Comparing Item Response Theory Subscoring Approaches: An Application to Senior High School Statistics and Probability Achievement Test

Kevin Carl P. Santos, Ph.D.

School of Statistics

University of the Philippines-Diliman

Armi S. Lantano

Center for Educational
Measurement, Inc.

14th NCS National Convention on Statistics
1-3 October 2019 | Crowne Plaza Manila Galleria
Organized by the Philippine Statistical System
Spearheaded by the Philippine Statistics Authority

Introduction

- In order to keep abreast with the curricular changes in the country, the Center for Educational Measurement, Inc. (CEM) – the pioneering testing and research institution in the Philippines, developed a series of achievement tests, the CEM K to 12 Achievement Tests
- These are standardized tests designed to measure knowledge and skills learned in school based on the national curriculum
- These include tests in English, Mathematics and Science from Kindergarten to Grades 11 and 12
- As of School Year 2017-2018, CEM has already released five achievement tests for Senior High School level including Statistics and Probability

Introduction

- The CEM K to 12 Achievement Test in Statistics and Probability, in particular, is composed of 60 multiple choice items partitioned in five content areas, namely:
  - (1) *Random Variables and Probability Distribution* (CA01)
  - (2) *Sampling and Sampling Distribution* (CA02),
  - (3) *Estimation of Parameters* (CA03),
  - (4) *Test of Hypothesis* (CA04) and
  - (5) *Correlation and Regression Analyses* (CA05)
- The reliability of the test is 0.83, whereas the concurrent validity of the test yields coefficients ranging from 0.38 to 0.65, indicating that the achievement test is reliable and valid.

Introduction

- The overall ability estimate on Statistics and Probability is useful for important decisions
- However, the domain ability estimates complement the overall ability estimating by providing finer grained diagnosis of examinees' strengths and weaknesses
- To make valid inferences about a student's attributes from the student's responses to items in the subtest domains, reliable subscores should be obtained
- Yet, because of the small number of items within the subtest sections, lack of sufficient reliability is the primary impediment for generating and reporting subtest scores

Objectives

- The primary objective of this research was to examine and compare the results of the four subscoring methods in the item response theory (IRT) context
  - Multidimensional Scoring (MS; de la Torre and Patz, 2005)
  - Augmented Scoring (AS; Wainer et al., 2001)
  - Objective Performance Index scoring (OPI; Yen, 1987)
  - Higher order – IRT Approach (HO-IRT; de la Torre and Song, 2009)
- Subsequently, the other goal of this study was to profile the SHS students who took the Statistics and Probability achievement test

Data

- Responses from a total of 2,536 Filipino SHS students coming from 11 private schools nationwide who took the CEM K to 12 Achievement Test in Statistics and Probability for Grades 11/12 in SY 2017-2018 were analyzed in this study

Location	***N***	**Percent (%)**
National Capital Region (NCR)	1,895	74.7
Luzon	235	9.3
Visayas	295	10.2
Mindanao	147	5.8
Total	2,536	100.0

Methodology

- Based on a simple structure assumption, the multidimensional model by Reckase (1996) reduces to the three-parameter logistic (3PL) model and is written in the following manner:

$$
\begin{aligned}
P_{j(d)}(\theta_{i(d)}) &= P(X_{ij(d)} = 1|\theta_{i(d)}, \alpha_{j(d)}, \beta_{j(d)}, \gamma_{j(d)}) \\
&= \gamma_{j(d)} + (1 - \gamma_{j(d)}) \frac{1}{1 + exp[-1.7\alpha_{j(d)}(\theta_{i(d)} - \beta_{j(d)})]},
\end{aligned}
$$

$P(X_{ij(d)} = 1|\theta_{i(d)}, \alpha_{j(d)}, \beta_{j(d)}, \gamma_{j(d)})$ is the probability of examinee i answering item j of dimension d correctly;

$\theta_{i(d)}$ is the d^{th} component of the ability vector $\boldsymbol{\theta}_i$;

$\alpha_{j(d)}, \beta_{j(d)}$, and $\gamma_{j(d)}$ are the discrimination, difficulty, and guessing parameters, respectively, of the j^{th} item of dimension d;

$i = 1, \ldots, I$ (the total number of examinees);

$j = 1, \ldots, J$ (the total number of items);

$d = 1, \ldots, D$ (the total number of dimensions);

$j(d) = 1(d), \ldots, J(d)$; and

$\sum_{d=1}^{D} J(d) = J$.

Methodology

Multidimensional Scoring

- The multidimensional approach (de la Torre & Patz, 2009) to simultaneously estimating abilities can be viewed as a more general framework for obtaining expected a posteriori (EAP) estimates of ability
- Improvement in the domain abilities can be observed when the abilities are correlated, particularly when there are multiple short tests and the underlying correlation is high
- Aside from generating better ability estimates, the hierarchical formulation allows the direct estimation of the correlation between the abilities

$$\Sigma \sim Inv - Wishart_{\nu_0}(\Lambda_0^{-1})$$

$$\theta_i|\Sigma \sim MVN(0, \Sigma).$$

Methodology

Augmented Scoring

- The procedure proposed by Wainer et al. (2001) uses the test reliabilities and intertest correlations in estimating the correlations among the abilities
- Their procedure relies on the test reliabilities and intertest correlations to estimate the correlations between the abilities
- These are used to compute the empirical Bayes ability estimates afterwards
- The method is a multivariate extension of Kelly’s (1927) regressed scores and can be used in conjunction with a variety of score types: conventional summed score, scale score, and IRT score

Methodology

Objective Performance Index Scoring

- Yen's (1987) scoring method, on the other hand, does not utilize the correlations between the abilities
- Instead, it employs the examinee's performance on the overall test to improve scores on the subsections of the test
- In particular, the overall ability estimate is first computed and used as "prior information" (i.e., based on Beta distribution) to improve the estimation of the true score (proportion correct) in a specific test objective
- This subsequently results in objective scores (called the objective performance indexes (OPI)) with smaller standard errors

Methodology

Higher-order IRT Approach

- The HO-IRT scoring method (de la Torre & Song, 2009) formulates a higher-order linear factor model that is used to relate the overall and domain abilities
- Although more constrained in some respects, the HO-IRT model is consistent with the hierarchical ability structure well accepted in psychological research and practice
- The higher-order (HO) scoring approach is based on a hierarchical Bayesian framework given by the following formulation

$$\theta_i \sim N(0,1)$$

$$\lambda_d \sim U(-1,1)$$

$$\theta_{i(d)}|\theta_i, \lambda_d \sim N(\lambda_d\theta_i, 1-\lambda_d^2),$$

where

θ is the overall ability of examinee i

λ_d is the coefficient in regression $\theta_{(i)d}$ on θ

Methodology

- The abilities and the corresponding proportion correct for each examinee on the four content areas were estimated using the Expected a Posteriori (EAP) and four subscoring methods
- In the absence of the true ability and proportion correct, the different methods were compared using the characteristics of the distribution of the ability estimates
- Specifically, summary statistics based on moments (mean and standard deviation) and quantiles (0.05, 0.50, and 0.95) were computed and compared

Results

Correlational Structure among the Content Domains

Domains	CA02	CA03	CA04	CA05
CA01	0.79	0.87	0.87	0.78
CA02		0.82	0.81	**0.73**
CA03			**0.90**	0.81
CA04				0.80

- The correlations among the content areas yielded coefficients ranging 0.73 to 0.90, with association between *Estimation of Parameters* (CA03) and *Test of Hypothesis* (CA04) as the strongest whereas association between *Sampling and Sampling Distribution* (CA02) and *Correlation and Regression Analyses* (CA05) as the weakest

Results

Summary Statistics for $\theta_{(d)}$ (Domain abilities)

Statistics		Method	CA01	CA02	CA03	CA04	CA05
Moment	*Mean*	MS	.14	.16	.18	.17	.16
		AS	.10	.18	**.25**	.11	**.20**
		OPI	.12	.15	**.12**	.13	**.11**
		HO	.13	.15	.17	.16	.15
	SD	MS	.73	.77	.79	.82	.72
		AS	.83	.81	.85	.89	.81
		OPI	**1.06**	**1.21**	**1.25**	**1.21**	**1.14**
		HO	.81	.75	.80	.80	.77
Quantile	*5th*	MS	-1.00	-1.05	-1.05	-1.11	-.96
		AS	-1.10	-.93	-.92	-1.09	-.91
		OPI	**-2.56**	**-3.00**	**-3.00**	**-3.00**	**-3.00**
		HO	-1.20	-1.02	-1.11	-1.12	-1.03
	50th	MS	.11	.12	.14	.13	.11
		AS	**.00**	**.06**	**.12**	**-.03**	**.08**
		OPI	**.27**	**.21**	**.34**	**.25**	**.24**
		HO	.10	.11	.13	.12	.10
	95th	MS	1.38	1.52	1.54	1.62	1.42
		AS	**1.61**	**1.69**	**1.86**	**1.81**	**1.71**
		OPI	**1.61**	**1.87**	**1.79**	**1.83**	**1.72**
		HO	1.48	1.45	1.55	1.57	1.49

Results

- For both measures of central tendency, HO and MS showed a more similar pattern compared to the AS and OPI estimates
- The median ability estimated under AS was the lowest (near zero) whereas the median ability estimated under OPI was the highest
- The variabilities of the OPI estimates had the largest SD, followed by the AS estimated, then by the HO estimates
- Upon examining the 5th percentiles, OPI ability estimates behaved very differently from the other subscoring methods (i.e., most extreme), whereas the ability estimates of the three methods were close to each other
- Across the four subscoring methods at the opposite end of the scale (i.e., 95th percentile), AS and OPI abilities estimates were higher than those of MS and HO

Results

Summary Statistics for $\pi_{(d)}$ (Expected Proportion Correct)

Statistics		Method	CA01	CA02	CA03	CA04	CA05
Moment	*Mean*	MS	.38	.40	.31	.35	.39
		AS	.37	.40	.32	.35	.40
		OPI	.39	.41	.33	.36	.42
		HO	.38	.39	.31	.35	.40
	SD	MS	.12	.09	.10	.10	.14
		AS	.13	.11	.13	.11	.16
		OPI	.14	.13	.12	.11	.17
		HO	.13	.09	.10	.09	.15
Quantile	*5th*	MS	.23	.29	.22	.27	.22
		AS	.23	.30	.23	.27	.23
		OPI	**.19**	**.24**	**.21**	**.24**	**.18**
		HO	.22	.29	.22	.27	.22
	50th	MS	.35	.37	.28	.32	.36
		AS	.33	.37	.28	.31	.35
		OPI	.38	.38	.30	.32	.39
		HO	.35	.37	.28	.31	.36
	95th	MS	.60	.59	.50	.56	.68
		AS	**.64**	**.63**	**.61**	**.61**	**.75**
		OPI	**.65**	**.67**	**.58**	**.62**	**.75**
		HO	.62	.58	.50	.55	.70

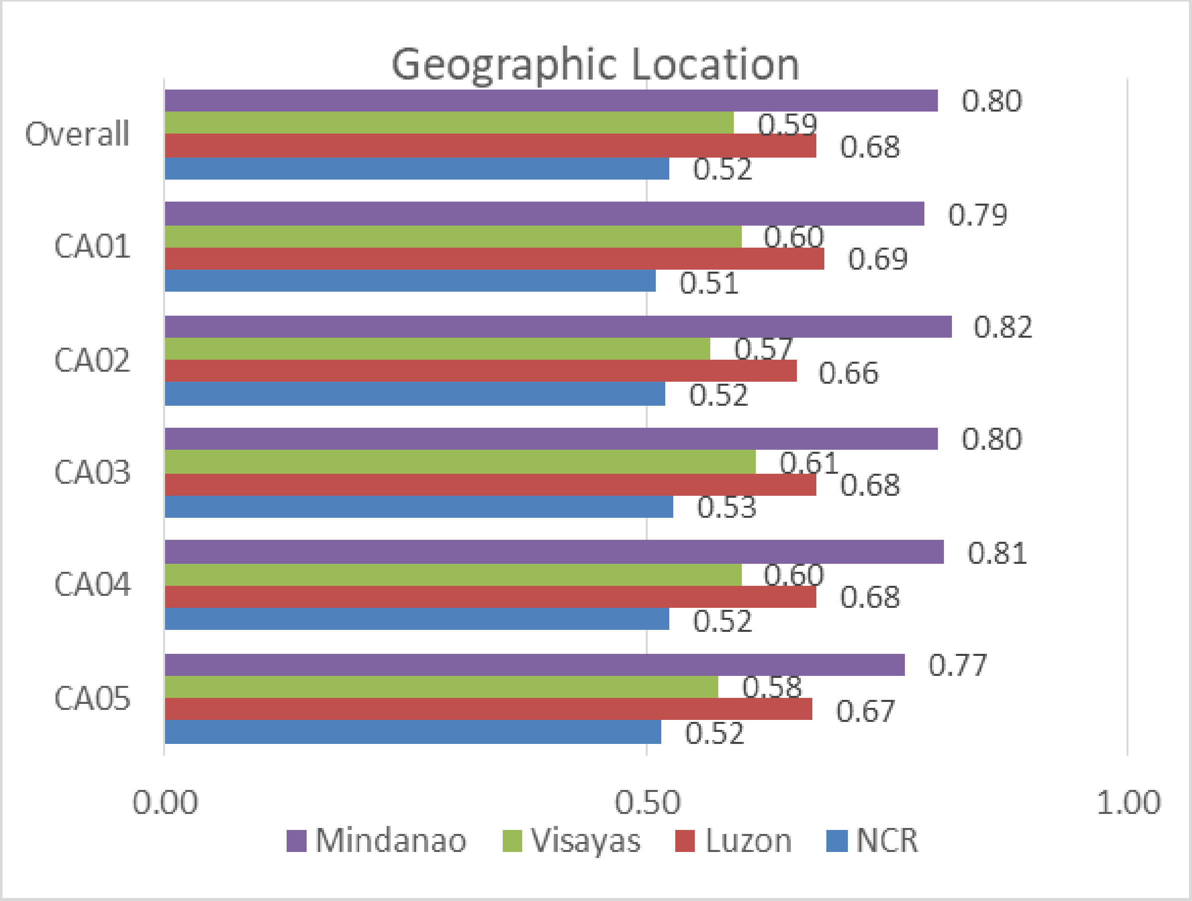
Results

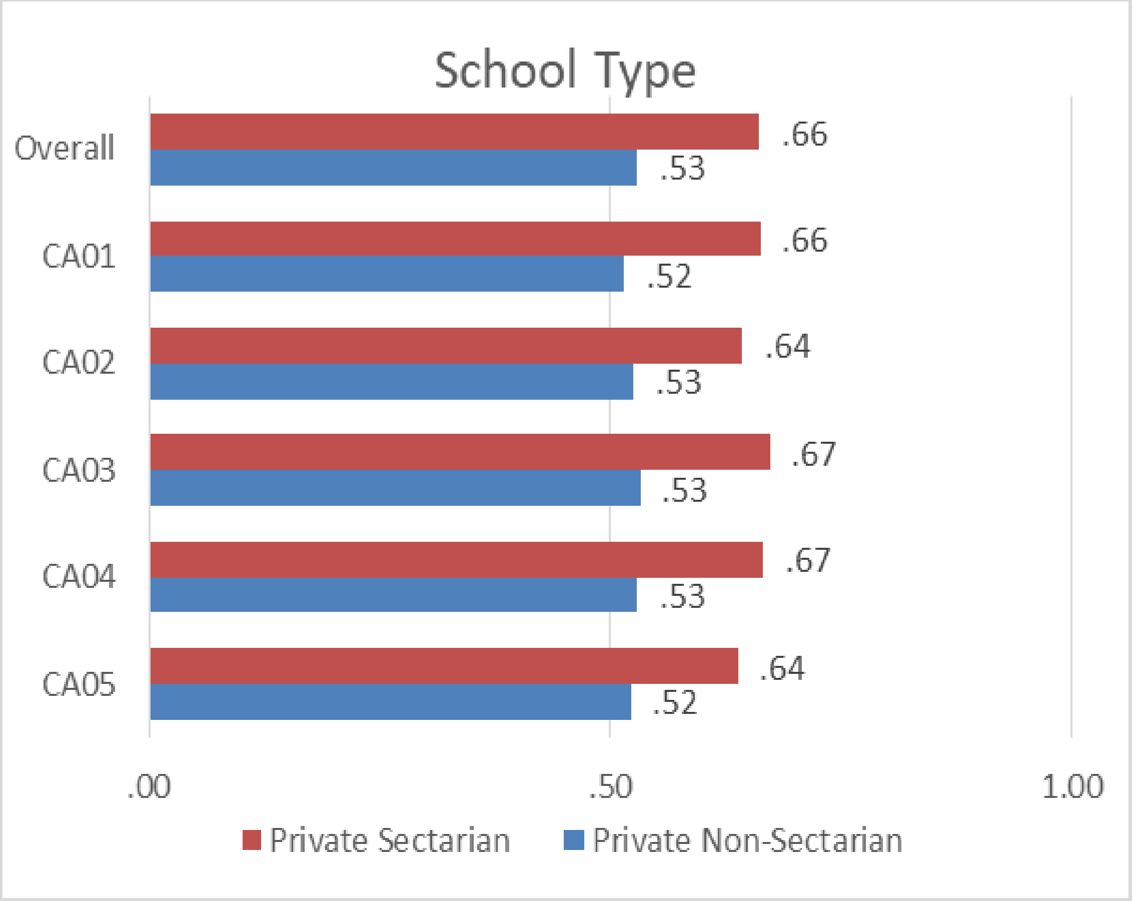
- The different methods showed fewer discrepancies when compared in terms of expected proportion correct
- Their mean and median estimates did not differ by more than 0.05 in absolute terms
- For the examinees at the 5th percentile, the estimated proportions correct under OPI was the lowest across the four subscoring methods
- At the 95th percentile, OPI and AS showed similar domain estimates that were higher than those under MS and HO

Discussion

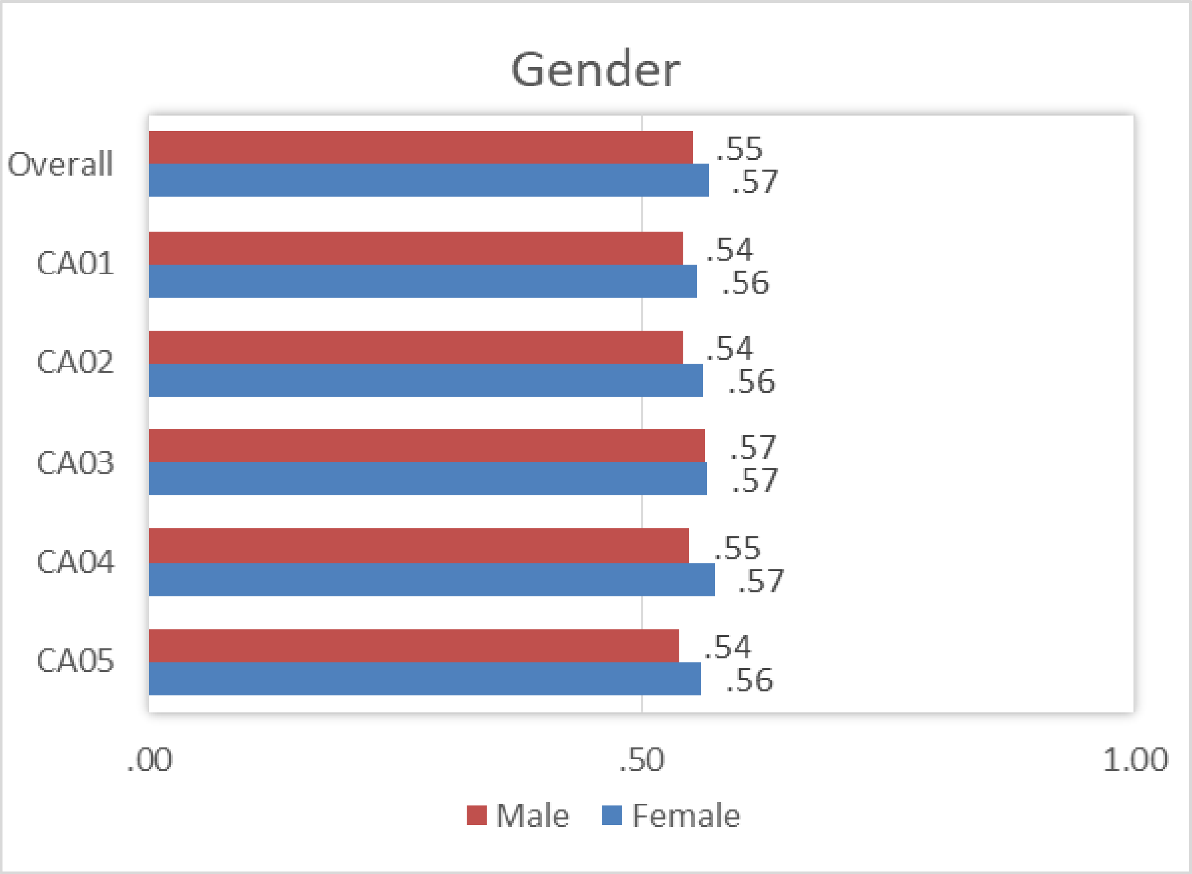
- The findings demonstrated that the ability estimates obtained using the four subscoring methods were not too different from each other in general
- However, it was revealed that they differed in terms of variability and estimation of the low-ability examinees
- Specifically, the OPI method demonstrated a greater tendency of yielding more extreme results in these respects
- Based on this empirical dataset, it was found that MS and HO produced highly comparable results
- Given that MS and HO performed similarly, HO should be the model of choice if a unified framework for obtaining the overall and domain ability estimates is of interest

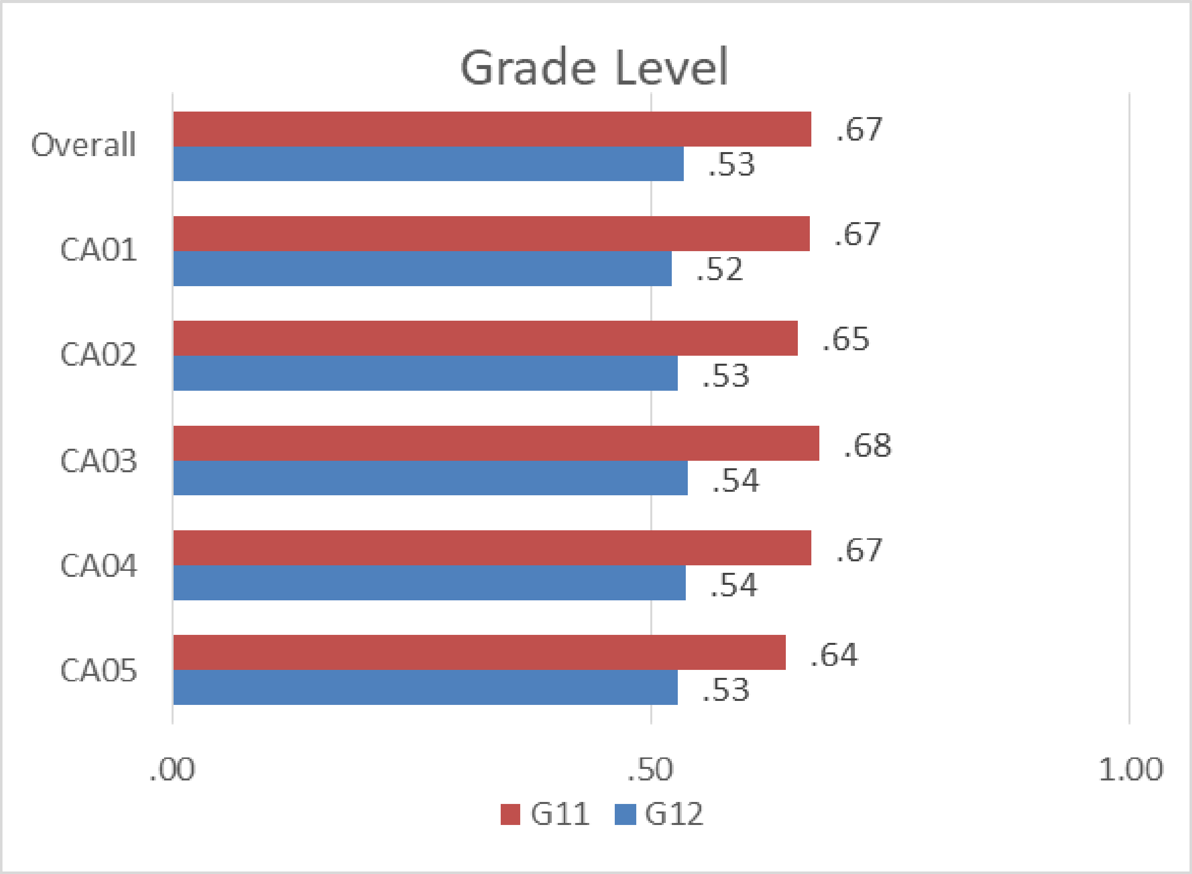
HO-IRT Approach Results





HO-IRT Approach Results





Fin.

Thank you very much!

kpsantos1@up.edu.ph