Hipotenusa : Journal of Mathematical Society
Volume 8 Issue 1Year 2026
ISSN: 2716-3156
Website: https://hipotenusa.iainsalatiga.ac.id/index.php/hipotenusa/index

Are We Measuring Ability or Guessing?: CTT and IRT Evidence from a Multiple-Choice Assessment in Econometric Test

Ajeng Wahyuni [1]*, Yunaita Rahmawati[2], Maulida Nurhidayati[3], Muhtadin Amri[3]
[1,2,3,4] Fakultas Ekonomi dan Bisnis Islam, Universitas Islam Negeri Kiai Ageng Muhammad Besari Ponorogo, Indonesia.
*Corresponding Author. E-mail: ajeng@uinponorogo.ac.id
DOI: 10.18326/hipotenusa.v8i1.7088

Article submitted: June 04, 2026
Article reviewed: June 28, 2026
Article published: June 30, 2026

Abstract

Multiple-choice tests are widely used in mathematics-related higher education courses because they are practical for assessing broad learning outcomes. Correct responses may not always indicate full conceptual mastery, as students may answer correctly through partial knowledge, distractor elimination, unintended item cues, or pseudo-guessing. This study evaluates the quality of a 20-item four-option multiple-choice econometrics assessment using Classical Test Theory (CTT) and Item Response Theory (IRT). The test was administered to 108 undergraduate students and was designed to measure econometrics competence as an applied mathematics construct, including quantitative and statistical reasoning, regression and model interpretation, hypothesis testing and inference, model assumptions and diagnostics, and data-based decision-making. CTT was used to examine item difficulty, item discrimination, while IRT was used to compare the 1PL, 2PL, and 3PL models and to diagnose pseudo-guessing. The results showed a mean score of 13.85 out of 20, KR-20 of 0.681, and Cronbach's alpha of 0.677, indicating moderate but not strong internal consistency. CTT identified no difficult items, nine easy items, and three items with poor discrimination. The 1PL model had the lowest BIC and was therefore the most fit model, while the 3PL model was retained diagnostically because it estimates pseudo-guessing. Eight items, namely I01, I05, I06, I12, I15, I16, I18, and I20, had pseudo-guessing parameters above 0.25. These findings suggest that some correct responses may have been influenced by non-mastery factors. This study contributes to mathematics education by demonstrating how integrated CTT and IRT diagnostics can improve the validity of econometrics assessment as a measure of quantitative and statistical reasoning.

Keywords: classical test theory, item response theory, pseudo-guessing, item discrimination, econometrics assessment

INTRODUCTION

Measurement accuracy in educational assessment is increasingly critical in ensuring the validity of student competences (Harris, 2023), without exception in mathematics education because test scores are often used to infer students' quantitative reasoning, statistical literacy, and ability to apply mathematical models in problem-solving contexts. In higher education, econometrics represents an applied mathematical domain in which students are expected to connect statistical concepts, regression models, inference, and empirical interpretation. Therefore, inaccurate measurement may lead to misleading conclusions about students' actual competence (Newton, 2005) in quantitative reasoning rather than merely producing technical scoring errors. This issue is especially relevant when assessment results are used for instructional feedback, course evaluation, and academic decision-making within economics and business education. Thus, improving measurement validity in econometrics assessment directly contributes to applied mathematics education by strengthening the quality of evidence used to evaluate applied mathematical learning.

In this study, econometrics competence is defined as students' ability to use quantitative and statistical reasoning to interpret regression models, understand hypothesis testing, evaluate model assumptions, and draw evidence-based conclusions from empirical information (Serbenyuk, 2021). This definition is important because econometric tests should not only measure factual recall but also assess students' mathematical reasoning, statistical interpretation, and model-based thinking. The intended in the assessment are quantitative reasoning, statistical reasoning, and mathematical modeling (Ajilore, 2006). By defining these domains explicitly, this study positions econometrics as an applied context of mathematics education.

Multiple-choice assessments are commonly used in quantitative courses because they are efficient, practical, and suitable for evaluating a broad range of learning outcomes (Fuhrman, 1996). However, in subjects such as econometrics, correct answers in multiple-choice items may not always indicate full conceptual understanding or genuine quantitative reasoning (Haladyna, 2022). Students may answer correctly through partial knowledge, elimination strategies, test-wise behavior, or random guessing, particularly when distractors are weak or item cues are easily recognized . This creates a validity concern because observed scores may overestimate students' actual econometrics competence. Consequently, econometrics assessments require item-level diagnostic evidence to determine whether correct responses are more likely to reflect ability or non-mastery factors.

Classical Test Theory (CTT) provides an important first step in evaluating the quality of classroom and institutional assessment. CTT offers practical indices such as item difficulty, item discrimination, and internal consistency reliability, which help identify whether a test functions appropriately at the observed score level (Andrich & Marais, 2019). These indices are useful because they are interpretable and can guide basic item revision in classroom and institutional assessment contexts . However, CTT is limited because it depends on observed scores and does not explicitly model latent ability or the probability that low-ability students will answer items correctly by chance or

through non-mastery strategies (Frey, 2020). Therefore, although CTT can indicate whether an item is easy or difficult, it cannot fully explain whether correct responses represent genuine econometric reasoning or possible guessing-related success.

Item Response Theory (IRT) offers a more detailed diagnostic framework because it models the relationship between students' latent ability and the probability of answering each item correctly (Schmidt & Embretson, 2012). In multiple-choice testing, the three-parameter logistic model is particularly relevant because it estimates item discrimination, item difficulty, and the lower asymptote or pseudo-guessing parameter. The guessing parameter does not prove that students intentionally guessed, but it can indicate that an item allows correct responses among students with low estimated ability (Brown, 2016) (Hambleton et al., 1992). In a four-option multiple-choice test, a pseudo-guessing estimate substantially above 0.25 may suggest weak distractors, item cues, or other non-mastery factors that increase the probability of correct responses (Reise & Revicki, 2014) (Schmidt & Embretson, 2012). Thus, the 3PL model provides evidence that is especially valuable for diagnosing validity threats in econometrics assessment.

Although CTT and IRT have been widely applied in educational measurement (Xie & Liu, 2025), language assessment (Fergadiotis et al., 2023), and general psychometric studies, few studies have examined whether multiple-choice assessments in econometrics can distinguish genuine quantitative reasoning from pseudo-guessing. This gap is important because econometrics is not a purely factual subject; it requires students to integrate mathematical reasoning, statistical inference, model interpretation, and empirical decision-making. Previous applications of CTT and IRT have contributed to broader assessment practices in educational measurement and language testing; however, domain-specific evidence in undergraduate econometrics assessment remains limited (K. M. Schmidt & Embretson, 2012). Therefore, examining CTT and 3PL IRT evidence in econometric assessments is necessary to strengthen the validity of quantitative assessments in higher education mathematics-related learning.

The novelty of this study lies in the integration of CTT and 3PL IRT to diagnose threats to validity in a domain-specific multiple-choice assessment of economics competence. Rather than treating test evaluation only as a reliability issue, this study examined whether item performance reflects students' ability or is partly influenced by pseudo-guessing and weak item functioning. This approach allowed the study to identify items that may inflate student scores without accurately measuring the intended construct of quantitative and statistical reasoning. Therefore, this study contributes to mathematics education by positioning econometrics assessment as an applied measure of quantitative reasoning, statistical reasoning, and mathematical modeling, while also providing practical evidence for revising multiple-choice items that may weaken measurement validity.

METHODS

This study used a quantitative psychometric design to evaluate the quality of a multiple-choice econometrics assessment. The analysis was diagnostic because item-level CTT and IRT evidence were used to explain whether correct responses were more consistent with students' latent econometrics ability or with potential non-mastery

response processes. The unit of analysis was the test item, and the person-level response patterns provided the empirical basis for estimating the item parameters.

The dataset contained dichotomously scored responses from 108 undergraduate students who completed a 20-item four-option econometrics test. The test was intended to assess students' ability to understand regression-related concepts, interpret estimated coefficients, evaluate statistical significance, recognize model assumptions, and draw conclusions from empirical results. These competencies are relevant to mathematics education because they require students to apply mathematical and statistical reasoning to real data and formal models. The 20 items covered ECM, ARCH/GARCH, and VAR/VECM.

Table 1. Construct Domains of the Econometrics Test

Construct domain	Operational indicator
Quantitative and statistical reasoning	Recognizing numerical patterns, probability-based reasoning, and basic statistical logic in econometric contexts
Regression and model interpretation	Interpreting regression coefficients, signs, magnitudes, and model outputs
Hypothesis testing and inference	Understanding significance tests, p-values, and decisions about estimated parameters
Model assumptions and diagnostics	Identifying assumptions, possible violations, and implications for interpretation
Data interpretation and decision making	Drawing conclusions from empirical information and connecting results to substantive problems

CTT analysis estimated item difficulty as the proportion of students who answered correctly and item discrimination (Wu, 2012). The upper-lower index was computed from the upper and lower 27% performance groups, with 28 respondents in each group (Scheuneman & Steinhaus, 1987). Corrected item-total correlations below 0.20 were treated as weak and requiring review, following common item analysis practice (Jeter et al., 2024).

IRT analysis was conducted in R 4.5.2 using the mirt package (Susanto et al., 2025). 1PL, 2PL, and 3PL models were estimated and compared using log-likelihood, AIC, BIC, sample-size-adjusted BIC, and model fit indices (Reise & Revicki, 2014). The 3PL model was used specifically to estimate item discrimination (a), difficulty (b), and the lower asymptote or pseudo-guessing parameter (c) (Hambleton et al., 1992).

$$P(X_i = 1 \mid \theta) = c_i + (1 - c_i) / [1 + \exp\{-a_i(\theta - b_i)\}]$$

In this model, a_i indicates how sharply the item differentiates students around the difficulty location, b_i indicates the ability level at which the item is more likely to be answered correctly, and c_i indicates the lower asymptote or pseudo-guessing probability. Because the test used four response options, $c_i > 0.25$ was interpreted as elevated pseudo-guessing. However, the 3PL estimates were interpreted diagnostically rather than as definitive population estimates because the sample size was relatively small for the unconstrained 3PL calibration.

RESULTS AND DISCUSSION

Descriptive Statistics and Reliability

The dataset contained 108 student response entries. The econometrics test consisted of 20 four-option multiple-choice items designed to assess students' quantitative reasoning, statistical reasoning, and interpretation of econometric models. The mean score was 13.850 out of 20 (SD = 3.364), with a median score of 14.000. Scores ranged from 4 to 20, and the mean proportion correct was 0.693, indicating that the test was, on average, moderately to relatively easy for the respondents.

Table 1. Descriptive Statistics and Reliability of the Test

Indicator	Result
Initial rows in the dataset	108
Valid respondents analyzed	108
Number of items	20
Mean score	13.850
Median score	14
Standard deviation	3.364
Minimum score	4
Maximum score	20
Mean proportion correct	0.693
KR-20	0.681
Cronbach's alpha	0.677
Standardized alpha	0.689
SEM based on alpha	1.910

Cronbach's alpha was 0.677, and KR-20 was 0.681. These values indicate moderate internal consistency and not strong reliability. Therefore, the test can be considered able to capture part of students' econometrics-related quantitative reasoning, but the score interpretation should remain cautious. The reliability evidence suggests that the instrument still requires item-level revision before it can be treated as a highly stable measure of econometric competence.

Classical Test Theory Analysis

Item Difficulty

Item difficulty was represented by the proportion of students who answered each question correctly. The categorization used in this study was difficult ($p < 0.30$), moderate ($0.30 <= p <= 0.70$), and easy ($p > 0.70$). Based on this criterion, no item was classified as being difficult. Eleven items were rated as moderate (I01, I03, I04, I10, I13, I14, I15, I16, I17, I18, and I20), while nine items were rated as easy (I02, I05, I06, I07, I08, I09, I11, I12, and I19). Item I09 was particularly easy ($p = 0.953$), indicating that it may provide limited information for distinguishing students' quantitative reasoning abilities.

Item Discrimination

Two discrimination indicators were considered: the upper-lower discrimination index and the corrected item-total correlation. For the upper-lower index, the criteria were poor ($D < 0.20$), moderate ($0.20 <= D < 0.30$), good ($0.30 <= D < 0.40$), and very good ($D >= 0.40$).

The item discrimination index suggested that many items could separate high- and low-performing students. Items with poor item discrimination are I07, I09, and I11, while others had good and very good item discrimination values.

The item discrimination index is important because the test is intended to measure students' integrated ability in quantitative reasoning, statistical reasoning, and interpretation of econometric models. Items with weak item discrimination may measure a different subskill, contain ambiguous wording, provide unintended clues, or function through non-mastery response processes rather than through the intended academic reasoning.

Table 2. Classical Test Theory Item-Level Results

Item	Item Difficulty	Diff.	Item Discrimination	Interpretaion
I01	0.505	Mod.	0.429	Very Good
I02	0.785	Easy	0.464	Very Good
I03	0.514	Mod.	0.393	Very Good
I04	0.514	Mod.	0.679	Very good
I05	0.766	Easy	0.464	Very good
I06	0.785	Easy	0.500	Very Good
I07	0.822	Easy	0.143	Poor
I08	0.729	Easy	0.286	Moderate
I09	0.953	Easy	0.179	Poor
I10	0.645	Mod.	0.607	Very Good
I11	0.813	Easy	0.107	Poor
I12	0.701	Easy	0.536	Very Good
I13	0.645	Mod.	0.429	Very Good
I14	0.570	Mod.	0.321	Good
I15	0.617	Mod.	0.571	Very Good
I16	0.692	Mod.	0.321	Good
I17	0.645	Mod.	0.464	Very Good
I18	0.664	Mod.	0.429	Very Good
I19	0.804	Easy	0.464	Very Good
I20	0.682	Mod.	0.571	Very Good

Item Response Theory Analysis

Unidimensionality and Local Independence

The dimensionality check indicated that the first eigenvalue was 4.735 and the second eigenvalue was 2.317, producing a lambda1/lambda2 ratio of 2.044. The first factor explained 0.237 of common variance. Because the ratio was below the practical benchmark of 3.000, the unidimensionality evidence was not strong and required further examination. The one-factor fit also suggested caution (RMSEA = 0.517, TLI = -0.060, and RMSR = 0.102). Therefore, all IRT results in this study are interpreted as diagnostic evidence, not as definitive population-level calibration.

The local independence analysis also produced mixed results. Several Yen's Q3 item pairs exceeded the absolute diagnostic threshold of approximately 0.20, whereas the LD X2 output did not indicate substantial dependence. This suggests that some item pairs may share residual response patterns, possibly because they require similar procedures or

interpretations of econometric concepts. These findings further justify cautious diagnostic interpretations.

Model Fit and Model Selection

Three IRT models were compared: 1PL, 2PL, and 3PL, to determine which models fit the best. Based on BIC, the 1PL model was the most parsimonious model (BIC = 2504.801), while the 3PL model had the largest BIC (2630.737) because it estimated more parameters. However, the 3PL model showed stronger absolute fit indices, including M2 p = 0.239, RMSEA = 0.027, TLI = 0.938, and CFI = 0.951. The 3PL model was retained for diagnostic interpretation because the focus of this study was pseudo-guessing, and the 3PL model is the only model that estimates the lower-asymptote parameter c. Nevertheless, because the sample size was small for stable 3PL estimation and the EM cycles reached the maximum iteration warning, the 3PL estimates should not be interpreted as final population estimates.

Table 3. IRT Model Fit and Model Selection

Model	LogLik	AIC	BIC	SABIC	M2 p	RMSEA	SRMSR	TLI	CFI
1PL	-1203.336	2448.671	2504.801	2438.451	0.031	0.044	0.110	0.843	0.844
2PL	-1185.357	2450.713	2557.626	2431.245	0.067	0.040	0.091	0.869	0.883
3PL	-1175.184	2470.367	2630.737	2441.165	0.239	0.027	0.091	0.938	0.951

Item Characteristic Curve and Information Function

The item characteristic curves and test information function from R Analysis output are displayed below. The ICC plot shows that several items had high lower asymptotes under the 3PL model. In substantive terms, this does not prove actual guessing behavior. Rather, it suggests elevated pseudo-guessing estimates that may reflect non-mastery factors such as partial knowledge, weak distractors, clueing, or distractor elimination.

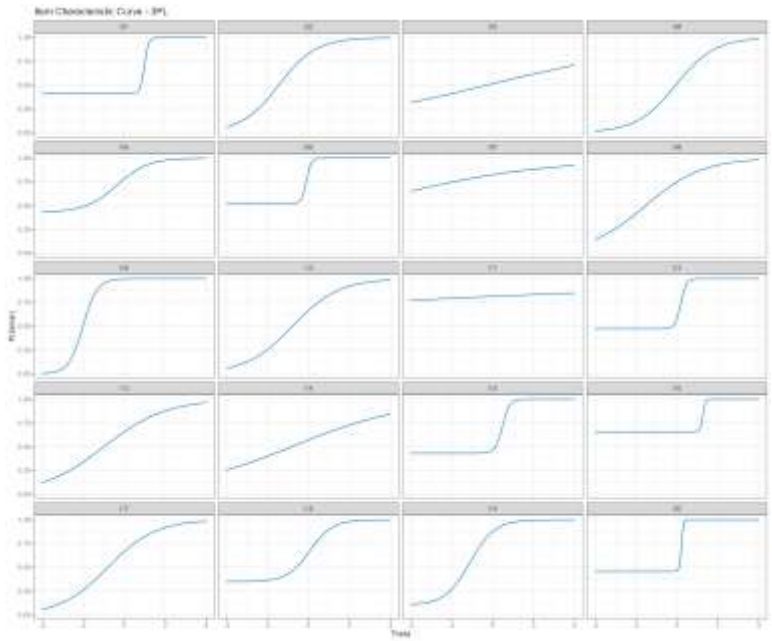

Figure 1. Item Characteristic Curves under the 3PL model

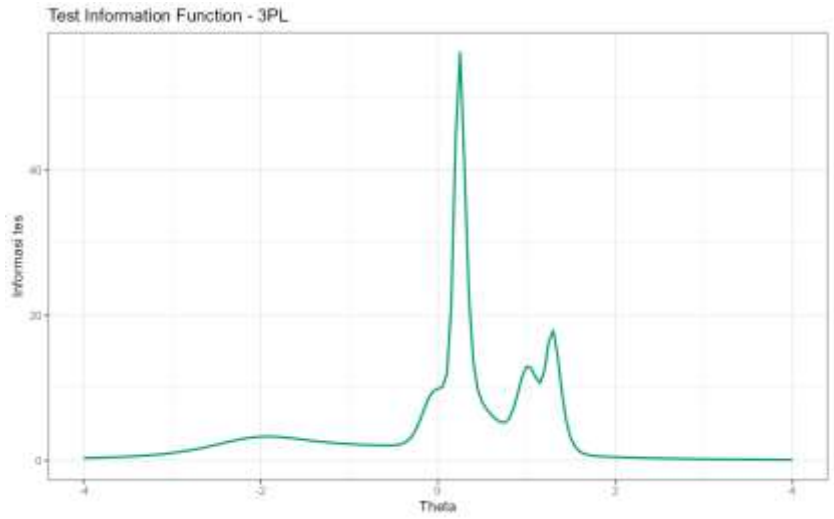

Figure 2. Test Information Function under the 3PL model

The 3PL test information function is concentrated within a relatively narrow ability range and contains sharp peaks. This pattern should not be interpreted as evidence that the test is highly precise across all ability levels. Instead, it may reflect the influence of extreme item discrimination estimates and the limited stability of the 3PL estimation in a small sample.

Item Fit Analysis

Most 3PL item-fit statistics were acceptable. However, I13 ($p = 0.028$), I18 ($p = 0.005$), and I19 ($p = 0.043$) showed item fit concerns at the 0.05 level. Item I09 produced a numerical warning in the item-fit output, which is likely related to its very high proportion of correct answers. These items should be reviewed in relation to content, distractor functioning, and the possibility that the item is too easy or not aligned with the intended econometric reasoning construct.

Item Parameter Estimation (a, b, c)

The 3PL parameter estimates identified eight items with pseudo-guessing parameters above the theoretical chance level of 0.25 for four-option items: I01, I05, I06, I12, I15, I16, I18, and I20. These items do not automatically prove that students are guessing. The more appropriate interpretation is that some correct responses may have been influenced by other factors, such as pseudo-guessing, partial knowledge, weak distractors, or distractor elimination.

Several item discrimination estimates were extremely high, including I01 ($a = 9.991$), I06 ($a = 8.810$), I12 ($a = 7.231$), I16 ($a = 15.958$), and I20 ($a = 21.658$). These values should not be mechanically interpreted as evidence that the items are excellent discriminators. In the context of a small sample and a 3PL model with three estimated item parameters, such extreme values more plausibly indicate estimation instability, quasi-separation or item-specific response patterns. Thus, these items should be inspected qualitatively rather than accepted solely because of their high IRT discrimination values.

Table 4. 3PL Item Parameter Estimates and Item-Fit Diagnostics

Item	a	b	c	c interpretation
I01	9.991	0.981	0.409	High Guessing
I02	1.089	-1.449	0.000	Low Guessing
I03	0.209	-0.226	0.005	Low Guessing
I04	1.033	-0.057	0.000	Low Guessing
I05	1.304	-0.362	0.429	High Guessing
I06	8.810	-0.104	0.520	High Guessing
I07	0.234	-6.566	0.011	Low Guessing
I08	0.719	-1.520	0.001	Low Guessing
I09	2.880	-2.017	0.001	Low Guessing
I10	0.878	-0.782	0.001	Low Guessing
I11	0.065	-21.259	0.075	Low Guessing
I12	7.231	0.189	0.472	High Guessing
I13	0.654	-0.990	0.002	Low Guessing
I14	0.353	-0.699	0.023	Low Guessing
I15	5.757	0.476	0.432	High Guessing
I16	15.958	1.264	0.658	High Guessing
I17	0.861	-0.794	0.001	Low Guessing
I18	1.783	0.071	0.349	High Guessing
I19	1.649	-1.110	0.102	Low Guessing
I20	21.658	0.221	0.455	High Guessing

Comparison between CTT and IRT Results

The CTT and IRT results indicate a mixed measurement condition. Some items performed reasonably well under CTT and did not show elevated pseudo-guessing under 3PL. For example, I02, I03, I04, I10, I13, I17, and I19 had very good item discrimination indices and low pseudo-guessing estimates, although I19 still showed a 3PL item-fit concern. Other items required different types of results. I01, I05, I06, I12, I15, I18, and I20 had elevated pseudo-guessing estimates but had very good item discrimination. These items should not be automatically removed; instead, their distractors and wording should be reviewed. In contrast, I07 and I11 had poor item discrimination but low or moderate pseudo-guessing estimates; therefore, they should be described as low-discrimination items, not as high pseudo-guessing items.

Pseudo-Guessing and Item Discrimination Analysis

A correlation analysis was conducted to examine the association between the 3PL pseudo-guessing parameter and the item discrimination indicators. The results showed that the association between c and item discrimination was not significant. Pearson's r was 0.058 ($p = 0.807$), whereas Spearman's rho was -0.083 ($p = 0.729$). The association between c and the upper-lower discrimination index was also not significant (Pearson's r = 0.258, $p = 0.272$; Spearman's rho = 0.055, $p = 0.817$). Therefore, this study does not claim a strong substantive relationship between pseudo-guessing and CTT discrimination.

Table 6. Association between Pseudo-Guessing and Discrimination Indicators

Association	Pearson r	Pearson p	Spearman rho	Spearman p
c vs IRT a	0.762	<0.001	0.621	0.003
c vs corrected item-total r	0.058	0.807	-0.083	0.729
c vs upper-lower D	0.258	0.272	0.055	0.817

The positive correlation between c and the IRT discrimination parameter a was statistically significant (Pearson's r = 0.762, p < 0.001; Spearman's rho = 0.621, p = 0.003). Cautious interpretation is that both parameters may be affected by small-sample estimation sensitivity, extreme item slopes, and item-specific response patterns in the 3PL model.

Implication in Educational Assessment

These findings have direct implications for mathematics education and quantitative assessment. Econometrics items are not only technical economics questions; they involve quantitative reasoning, statistical reasoning, mathematical modeling, and interpreting regression results. If multiple-choice items allow students to obtain correct answers through weak distractors or clue-based elimination, the resulting scores may overestimate students' ability to reason mathematically and statistically.

For instructional practice, item analysis suggests several revision strategies. First, distractors should be made more homogeneous in terms of structure, length, and plausibility so that the correct answer is not visually or linguistically obvious. Second, unintended clues in the item stems or response options should be removed. Third, items should be aligned with the intended cognitive level: some items should assess basic recognition, while others should require the interpretation of coefficients, model assumptions, hypothesis testing, and regression output. Fourth, items with weak corrected item-total correlations should be reviewed to ensure that they measure the same quantitative reasoning construct as the overall test.

Overall, the integration of CTT and 3PL IRT helps lecturers move beyond total scores and identify whether an econometrics test captures students' quantitative reasoning or is partly affected by non-mastery response processes. This is especially important in mathematics education contexts, where assessment should support the valid interpretation of students' reasoning, not merely count correct answers.

CONCLUSION

This study investigated whether a multiple-choice econometrics assessment measured students' ability or was partly influenced by pseudo-guessing and other non-mastery response processes. The findings indicate that the test measured students' econometrics-related quantitative reasoning to a moderate extent, but its measurement quality was not sufficiently strong to support highly stable score interpretation. This conclusion is supported by the moderate reliability coefficients, with KR-20 = 0.681 and Cronbach's alpha = 0.677, as well as the presence of several items that showed acceptable discrimination under Classical Test Theory. However, the test also contained several items that were too easy, poor item discrimination, or high pseudo-guessing value.

The CTT results showed that no item was classified as difficult, while nine items were categorized as easy and 11 as moderate. Items I07, I09, and I11 showed poor discrimination, indicating that they were less effective in distinguishing students with higher and lower levels of econometrics-related reasoning abilities. These items require careful review because poor item discrimination may result from ambiguous wording,

overly obvious options, poor distractors, or misalignment with the intended constructs of quantitative reasoning, statistical reasoning, and model interpretation.

The IRT model comparison showed that the 1PL model was the most parsimonious model based on the BIC. Nevertheless, the 3PL model remained useful as a diagnostic tool because it provided information about the lower asymptote or pseudo-guessing parameter. The 3PL analysis identified eight items with pseudo-guessing parameters above 0.25: I01, I05, I06, I12, I15, I16, I18, and I20. These elevated values should not be interpreted as direct evidence that the students intentionally guessed. Rather, they indicate that some correct responses may have been influenced by partial knowledge, weak distractors, distractor elimination or unintended item clues.

This study contributes to mathematics education by positioning econometrics assessment as an applied measure of quantitative reasoning, statistical reasoning, mathematical modeling and regression interpretation. The integration of CTT and IRT allows lecturers and test developers to move beyond total scores and examine whether correct responses genuinely represent students' reasoning abilities. Practically, the findings suggest that item revision should focus on improving distractor quality, removing unintended cues, balancing item difficulty, strengthening item discrimination, and aligning each item with the intended econometric competence domain.

Several limitations for this study are that the sample size was relatively small for stable 3PL calibration, several IRT discrimination estimates were extreme, the EM estimation reached the maximum number of cycles, and the unidimensionality evidence was not strong. Therefore, the 3PL results should be interpreted as diagnostic evidence rather than definitive population-level estimates. Future research should recalibrate the instrument using a larger sample, conduct distractor analysis, examine students' response processes, compare IRT models through sensitivity analysis, and validate the revised items after improvement. Through these steps, the econometrics test can become a stronger and more valid instrument for assessing students' quantitative and statistical reasoning in mathematics-related higher education contexts.

References

Ajilore, O. (2006). Econometric Issues in Education Finance. *Review of Regional Studies*, *36*(2). https://doi.org/10.52324/001c.8317

Andrich, D., & Marais, I. (2019). *Classical Test Theory* (pp. 29–39). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-7496-8_3

Brown, G. (2016). Item Response Theory: Complicated but better. *Figshare*. https://doi.org/10.17608/k6.auckland.3827082.v4

Fergadiotis, G., Casilio, M., Dickey, M. W., Steel, S., Nicholson, H., Fleegle, M., Swiderski, A., & Hula, W. D. (2023). Item Response Theory Modeling of the Verb Naming Test. *Journal of Speech, Language, and Hearing Research*, *66*(5), 1718–1739. https://doi.org/10.1044/2023_jslhr-22-00458

Frey, F. (2020). Test Theory and Classical Test Theory. In *The International Encyclopedia of Media Psychology* (pp. 1–6). Wiley. https://doi.org/10.1002/9781119011071.iemp0047

Fuhrman, M. (1996). Developing Good Multiple-Choice Tests and Test Questions. *Journal of Geoscience Education*, *44*(4), 379–384. https://doi.org/10.5408/1089-9995-44.4.379

Haladyna, T. (2022). Creating multiple-choice items for testing student learning. *International Journal of Assessment Tools in Education*, *9*(Special Issue), 6–18. https://doi.org/10.21449/ijate.1196701

Hambleton, R., Swaminathan, H., & Rogers, H. (1992). Fundamentals of item response theory. *Choice Reviews Online*, *29*(07), 29–4185. https://doi.org/10.5860/choice.29-4185

Harris, D. J. (2023). *Theory and Principles of Educational Measurement* (pp. 27–45). Routledge. https://doi.org/10.4324/9781003444534-3

Jeter, R., Chamberlain, D., & Rozier, K. (2024). *An Integrated Methodology for Assessing Item Discrimination in Mathematics Assessments*. Center for Open Science. https://doi.org/10.31235/osf.io/xvh7y

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, *31*(4), 419–442. https://doi.org/10.1080/01411920500148648

Reise, S. P., & Revicki, D. A. (2014). *Handbook of Item Response Theory Modeling*. Routledge. https://doi.org/10.4324/9781315736013

Scheuneman, J. D., & Steinhaus, K. S. (1987). A Theoretical Framework For The Study Of Item Difficulty And Discrimination. *ETS Research Report Series*, *1987*(2), i–35. https://doi.org/10.1002/j.2330-8516.1987.tb00248.x

Schmidt, K. M., & Embretson, S. E. (2012). Item Response Theory and Measuring Abilities. In *Handbook of Psychology, Second Edition*. John Wiley Sons. https://doi.org/10.1002/9781118133880.hop202016

Schmidt, S., Zlatkin-Troitschanskaia, O., & Shavelson, R. J. (2023). Modeling and Measuring Domain-Specific Quantitative Reasoning in Higher Education Business and Economics. *Frontline Learning Research*, *11*(1), 40–56. https://doi.org/10.14786/flr.v11i1.885

Serbenyuk, S. (2021). On Some Aspects of the Examination in Econometrics. *Journal of Vasyl Stefanyk Precarpathian National University*, *8*(3), 7–16. https://doi.org/10.15330/jpnu.8.3.7-16

Susanto, H. P., Abadi, A. M., , H., Retnawati, H., Ali, R. M., & Djidu, H. (2025). Development of irtawsi: A User-Friendly R Package for IRT Analysis. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, *14*(1), 1–23. https://doi.org/10.15408/jp3i.v14i1.32091

Wu, M. (2012). *Using Item Response Theory as a Tool in Educational Measurement* (pp. 157–185). Springer Netherlands. https://doi.org/10.1007/978-94-007-4507-0_9

Xie, L., & Liu, X. (2025). Exploring the Role of Response Time in Item Response Theory: Rethinking the PISA 2022 Creative Thinking Assessment. *The Journal of Creative Behavior*, *59*(4). https://doi.org/10.1002/jocb.70072