of the *ICGA Journal*.