Systemic Analytics

www.sa-journal.org

Syst. Anal. Vol. 4, No. 1 (2026) 68–80.

Paper Type: Original Article

Behavioral Validity of Confidence-Based Knowledge State Reporting in Multiple Choice Examinations

Janardan Behera[1,*], Prabhudarshan Sahoo[2], Raja Kishaor Prusty[2]

[1] Department of Statistics, Ravenshaw University, Cuttack, India; janardan@ravenshawuniversity.ac.in.

[2] Department of Psychology, Ravenshaw University, Cuttack, India; prabhudarsansahoo@yahoo.com; rajaprustybgr@gmail.com.

Citation:

Received: 13 July 2025 *Revised: 09 November 2025* *Accepted: 19 January 2026*	Behera, J., Sahoo, P., & Kishaor Prusty, R. (2026). Behavioral validity of confidence-based knowledge-state reporting in multiple-choice examinations. *Systemic Analytics, 4*(1), 68-80.

Abstract

Multiple-choice examinations remain among the most widely deployed instruments of academic and professional assessment, yet their fundamental architecture invites systematic distortion: examinees who lack genuine knowledge may still earn credit through random selection among alternatives. Classical scoring mechanisms offer no incentive for examinees to reveal the quality of their knowledge, and consequently, the observed score conflates genuine knowledge with fortunate guessing. This article introduces and empirically investigates a structured self-reporting framework in which each examinee, for every item, declares one of three knowledge states, namely Full Knowledge (FK), Partial Knowledge (PK), or No Knowledge (NK), alongside their chosen answer. The scoring mechanism is constructed such that truthful declaration of one's epistemic state constitutes the uniquely optimal strategy in expectation, rendering guessing strictly suboptimal. The central empirical question is whether examinees, when placed within this incentive-compatible framework, do in fact report their knowledge states truthfully or whether systematic behavioral deviations, rooted in overconfidence, risk aversion, or strategic misrepresentation, emerge. Drawing on psychometric theory, Bayesian probability modeling, and the cognitive psychology of metacognition, this article formalizes the theoretical relationships between declared knowledge states and observed response accuracy, derives testable hypotheses, proposes an experimental design, and specifies a complete statistical inference framework for behavioral validation. The contribution is threefold: a formal probabilistic model linking epistemic-state declarations to correctness probabilities, a rigorous hypothesis-testing architecture, and an experimentally grounded methodology for assessing metacognitive honesty under incentive-compatible conditions.

Keywords: Multiple-choice examination, Knowledge state reporting, Metacognition, Incentive compatibility, Proper scoring rules, Calibration, Guessing behavior, Psychometrics.

1 | Introduction

The design of fair, valid, and knowledge-revealing examination systems has occupied a central position in educational measurement research for well over a century. Multiple choice questions, by virtue of their scoring

Corresponding Author: janardan@ravenshawuniversity.ac.in

https://doi.org/10.31181/sa41202672

objectivity and administrative convenience, dominate standardized assessment worldwide. Yet the psychometric literature has long recognized a structurally embedded problem. In the absence of any penalizing mechanism or epistemic accountability, the optimal rational strategy for an uninformed examinee is simply to guess, thereby obtaining nonzero expected credit irrespective of actual knowledge. Number-correct scoring assigns uniform credit to a correct response, whether it results from genuine recall, partial elimination of distractors, or pure chance. This conflation of knowledge and luck undermines score validity, distorts rank ordering, and renders the examination score a noisy and arguably biased proxy for the underlying latent trait it is meant to measure.

Prior remedies have followed two principal directions. The first is formula scoring, which subtracts a fraction of wrong answers from the count of correct ones, thereby making random guessing actuarially neutral rather than profitable. The second is confidence-weighted scoring, in which examinees assign numerical probabilities to each option and are scored according to proper scoring rules, such as the Brier score or the logarithmic rule. Both approaches carry limitations. Formula scoring eliminates the expected gain from guessing but does not differentiate between types of Partial Knowledge (PK); it treats a well-reasoned wrong answer the same as a random one, and its effectiveness depends critically on the assumption of risk neutrality, which behavioral evidence consistently refutes [1]. Confidence-weighted scoring based on continuous probability reports requires a level of probabilistic literacy that is unrealistic in most testing populations and imposes a significant response burden.

The framework proposed in the present article charts a different course. Rather than eliciting numerical probabilities, it asks examinees to classify their epistemic state into one of three ordered categories: Full Knowledge (FK), PK, or No Knowledge (NK). This trichotomy is grounded in cognitive realism. There is abundant evidence from the metacognitive literature that individuals can meaningfully distinguish between items they know with certainty, items on which they can eliminate some alternatives or reason approximately, and items on which they have no useful information whatsoever. The discrimination is, of course, imperfect, and the degree of imperfection is itself an empirically important quantity. The present article takes this imperfection seriously rather than assuming it away.

The scoring mechanism assigns state-dependent payoffs designed to make truthful reporting the dominant strategy. Under the FK declaration, a correct answer yields the maximum reward. In contrast, an incorrect answer under this declaration incurs a substantial penalty, reflecting the implausibility of error when genuine complete knowledge is claimed. Under PK, correct responses are rewarded moderately and incorrect responses penalized mildly, consistent with the probabilistic nature of partial information. Under NK, a correct response yields a small positive value, and an incorrect response yields zero or a mild penalty, since neither outcome is informationally diagnostic. The scoring parameters are constrained by incentive-compatibility conditions derived from expected utility theory, ensuring that, for a rational, risk-neutral examinee, truthful declaration weakly dominates all misreporting strategies.

The central contribution of this article is empirical. Knowing that the system is theoretically incentive-compatible is necessary but not sufficient for establishing its validity as an assessment tool. Whether examinees actually report truthfully is a behavioral question, and behavioral responses to incentive structures are known to deviate systematically from the predictions of classical rationality. Overconfidence, in which examinees declare FK when their actual accuracy is consistent with PK or NK, is among the most robustly documented findings in cognitive psychology and decision research. Underconfidence, though less common, also appears in high-stakes testing environments, particularly among high-achieving examinees who approach ambiguous items with epistemic caution. Risk aversion can induce NK declarations from examinees who have genuine PK but prefer the certainty of a small guaranteed payoff over the variance of the PK reward structure. These behavioral phenomena, if sufficiently pronounced, can compromise the validity of the knowledge state signal even when the scoring mechanism is theoretically sound.

2 | Literature Review

The question of whether individuals can accurately monitor and report their own cognitive states is one of the foundational inquiries of metacognitive research, a field whose systematic study was inaugurated by Flavell's pioneering work in the 1970s and has been substantially developed in the subsequent decades [2]. Metacognition, broadly construed, encompasses both the knowledge individuals hold about their own cognitive processes and the regulatory monitoring that governs those processes in real time. In the domain of memory and knowledge, the most studied metacognitive judgment is "feeling of knowing," which refers to the phenomenological sense of being able to retrieve information or recognize the correct answer even in the absence of immediate recall. The accuracy of feeling-of-knowing judgments, operationalized as the correlation between the declared judgment and subsequent retrieval or recognition performance, is positive but modest, varying substantially across individuals, domains, and testing conditions.

For the present framework, the critical metacognitive capacity is the ability to discriminate, at the moment of item response, among states of complete knowledge, PK, and ignorance. This discrimination maps onto the three declared states: FK, PK, and NK. The psychometric literature on confidence-based marking, developed most systematically by Gardner-Medwin and Gahan [3], Bruno and Dirkzwager [4], has consistently shown that examinees can make meaningful distinctions at a coarser level of confidence categorization than continuous probability reports demand and that these categorical confidence reports carry incremental predictive value for correctness beyond the answer choice itself. The present framework leverages this result.

The most extensively documented deviation from truthful self-reporting in the cognitive and judgment literature is overconfidence: the systematic tendency to express greater certainty in one's judgments than warranted by their accuracy rates. Overconfidence has been documented across a wide range of domains, including general knowledge questions, medical diagnosis, financial forecasting, and legal judgment. In the examination context, overconfidence would manifest as a tendency for examinees to declare FK when their actual correctness rate on those items is significantly below p_{FK} , or to declare PK on items where their correctness rate is closer to the random guessing baseline. The Overconfidence Bias Index, typically operationalized as the mean difference between declared confidence and observed accuracy rate, provides a tractable summary statistic that can be directly estimated from the experimental data in the present study [5].

Although less common, underconfidence also merits attention. Kruger and Dunning's [6] influential findings, subsequently refined by extensive follow-up work, suggest that high-achieving individuals sometimes underestimate their relative performance, which, in a testing context, could lead to a tendency toward PK or NK declarations on items where the examinee actually possesses complete knowledge. In the present scoring framework, underconfidence leads to a suboptimal expected score rather than a distorted ordinal ranking, and its prevalence in high-stakes academic populations is therefore of both theoretical and applied significance.

The incentive compatibility analysis presented assumes risk neutrality. Behavioral economics and decision theory have established beyond a reasonable doubt that human preferences deviate from risk neutrality in systematic ways. Kahneman and Tversky's [7] Prospect Theory, perhaps the most influential descriptive theory of choice under uncertainty, proposes that individuals are risk-averse in the domain of gains and risk-seeking in the domain of losses relative to a psychologically determined reference point. In the context of the present scoring system, an FK declaration involves both the highest potential gain and the highest potential loss, making it an intrinsically high-variance choice. A risk-averse examinee might therefore prefer a PK or NK declaration even when genuine FK is possessed, not because they are attempting to misrepresent their state but because the utility of the high-variance FK payoff structure is genuinely lower for them than the utility of the lower-variance alternatives. This type of behavioral deviation is qualitatively different from strategic misrepresentation and must be distinguished from it in empirical analysis.

Loss aversion, a component of prospect theory that has received particular empirical support, further complicates the picture. The penalty b associated with FK errors looms disproportionately large relative to

the reward a in the subjective utility calculation of a loss-averse examinee. This argument creates a bias toward conservative declarations, particularly FK-to-PK migration, which could systematically distort the knowledge state signal in ways that the expected-value analysis does not capture. The empirical investigation must therefore be designed to distinguish truthful reporting from risk-aversion-induced conservatism and from strategic misrepresentation motivated by an understanding of the scoring incentives.

A modest but growing body of empirical work has examined the effects of confidence-based and knowledge-state-based scoring on examinee behavior and score validity. Studies using continuous confidence ratings have generally found improved calibration relative to standard right-wrong scoring when proper incentives are in place. Still, the cognitive burden of eliciting numerical probabilities remains a practical limitation. Studies using binary confidence categorization (know versus guess) have found that the know-versus-guess distinction is behaviorally meaningful, with substantially higher correctness rates on known-declared items. Still, two-category systems capture less metacognitive information than a three-category scheme. The three-category scheme investigated here, with its formally grounded intermediate PK state, has not been rigorously validated in the psychometric literature, and the present article fills this gap [8].

3 | Theoretical Framework

3.1 | Knowledge State Definitions and Probabilistic Characterization

Let $\Omega = \{FK, PK, NK\}$ denote the set of possible knowledge states available for declaration by an examinee at any given item. These states are intended to capture a subjective epistemic classification that, if reported truthfully, carries diagnostic information about the probability of a correct response. To give this structure empirical and theoretical content, we define for each true knowledge state $s \in \Omega$ the conditional probability of a correct response, denoted p_s.

For an item with k alternatives, the theoretical floor on the probability of a correct response under pure random guessing is $\frac{1}{k}$. We define the three conditional correctness probabilities as satisfying the strict ordering.

$$p_{FK} > p_{PK} > p_{NK}, \quad (1)$$

With the following interpretations. Under FK, the examinee possesses sufficient information to identify the correct answer with near certainty, so p_{FK} is assumed to be close to 1, formally $p_{FK} \in (p^*, 1]$ for some threshold p^* near unity. Under NK, the examinee has no basis for preferring any one option over the others, so $p_{NK} = \frac{1}{k}$, reflecting uniform random selection. Under PK, the examinee possesses some relevant information, perhaps sufficient to eliminate one or more distractors but insufficient to uniquely identify the correct response with certainty, so $p_{PK} \in \left(\frac{1}{k}, p_{FK}\right)$.

This parameterization must be understood as characterizing the average examinee behavior associated with each state. Individual variation around these means is expected, and the empirical investigation of this variation is a primary goal of the present research program.

3.2 | The Scoring Mechanism

Let each item have k possible response options. For each item, the examinee declares a state $\hat{s} \in \Omega$ and selects an answer. Let the binary outcome variable $C \in \{0,1\}$ denote whether the selected answer is correct ($C = 1$) or incorrect ($C = 0$). The scoring function $\sigma(\hat{s}, C)$ assigns a real-valued score to each combination of declarations and correctness outcomes. The complete scoring table is defined as follows.

For FK declarations:

$$\sigma(FK, 1) = a, \sigma(FK, 0) = -b, \quad (2)$$

where $a > 0$ and $b > 0$ with b sufficiently large to penalize incorrect responses under FK. For PK declarations:

$$\sigma(\mathrm{PK}, 1) = c, \sigma(\mathrm{PK}, 0) = -d, \quad \textbf{(3)}$$

where

$$c > 0, d \geq 0 \text{ and } c < a,$$

for NK declarations:

$$(\mathrm{NK}, 1) = e, \sigma(\mathrm{NK}, 0) = f, \quad \textbf{(4)}$$

where $e \geq 0$ and $f \geq 0$, with both values small relative to a and c. The expected score under the declared state $\hat{s}$ when the true knowledge state is s (so the correctness probability is p_s) is:

$$\mathbb{E}[\,\sigma(\hat{s}, C) \mid s\,] = p_s \cdot \sigma(\hat{s}, 1) + (1 - p_s) \cdot \sigma(\hat{s}, 0). \quad \textbf{(5)}$$

3.3 | Incentive Compatibility Conditions

The scoring mechanism is incentive-compatible if and only if truthful reporting is an expected-value-maximizing strategy for each type of examinee. That is, for each true state $s \in \Omega$ and each alternative declaration $\hat{s} \neq s$, we require the following:

$$\mathbb{E}[\,\sigma(s, C) \mid s\,] \geq \mathbb{E}[\,\sigma(\hat{s}, C) \mid s\,]. \quad \textbf{(6)}$$

This argument yields a system of inequalities that constrains the scoring parameters. For an FK examinee ($p_s = p_{FK}$), truthful FK declaration must dominate both PK and NK declarations:

$$p_{FK} \cdot a - (1 - p_{FK}) \cdot b \geq p_{FK} \cdot c - (1 - p_{FK}) \cdot d, \quad \textbf{(7)}$$

$$p_{FK} \cdot a - (1 - p_{FK}) \cdot b \geq p_{FK} \cdot e + (1 - p_{FK}) \cdot f. \quad \textbf{(8)}$$

For a PK examinee ($p_s = p_{PK}$), truthful PK declaration must dominate both FK and NK:

$$p_{PK} \cdot c - (1 - p_{PK}) \cdot d \geq p_{PK} \cdot a - (1 - p_{PK}) \cdot b, \quad \textbf{(9)}$$

$$p_{PK} \cdot c - (1 - p_{PK}) \cdot d \geq p_{PK} \cdot e + (1 - p_{PK}) \cdot f. \quad \textbf{(10)}$$

For an NK examinee $\left(p_s = p_{NK} = \frac{1}{k}\right)$, truthful NK declaration must dominate both FK and PK:

$$\frac{1}{k} \cdot e + \frac{k-1}{k} \cdot f \geq \frac{1}{k} \cdot a - \frac{k-1}{k} \cdot b, \quad \textbf{(11)}$$

$$\frac{1}{k} \cdot e + \frac{k-1}{k} \cdot f \geq \frac{1}{k} \cdot c - \frac{k-1}{k} \cdot d. \quad \textbf{(12)}$$

Together, these six inequalities define a feasible polytope in the parameter space (a, b, c, d, e, f). Scoring vectors lying in this polytope guarantee that a rational, risk-neutral examinee has no expected-value incentive to misreport. The design task explored in detail in Article 4 of this series is to select from within this polytope in a manner that maximizes additional desiderata, such as score variance, discriminatory power, and robustness to risk preferences. The present article takes a specific admissible parameter vector as given and investigates behavioral compliance with the incentive structure.

3.4 | Connection to Proper Scoring Rules

The scoring mechanism developed above can be situated within the theoretical literature on proper scoring rules. A scoring rule σ defined over a finite outcome space is called proper if it incentivizes an agent to report their true subjective probability distribution over outcomes rather than any strategically distorted distribution.

The classic examples are the Brier (quadratic) and logarithmic scoring rules, both defined for continuous probability reports. The present mechanism constitutes a discretized, coarsened form of a proper scoring rule in which the examinee's probability report is restricted to three representative probability levels corresponding to FK, PK, and NK rather than the full unit interval. The incentive compatibility conditions derived above are analogous, in this discrete setting, to the properness condition in the continuous case. This framing places the present framework within established decision-theoretic foundations while acknowledging that the discretization introduces approximation relative to full probability elicitation.

4 | Research Questions and Hypotheses

4.1 | Research Questions

The empirical investigation is organized around four primary research questions, each targeting a distinct aspect of behavioral validity.

The first research question concerns the discriminative validity of knowledge state declarations. When examinees declare a given state, do the observed correctness rates differ significantly across the three declared states and in the theoretically predicted direction? This question probes the most basic diagnostic claim of the framework: that the declared state carries genuine information about the probability of a correct response.

The second research question concerns the calibration of knowledge state declarations. Do the observed correctness probabilities conditional on each declared state align quantitatively with the theoretical values p_{FK}, p_{PK}, and p_{NK} specified in the scoring design? This question goes beyond qualitative ordering to ask whether the declarations are quantitatively accurate and, thereby, whether the scoring parameters are empirically grounded.

The third research question concerns the prevalence and direction of systematic misreporting. Is there evidence of systematic overconfidence, specifically a tendency to declare FK on items where the realized correctness rate is substantially below p_{FK}? Conversely, is there evidence of underconfidence or risk-aversion-driven conservatism in the declaration of NK on items where PK appears to be present?

The fourth research question concerns individual differences in reporting accuracy. Do metacognitive monitoring skills, as measured by validated instruments, moderate the correspondence between declared knowledge states and realized correctness? That is, are examinees with higher metacognitive ability more reliably calibrated?

4.2 | Formal Hypotheses

The hypotheses are presented in both null and alternative forms, organized by research question.

Hypothesis Set 1. Discriminative Validity

Let μ_s denote the population mean correctness rate conditional on the declaration of state $s \in \{FK, PK, NK\}$. The null Hypothesis is:

$$H_{0,1}: \mu_{FK} = \mu_{PK} = \mu_{NK}. \quad (13)$$

It asserts that there is no systematic relationship between the declared state and the probability of correctness. The alternative Hypothesis is:

$$H_{A,1}: \mu_{FK} > \mu_{PK} > \mu_{NK}. \quad (14)$$

Asserting a strictly ordered relationship consistent with the theoretical specification.

Hypothesis Set 2. Calibration Accuracy

Let p_s^* denote the theoretically specified correctness probability under state s as embedded in the scoring design, and let $\hat{\mu}_s$ denote the sample estimate of μ_s. The null Hypothesis of exact calibration is the following:

$$H_{0,2}: \mu_s = p_s^* \text{ for all } s \in \Omega. \quad \textbf{(15)}$$

The alternative Hypothesis is that systematic miscalibration is present in at least one state:

$$H_{A,2}: \exists s \in \Omega \text{ such that } \mu_s \neq p_s^*. \quad \textbf{(16)}$$

Hypothesis Set 3. Direction of Systematic Bias

Let $\Delta_s = \hat{\mu}_s - p_s^*$ denote the calibration residual for state s. The overconfidence hypothesis asserts that FK declarations are systematically associated with correctness rates below the theoretical specification, i.e.,

$$H_{A,3a}: \Delta_{FK} < 0. \quad \textbf{(17)}$$

The conservatism or underconfidence hypothesis, pertaining primarily to NK declarations, asserts the following:

$$H_{A,3b}: \Delta_{NK} > \frac{1}{k}. \quad \textbf{(18)}$$

This indicates that examinees who declare NK are performing above chance, suggesting that their PK is being mischaracterized as ignorance.

Hypothesis Set 4. Metacognitive Moderation

Let M_i denote the metacognitive monitoring score of examinee i, measured by a validated instrument. Let ρ_i denote the individual calibration accuracy, quantified as the correlation between their binary correctness outcomes and an indicator variable for FK declaration across items. The moderation hypothesis is

$$H_{A,4}: \mathrm{Cor}(M_i, \rho_i) > 0. \quad \textbf{(19)}$$

Asserting a positive relationship between metacognitive ability and reporting accuracy.

5 | Experimental Design

The experimental design is a single-group pretest observation study in which all participants complete a knowledge-state-declared MCQ examination under the proposed incentive-compatible scoring mechanism. A single-group design is appropriate for the primary goal of validating the behavioral meaning of declared knowledge states; no comparison to a control group is necessary for *Hypothesis Sets 1* through 4, though a two-group extension would be valuable for examining behavioral differences relative to standard scoring. The experiment is designed with high internal validity as its primary priority, and the ecological validity of the examination setting is preserved through the use of genuine academic content and appropriately high stakes.

Ethical approval will be sought from the relevant institutional review board before data collection. All participants will provide written informed consent. The incentive structure will be fully disclosed to participants before the examination begins, and the scoring mechanism will be explained in a standardized briefing session. Examinees will be given a practice exercise of sufficient length to ensure familiarity with the declaration procedure before the actual examination.

5.1 | Participants

Participants will be undergraduate students enrolled in an introductory statistics or research methods course at a university where the experiment is approved and administered. The choice of population is motivated by the requirement that participants possess a meaningful but variable range of domain knowledge across the examination items, enabling the full range of knowledge states to be observed across individuals and items.

A target sample of $N = 300$ participants is proposed. This sample size is determined by a prospective power analysis for the primary inferential procedures described in Section 7. It ensures adequate statistical power to detect medium-effect-sized deviations from the null hypotheses at the α=0.05 significance level, with $1 - \beta \geq 0.85$. Participants will be excluded from analysis if they fail a comprehension check confirming understanding of the scoring mechanism, if they exhibit zero variance in their knowledge state declarations across all items (indicating either total ignorance or a failure to engage with the declaration task), or if their total response time falls below a minimum plausible threshold indicating random responding. These exclusions will be preregistered before data collection.

5.2 | Examination Instrument

The examination instrument will consist of $n = 60$ multiple-choice items drawn from the target course's content domain, each with $k = 4$ response alternatives. Items will be selected from a pretested bank based on a target range for item difficulty of $0.30 \leq \bar{p} \leq 0.80$ under standard scoring, where $\bar{p}$ denotes the proportion correct in a pilot administration. This range ensures representation of items spanning the difficulty spectrum, which is necessary to elicit a meaningful distribution of knowledge state declarations across the three categories. Items will additionally be selected to ensure discrimination parameters above 0.20 in a calibration using item response theory, confirming that the items provide reliable differentiation of examinee ability levels.

The 60 items will be randomized for each participant to preclude order effects. The examination will be administered on a secure digital platform that records response choice, declared knowledge state, and item-level response time for each participant-item combination. Response time is collected as an ancillary measure that may serve as a behavioral proxy for cognitive certainty: items on which an examinee deliberates for longer may, on average, be associated with lower subjective confidence, a conjecture that can be analyzed as a secondary validation of the declaration behavior.

5.3 | Scoring Parameters

Based on the incentive-compatibility analysis in Section 3, a specific parameter vector will be selected before the experiment and preregistered. For a four-alternative item $(k = 4)$, one admissible parameter vector that satisfies all six incentive compatibility inequalities under the theoretical values $p_{FK} = 0.95$, $p_{PK} = 0.60$, $p_{NK} = 0.25$ is the following:

$$a = 4, b = 2, c = 2, d = 1, e = 1, f = 0. \quad (20)$$

The expected scores under truthful reporting are then:

$$\mathbb{E}[\sigma(FK) \mid FK] = 0.95 \times 4 - 0.05 \times 2 = 3.70, \quad (21)$$

$$\mathbb{E}[\sigma(PK) \mid PK] = 0.60 \times 2 - 0.40 \times 1 = 0.80, \quad (22)$$

$$\mathbb{E}[\sigma(NK) \mid NK] = 0.25 \times 1 + 0.75 \times 0 = 0.25. \quad (23)$$

The verification that no declaration deviation from truth is beneficial under these parameters for each type of examinee can be confirmed by substituting into the six incentive compatibility inequalities. The published version of this article will provide a comprehensive report of these calculations.

Participants will be informed that their score on this examination contributes a defined proportion to their course grade, thereby ensuring the incentive structure is real rather than hypothetical. This argument is an important design consideration: the behavioral decision research literature has consistently found that hypothetical incentives produce systematically different responses from real stakes, particularly in tasks involving loss-bearing outcomes.

5.4 | Supplementary Instruments

In addition to the primary examination, participants will complete two supplementary instruments administered immediately afterward. The first is a validated measure of metacognitive monitoring ability, specifically the Metacognitive Awareness Inventory (MAI), which yields a score measuring the examinee's self-reported ability to monitor and regulate their cognitive processes. This instrument is included to enable the moderation analysis specified in *Hypothesis Set 4*. The second is a brief measure of risk preferences, administered via a standard lottery-choice task adapted for the classroom setting, yielding an individual-level risk aversion coefficient for each participant. This measure is included to control for confounding between risk preference and truthful reporting in the primary analysis.

6 | Variable Definitions and Measurement

6.1 | Primary Outcome Variables

The primary outcomes of interest are defined at the examinee-item pair level, yielding a dataset with $N \times n$ observations, where N denotes the number of participants and n the number of items. For participant i on item j, the following variables are recorded.

The declared knowledge state is a categorical variable $S_{ij} \in \{FK, PK, NK\}$ representing the state declared by participant i for item j. The correctness indicator is a binary variable $C_{ij} \in \{0,1\}$, where $C_{ij} = 1$ if the selected response is correct and $C_{ij} = 0$ otherwise. The item score is the real-valued variable $\sigma_{ij} = \sigma(S_{ij}, C_{ij})$, computed from the scoring function using the preregistered parameter vector.

6.2 | Derived Calibration Variables

From the primary variables, the following derived quantities are computed for each participant i.

The state-specific correctness rate is defined as:

$$\hat{\mu}_{is} = \frac{\sum_{j:S_{ij}=s} C_{ij}}{\#\{j: S_{ij} = s\}}, \tag{24}$$

for each $s \in \Omega$, where the numerator sums correct responses among all items declared under state s by participant i, and the denominator counts the number of items so declared. This variable is defined only when the denominator is positive, so participants with zero declarations in a given state will be excluded from state-specific analyses for that state.

The individual calibration error for state s is defined as:

$$\delta_{is} = \hat{\mu}_{is} - p_s^*. \tag{25}$$

where p_s^* is the theoretically specified correctness probability. Positive values of δ_{is} indicate underconfidence for state s, and negative values indicate overconfidence relative to the scoring design.

The overall individual calibration accuracy [5], denoted ρ_i, is computed as the point-biserial correlation between the correctness indicator C_{ij} and a recoded ordinal variable representing the declared state (FK coded as 2, PK as 1, and NK as 0) across all items for participant i:

$$\rho_i = \mathrm{Cor}\left(\{C_{ij}\}_{j=1}^{n}, \{D_{ij}\}_{j=1}^{n}\right), \tag{26}$$

where D_{ij} is the ordinal recode of S_{ij}. Higher values of ρ_i indicate that the participant's declarations are more strongly predictive of their actual performance, reflecting more accurate metacognitive monitoring.

6.3 | Item-Level Variables

At the item level, the following variables are computed from the pooled cross-examinee data for each item j.

The item difficulty under state s is $\bar{p}_{js} = \frac{1}{N_s}\sum_{i:S_{ij}=s} C_{ij}$, where N_s is the number of participants who declared state s on item j.

The state distribution for item j is the empirical proportion of participants declaring each state: $\hat{\pi}_{js} = \frac{N_s}{N}$ for $s \in \Omega$.

7 | Statistical Models and Inferential Framework

7.1 | Testing Hypothesis Set 1: Discriminative Validity

The discriminative validity of declared knowledge states is assessed through a mixed-effects logistic regression model in which the correctness indicator C_{ij} is regressed on the declared state S_{ij}, with participant i and item j included as crossed random effects to account for the multilevel structure of the data (multiple observations per participant and per item) [9], [10]. The model is specified as:

$$\log\frac{P(C_{ij}=1)}{P(C_{ij}=0)} = \beta_0 + \beta_1 \cdot 1[S_{ij}=FK] + \beta_2 \cdot 1[S_{ij}=PK] + u_i + v_j, \quad (27)$$

where NK is the reference category, $u_i \sim \mathcal{N}(0,\sigma_u^2)$ are participant random effects, and $v_j \sim \mathcal{N}(0,\sigma_v^2)$ are item random effects. The Hypothesis $H_{0,1}$ corresponds to $\beta_1 = \beta_2 = 0$, and $H_{A,1}$ corresponds to $\beta_1 > \beta_2 > 0$. The model is estimated by maximum likelihood using the Laplace approximation, and the fixed effects are tested by a likelihood ratio test comparing the full model to a null model with only random effects. Directional tests of the ordered inequality $\beta_1 > \beta_2 > 0$ are conducted using one-sided Wald tests with a false discovery rate of 0.05.

7.2 | Testing Hypothesis Set 2: Calibration Accuracy

For each knowledge state $s \in \Omega$, the calibration accuracy hypothesis is assessed by testing whether the population mean correctness rate μ_s equals the theoretically specified value p_s^*. Since the data structure is hierarchical, a one-sample test applied to the raw observations would violate independence assumptions [9]. Instead, a two-stage procedure is used. In the first stage, the participant-level mean correctness rate. $\hat{\mu}_{is}$ is computed for each participant-state combination. In the second stage, a one-sample t-test is applied to the participant-level means:

$$t_s = \frac{\overline{\hat{\mu}}_s - p_s^*}{\frac{s_s}{\sqrt{N_s'}}}, \quad (28)$$

where $\bar{\hat{\mu}}_s$ is the grand mean across participants, s_s is the standard deviation of the participant-level means, and N_s' is the number of participants with at least one declaration in state s. Bonferroni correction is applied across the three simultaneous tests to maintain a familywise error rate of 0.05.

7.3 | Testing Hypothesis Set 3: Systematic Bias Direction

The signed calibration residuals quantify the direction and magnitude of calibration bias δ_{is}. For each state s, the signed residuals δ_{is} are analyzed in a one-sample test of whether the population mean residual $\mathbb{E}[\delta_{is}]$ differs from zero, using the same two-stage approach as in Section 7. The alternative Hypothesis $H_{A,3a}$ (overconfidence in FK) corresponds to $\mathbb{E}[\delta_{i,FK}] < 0$, tested via a one-sided t-test. The alternative Hypothesis $H_{A,3b}$ (underreporting of PK as NK) corresponds to $\mathbb{E}[\delta_{i,NK}] > 0$, also tested via a one-sided t-test. To characterize the joint pattern of calibration bias across all three states simultaneously, a

multivariate calibration profile is constructed for each participant as the vector $\delta_i = (\delta_{i,FK}, \delta_{i,PK}, \delta_{i,NK})^\top$. Cluster analysis is then applied to the calibration profile matrix to identify empirically derived subgroups of examinees with qualitatively distinct misreporting patterns: systematic overconfidence, systematic underconfidence, and approximate truthfulness. This analysis serves an exploratory function and provides hypotheses for subsequent confirmatory research.

7.4 | Testing the Hypothesis Set 4: Metacognitive Moderation

The moderation of reporting accuracy by metacognitive monitoring ability is assessed through a hierarchical regression model [11]. In the first step, individual calibration accuracy ρ_i is regressed on the risk aversion coefficient r_i to control for the confounding of risk preference:

$$\rho_i = \gamma_0 + \gamma_1 r_i + \varepsilon_i \quad \textbf{(29)}$$

In the second step, the metacognitive monitoring score M_i is added as a predictor:

$$\rho_i = \gamma_0 + \gamma_1 r_i + \gamma_2 M_i + \varepsilon_i \quad \textbf{(30)}$$

The Hypothesis $H_{A,4}$ corresponds to $\gamma_2 > 0$, a tested factor of a one-sided t-test on the increment in R^2 associated with the addition of M_i. The hierarchical structure of the regression ensures that the association between metacognitive ability and calibration accuracy is estimated net of the confounding factor of risk preference.

7.5 | Supplementary Analysis: Response Time as Behavioral Validity Evidence

As a supplementary analysis, the response timeT_{ij} for participant i on item j is modeled as a function of the declared knowledge state, using a log-normal mixed-effects regression:

$$\log T_{ij} = \alpha_0 + \alpha_1 \cdot 1[S_{ij} = FK] + \alpha_2 \cdot 1[S_{ij} = PK] + u_i + v_j + \varepsilon_{ij} \quad \textbf{(31)}$$

The theoretical prediction is $\alpha_1 < 0$ and $\alpha_2 < 0$, indicating that FK declarations are associated with shorter response times than PK and NK declarations. This prediction follows from the assumption that genuine FK allows fast, fluent retrieval, while Partial or NK requires more deliberation. If this pattern is observed in the data, it provides convergent behavioral evidence that the declared states have genuine epistemic meaning instead of being assigned arbitrarily.

8 | Expected Results and Theoretical Implications

The theoretical model predicts a set of specific empirical outcomes that, together, would constitute a coherent validation of the knowledge state framework. The strongest and most fundamental prediction is that a declared knowledge state will be a significant positive predictor of correctness probability, with FK declarations associated with substantially higher correctness rates than PK declarations, which in turn exceed NK declarations. If this ordering is not observed, the foundational diagnostic claim of the framework is falsified, and no further utility of the declaration system as a signal of knowledge quality can be claimed.

Beyond the ordinal ordering, the quantitative calibration of the declarations carries substantial implications for the design of the scoring system. If the empirically observed correctness rates match well the theoretically specified values $p^*_{FK} = 0.95$, $p^*_{PK} = 0.60$, and $p^*_{NK} = 0.25$, then the incentive-compatibility analysis based on these parameter values is empirically well-founded. However, any systematic departures from these values would require recalibrating the scoring parameters to reflect the actual behavioral mapping from the reported state to the probability of correctness.

This dataset is also expected to show the pattern of overconfidence in FK declarations, as extensively documented in the broader calibration literature. The theoretically intriguing question is whether the incentive-compatible scoring structure, with its substantial penalty for incorrect responses under FK, partially

attenuates the overconfidence bias relative to environments without such penalties. A comparison of calibration statistics across conditions with and without the penalty structure would shed light on this question, though such a comparison is deferred to a subsequent experimental study. Regarding the moderating effect of metacognitive ability, the individual differences literature on metacognitive monitoring is sufficiently well established to warrant the prediction of a positive association between the MAI score and calibration accuracy. Still, the effect size for this relationship might be small, as the MAI assesses self-reported metacognitive tendencies rather than metacognitive performance directly. Thus, a moderate-to-small effect size would seem to be the best theoretically informed expectation, and the statistical design is sufficiently powered to detect effects of this magnitude.

9 | Limitations and Future Directions

Several limitations constrain the generalizability of the proposed study. The use of a single academic domain, introductory statistics, limits the external validity of the findings to populations with varying familiarity with quantitative material. It is plausible that calibration behavior differs substantially across domains with differing familiarity structures, such as historical knowledge versus mathematical reasoning, and such domain dependency is an important direction for future investigation. The sample restriction to undergraduates at a single institution similarly limits generalizability across demographics and ability levels. A second limitation concerns the confound between the novelty of the declaration task and the behavioral measures. Participants in any initial exposure to knowledge state reporting may exhibit more conservative or erratic declarations than they would after familiarization. A longitudinal design in which the same participants complete multiple examinations with the same scoring mechanism over a semester would enable the investigation of learning effects in declaration behavior, an important question for the practical deployment of the framework. Third, the calibration analysis assumes that the theoretically specified correctness probabilities p_s^* are appropriate population parameters against which to evaluate empirical calibration. In practice, these values may need to be estimated from a pilot study rather than specified a priori, and the statistical uncertainty in such estimation propagates into the calibration tests. A fully Bayesian treatment of this uncertainty, in which the p_s^* values are given prior distributions and updated from the data, which would provide a more principled framework for calibration assessment. Future research should extend the present design to a two-group randomized experiment in which one group receives the incentive-compatible scoring mechanism and a control group receives standard right-wrong scoring, enabling a direct comparison of declaration quality and score validity across conditions. This comparison is conceptually central to the research program but requires a more complex experimental infrastructure than is feasible within a single initial validation study.

10 | Conclusion

This article develops a comprehensive empirical framework for validating the behavioral validity of knowledge-state declarations under an incentive-compatible scoring mechanism for multiple-choice examinations. The theoretical contributions include a formal probabilistic model relating declared states to correctness probabilities, a rigorous derivation of the incentive compatibility conditions that guarantee truthful reporting as the expected-value-optimal strategy, and a connection of this framework to the theory of proper scoring rules. The empirical contributions include a preregistered experimental design, a complete set of operationally defined hypotheses, multilevel statistical models appropriate for the nested data structure, and a supplementary behavioral validity assessment using response times. Together, these elements constitute a methodologically rigorous foundation for a program of research on knowledge-state-based assessment that is grounded simultaneously in decision theory, psychometrics, and cognitive psychology. The results of this study, whether confirming or partially disconfirming the behavioral validity of the framework, will provide the empirical inputs necessary for subsequent articles in this series that address guessing reduction, metacognitive calibration, scoring parameter optimization, and fairness and reliability analysis.

Authors' Contributions

The authors conducted the research and prepared the manuscript, and have approved its final version.

Data Availability

The data are available from the corresponding authors upon reasonable request.

Funding

This work was carried out without financial support from any public, commercial, or non-profit organizations.

Conflict of Interest

There are no competing interests to declare.

References

[1] Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of educational measurement*, 7–11. https://doi.org/10.1111/j.1745-3984.1975.tb01003.x

[2] Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive--developmental inquiry. *American psychologist*, *34*(10), 906. https://psycnet.apa.org/doi/10.1037/0003-066X.34.10.906

[3] Gardner-Medwin, A. R., & Gahan, M. (2003). Formative and summative confidence-based assessment. https://tmedwin.net/~ucgbarg/tea/caa03a.pdf

[4] Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and psychological measurement*, *55*(6), 959–966. https://doi.org/10.1177/0013164495055006004

[5] Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

[6] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121. https://doi.org/10.1037//0022-3514.77.6.1121

[7] Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: part i* (pp. 99–127). World Scientific. https://doi.org/10.1142/9789814417358_0006

[8] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the american statistical association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

[9] Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing. *Learning, and instruction: National research council. national academy press*. http://www.nap.edu/catalog/10019.html

[10] Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied measurement in education*, *6*(2), 103–118. https://doi.org/10.1207/s15324818ame0602_1

[11] Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary educational psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033