Psychometric Features of the Engineering Learning Outcomes Exam from Different Perspectives

Dr Muhammad Khalid Naveed

Dr Abdullah Al - Sadaawi

National Canter for Assessment in Higher Education, Saudi Arabia

Introduction

- Context of the study
- CTT and IRT analysis
- Dimensionality
- Factor structure
- Differential item functioning
- Measurement invariance

Context of the study

- NCA has developed a number of assessment tools for certification and admission
- Engineering Learning Outcomes Exam (ELOE) is chosen as an example
- ELOE Comprised of 105 MCQs items
- ELOE Comprised of 11 domains

Structural Aspect of Validity

- To ensure essential unidimensionalty
- (a) **Model A**: a one-factor model,
- (b) **Model B**: a eleven-factor model with the eleven content-specific domains as correlated latent factors,
- (c) **Model C**: a eleven-factor model with the eleven content-specific domains as uncorrelated latent factors, and
- (d) **Model D**: a bifactor model, with one general factor loading on all items and eleven latent factors as content-specific aspects of the general factor

Reliability

- Using latent variable modeling (LVM) approach taking into account the binary nature of the item scores
- Cronbach Alpha
- Marginal / True score

Factorial Invariance

- **Configural invariance:** Configural invariance is achieved if the model of interest fits across the groups. Although the model is the same across groups, the unknown parameters of the model are assumed to be different across the groups.

- **Weak invariance:** Weak invariance for a measurement model implies that the model is the same across the groups and that the factor loadings are identical across the groups. Weak convergence is then achieved if this multiple group model fits the data.

- **Strong invariance:** The measurement model is said to have strong invariance if the model is the same with identical factor loadings and means across studied groups of interest.

- **Complete invariance:** A model is completely invariant across groups if the model is the same across the groups and all the parameters are identical across the groups.

Differential Item Functioning

Groups are compared on item performance after adjusting for overall performance on the measured trait

Mantel-Haenszel Statistics

Type A items - negligible DIF: items with $|\Delta\alpha_{MH}| < 1$ or MH test is not statistically significant.

Type B items - moderate DIF: items with $1 \leq |\Delta\alpha_{MH}| \leq 1.5$, and MH test is statistically significant.

Type C items - large DIF: items with $|\Delta\alpha_{MH}| > 1.5$, and MH test is statistically significant.

Results

Dimensionality and Factor Structure of ELOE

Fit Indices: (a) *Comparative fit index*: CFI > 0.95 ;

(b) *Incremental Fit Index*: IFI > 0.95;

(c) *Standardized root mean square residual*: SRMR = 0.00 (SRMR < 1.00 for an adequate fit); and

(d) *Root mean square error of approximation*: RMSEA = 0.00 (RMSEA ≤ 0.05 for an adequate data fit (e.g., Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004).

Dimensionality and Factor Structure of ELOE

Testing for Data Fit of Four CFA Models of the ELOE Structure

CFA Model	χ^2	*df*	CFI	IFI	SRMR	RMSEA	90% CI for RMSEA *LL*	90% CI for RMSEA *UL*
Model A	9718.886	5355	. 944	. 944	. 027	. 019	. 019	.020
Model B	8217.983	5300	. 959	. 959	. 025	. 016	.015	.017
Model C	9876.114	5355	. 860	. 861	. 066	. 036	. 036	. 037
Model D	7064.192	5184	. 971	. 971	. 023	. 013	.012	.014

Note. Model A = one-factor; Model B = eleven correlated factors; Model C = eleven uncorrelated factors; Model D = bifactor model.

Correlation

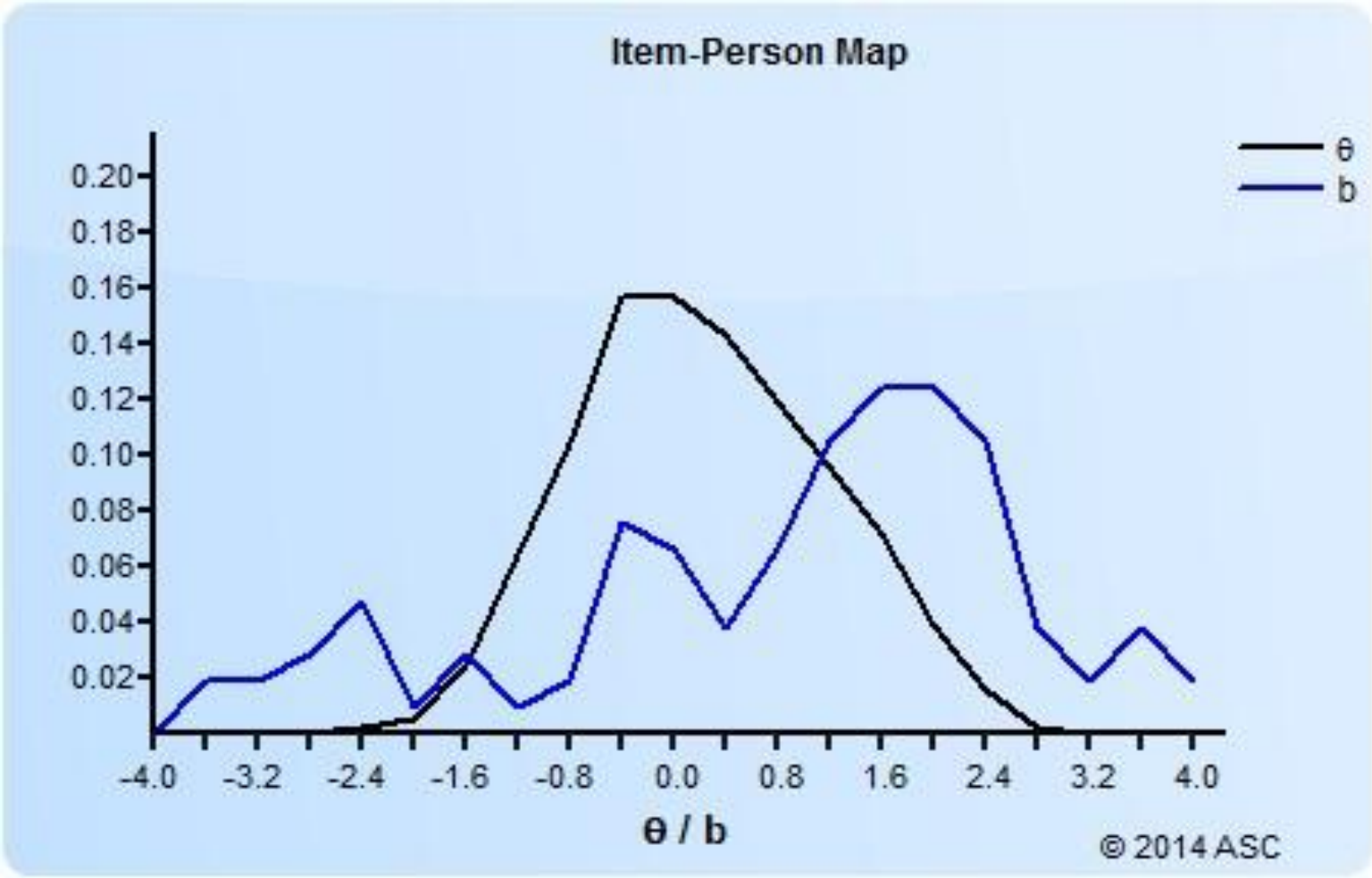
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
F1	1.000										
F2	0.813	1.000									
F3	0.635	0.581	1.000								
F4	0.901	0.590	0.583	1.000							
F5	0.998	0.815	0.825	0.831	1.000						
F6	0.994	0.663	0.572	0.955	0.795	1.000					
F7	0.672	0.641	0.787	0.509	0.764	0.506	1.000				
F8	0.644	0.614	0.764	0.506	0.725	0.618	0.877	1.000			
F9	0.708	0.655	0.788	0.543	0.798	0.648	0.897	0.938	1.000		
F10	0.760	0.561	0.861	0.565	0.651	0.560	0.881	0.782	0.982	1.000	
F11	0.755	0.639	0.846	0.574	0.807	0.581	0.887	0.927	0.995	0.945	1.000

IRT Analysis of ELOE

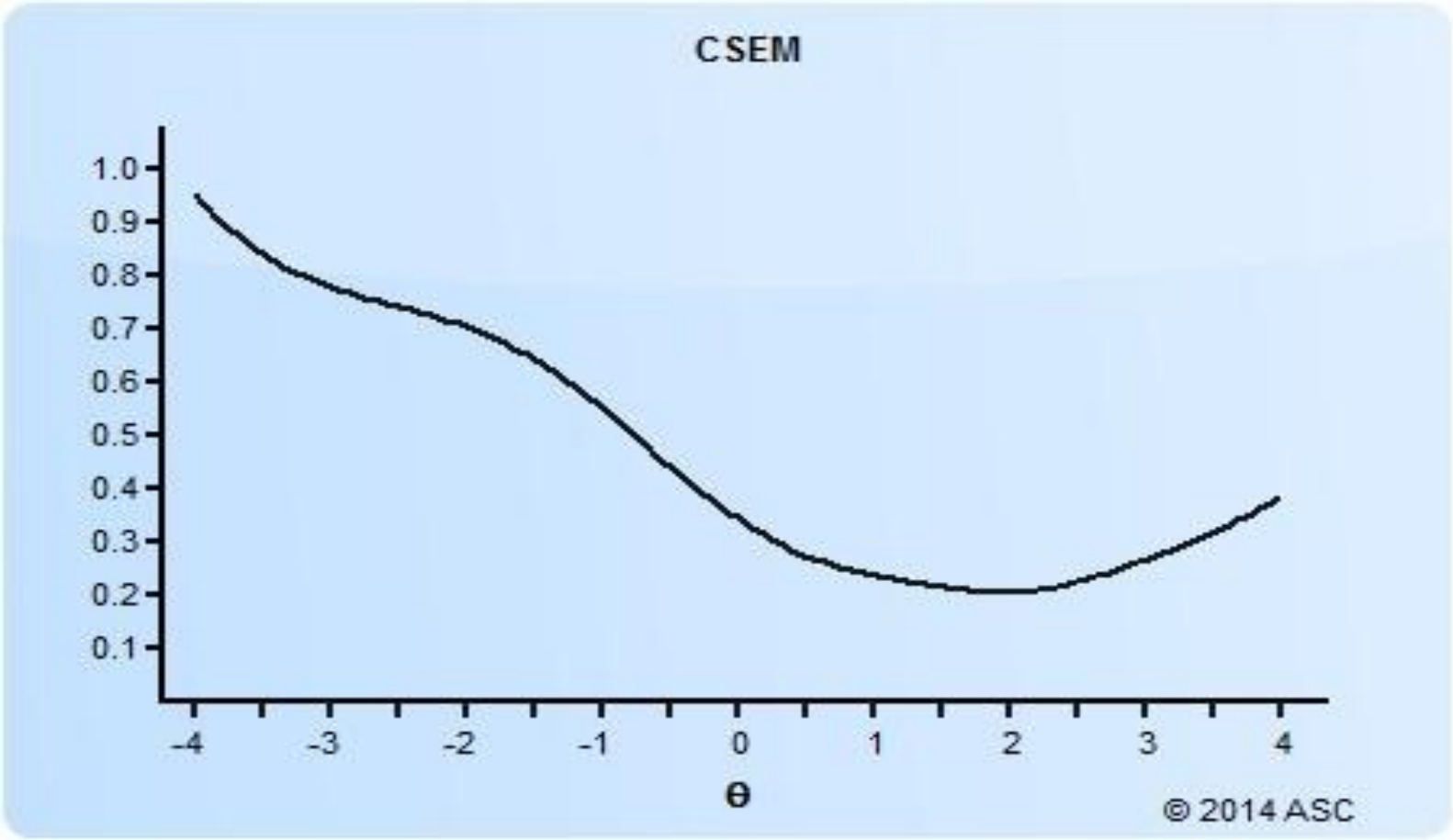
Range, Mean, and Standard Deviation of the Item Parameter Estimates Under the 3PL in IRT

Parameter	Items	Mean	SD	Min	Max
a	105	0.825	0.381	0.201	2.066
b	105	0.627	1.849	-3.671	4.000
c	105	0.240	0.041	0.128	0.386

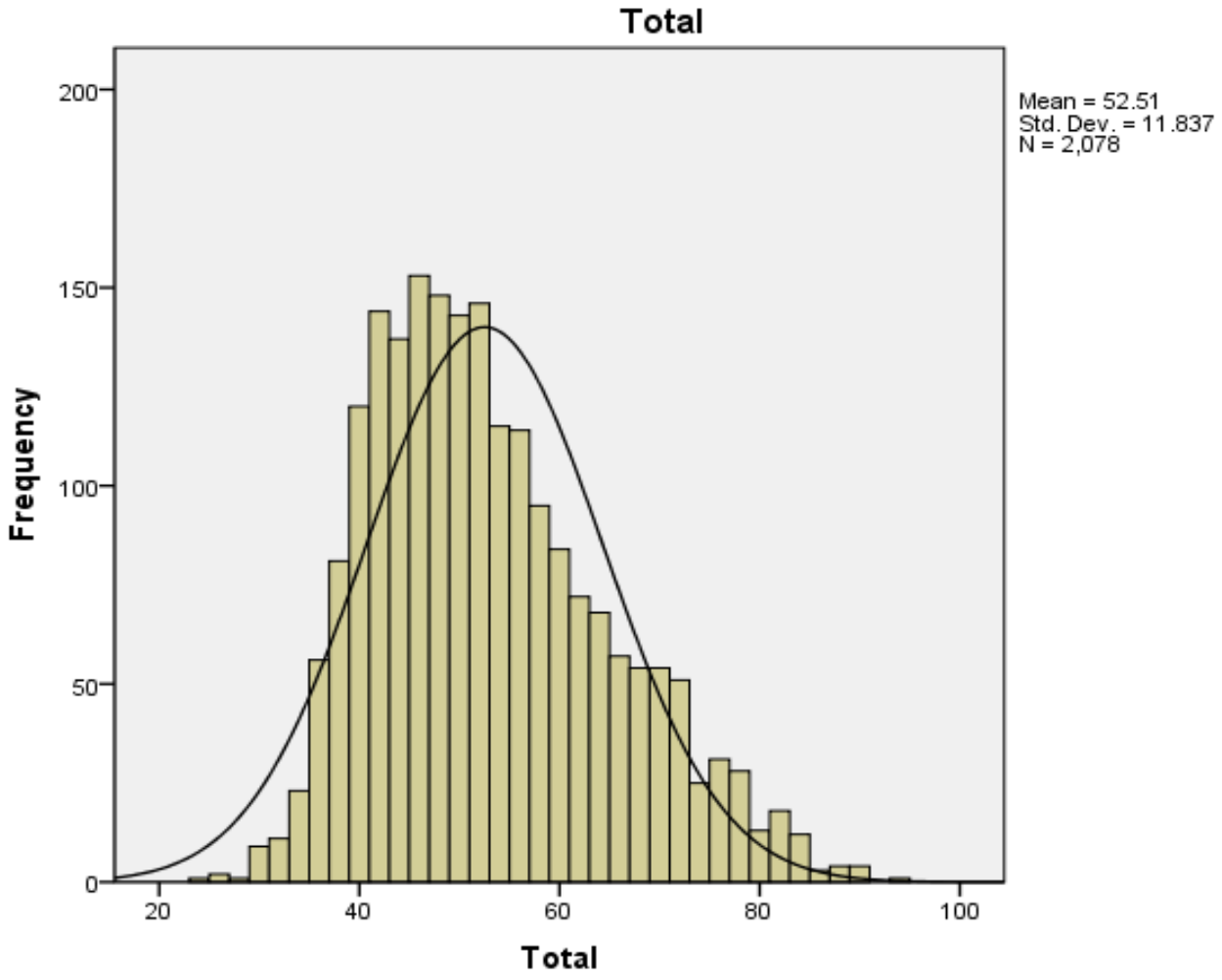
Item-Person Map



Conditional Standard Error of Measurement



Distribution of Number Correct Score



Reliability

- Reliability = true score variance/ true score variance + error variance
- The Cronbach's alpha for this test was 0.861
- True score / marginal reliability was 0.923

Differential Item Functioning

Table 4: Classification of DIF Items using MH Delta Index Across Gender

Matching Level	Category - C	Category - B	Category - A
Thin	5	5	95
Thick (Min. Freq.)	5	5	95

Table 5: Classification of DIF Items using MH Delta Index Across Experience

Matching Level	Category - C	Category - B	Category - A
Thin	2	3	100
Thick (Min. Freq.)	2	3	100

Magnitude of DIF will be evaluated as follows. Type A items - negligible DIF: items with $|\Delta\alpha_{MH}| < 1$ or *MH* test is statistically significant. Type B items - moderate DIF: items with $1 \leq |\Delta\alpha_{MH}| \leq 1.5$, and *MH* test is statistically significant. Type C items - large DIF: items with $|\Delta\alpha_{MH}| > 1.5$, and *MH* test is statistically significant.

Factorial Invariance

Testing for Data Fit of Four MI Models of the Engineering Test Structure across Groups

MI Model	χ^2	*df*	CFI	IFI	SRMR	RMSEA	90% CI for RMSEA *LL*	90% CI for RMSEA *UL*
Model A	13570.256	10645	.950	. 950	.033	.017	.016	.017
Model B	13973.617	10750	.949	.949	.034	.017	.016	.017
Model C	19750.017	10949	. 900	.899	.052	.027	.027	.028
Model D	20195.703	11064	.891	.890	.052	.028	.027	.028

Note. Model A = Configural; Model B = Weak; Model C = Strong; Model D = Complete.

Conclusions

- Engineering Learning Outcomes Exam data is essentially unidimensional thus allowing for the use of item response theory (IRT) calibration.
- The score reliability is quite high.
- Study support the configural invariance, weak factorial invariance and strong invariance of the measurement model across studied groups.

Conclusions

- Study found differential item functioning (DIF) across studied groups for few items.
- Distribution of test scores, both on the IRT logit scale and the classical number-correct scale, can be treated as normal.
- Classical and IRT analysis identified few items which have not desirable statistics.

Thanks for Listening

&

Soft Questions