Principles of Measurement - Final Exam

December 11, 2020

Task 1 (10 points)

Task 1a

Required for full marks:

- Compute the correlation between the test scores, $\hat{\rho} \approx 0.6245$.
- Adjust the correlation based on the reliabilities of the test scores, $\hat{\rho}_{\text{adj}} = \hat{\rho}/\sqrt{0.74 \times 0.68} \approx 0.8804$.
- Comment on the correlation in terms of the magnitude (very high) and state that it measures a linear relationship.
- Comment on the required assumption of uncorrelated errors between the test scores.

Task 1b

Required for full marks:

- Compute the estimated error variance from the reliability coefficient and the sample variance via $\sigma_E^2 = \sigma_Y^2 \times (1 - \rho_{Y,Y'})$. Estimate is 2.4. Also fine to directly compute the standard error of measurement.
- Compute the estimated standard error of measurement and the estimated 95% confidence interval (16.96, 23.04).

- State that the approach depends on normal distributed errors of the sum scores, which can be justified if there are many items with independent errors.
- Correct interpretation of the resulting confidence interval.

Task 1c

Required for full marks:

- Compute the linear equating function from the estimated means and standard deviations, $e_Y(x) = \frac{\sigma_Y}{\sigma_X}x + (\hat{\mu}_Y - \frac{\sigma_Y}{\sigma_X}\hat{\mu}_X)$. Plug in the value 12 and the estimated quantities to obtain the equated value 14.16.
- Compute the lower bounds to the reliability coefficients ($\hat{\alpha}_X \approx 0.855$, $\hat{\alpha}_Y \approx 0.779$) and reflect on indication of similar but not equal reliability, reflect that linear equating is symmetric, state that there is not much about equity and nothing about population invariance provided in the task.

Task 1d

Required for full marks:

- Choose model 2 based on the BIC (or the RMSEA). Evaluate model 2 in terms of RMSEA, SRMR and/or GFI, and state that it indicates good overall fit. Estimate coefficient omega from the factor loadings and error variances of model 2. Correct interpretation as the reliability coefficient for the sum scores.
- An unrestricted single factor model fits the data well, which means that coefficient alpha can be viewed as an appropriate lower bound to the reliability coefficient of the sum scores. However, the model fit of the unrestricted model is better which indicates that coefficient alpha is not equal to the reliability coefficient.

Task 2 (15 points)

General for all tasks to obtain full marks

- Accurate usage of terminology.
- Accurate classifications of sources of validity evidence.
- Appropriate weight in terms of what are the most important aspects of the context in question.

Task 2a

Answer should reflect the following to obtain full marks:

- A description of the test development process in brief terms.
- A presentation of a standard setting procedure to identify the required performance level.
- The following aspects should be covered in the response, but the explicit categorizations do not need to be included:
  - Reflection on content-oriented evidence: For example, construct representation by means of appropriate content sampling in terms of subjects, as well as items being of appropriate difficulty.
  - Reflection on response processes: For example, no irrelevant variance due to language barrier.
  - Reflection on internal structure: Reliability providing evidence of reproducibility of scores, and hence absence of random error variance.
  - Relationships to other variables: For example, predictive evidence of enhanced performance relative to individuals not subjected to accelerated teaching, construct representation.

Task 2b

Answer should reflect the following to obtain full marks:

- Reflection on content evidence: Can note that the content of the test is less important for the intended use of the test than its predictive capacity with respect to university success.
- Reflection on evidence based on internal structure: Note how score-reliability will influence prediction (i.e., correlation with predictor and criterion-measure scores).
- Relationships to other variables: Present a regression model, discuss requirements of some reliability of the test scores (reproducibility of scores, absence of construct-irrelevant variance). Demonstration that the test-scores predict success at university over and above the predictive capacity of high-school grades. Regression model could also include demographic variables to examine potential bias with respect to subgroups of the population.
- Reflection on consequences: Consider the impact on learning if grades are removed entirely or partly from college entrance decisions. Could, for example, cause students to direct their efforts at studying for that one particular test and focus less on their performance in each individual school subject.

Task 2c

Answer should reflect the following to obtain full marks:

- Content-oriented evidence: Examining and problematizing the extent to which the dimensions measured by the test-battery are appropriate with respect to representing the breadth of the potential dimensions along which schools can perform better or worse.
- Relationships to other variables, predictive evidence of validity: Examining the extent to which the incentive-structure causes lower performing schools to improve their teaching over a timespan.
- Relationships to other variables, convergent evidence of validity: Examining the extent to which the scores on the test-battery correlates with scores from another test-battery measuring a more diverse set of performance-related dimensions.

- Possible unintended consequences: Performance is related SES and other factors which may reinforce existing differences, more resources for better-performing schools most likely increases the divide between schools. An uproar among teachers because of the policy. Push towards studying for the tests rather than student learning.

Task 2d

Answer should reflect the following to obtain full marks:

- Description of how to collect data from the target population, evaluate the fit of a single factor model. Use scores in relation to status of individuals, decide diagnostic criterion based on the scores.
- Reflection on content-oriented evidence: Operationalization should be developed with indicators adhering to the theory of how panic disorder manifests. Expert panel could rate items with respect to this criterion.
- Reflection on internal structure: Reliability to demonstrate that the scores are not unduly influenced by random error variance.
- Reflection on relationships to other variables: Convergent and discriminant evidence with similar and different constructs measured using other tests. That is, scores on a test for the panic-disorder construct should correlate strongly with other tests also intended to measure panic disorder, and less strongly with tests intended to measure other constructs.
- Description of a standard-setting method that finds an appropriate cut-off point for delineating threshold for categorization.

Task 3 (25 points)

General for all tasks to obtain full marks

- Accurate usage of terminology.
- Accurate classifications of sources of validity evidence.
- Appropriate weight in terms of what are the most important aspects of the context in question.

a) (5 points)

Answer should be consistent with the following to obtain full marks:

- An analysis of the scale questions is provided.
- Consideration of potential for construct under-representation and construct contamination/irrelevant variance.
- Consideration of potential shift in content-meaning as a consequence of translation to Norwegian.
- Response can include reflections on what kind of evidence is required to rule out sources of non-validity.
- Reflection on the meaning of the scores produced.

b) (10 points)

Answer should include the following to obtain full marks:

- Fit a single factor model and evaluate it with respect to the fit indices, see Table 1 for the statistics for each candidate ID (values can vary somewhat with different estimation methods).
- Make a judgment about the model fit, which should indicate an acceptable but not good fit.
- Conduct a residual analysis by looking at the residual matrix and noting if there are entries larger than 0.1 or some other well-motivated approach.
- Compute the reliability coefficient (see Table 1) and interpret it appropriately, with a reflection that the model is not entirely appropriate which may make coefficient omega not reflective of the reliability coefficient.
- Provide comments on the results that make sense.

c) (2 points)

Answer should be consistent with the following to obtain full marks:

- Reflection on the impact of administering the scale on the context in which it is used.
- Consider both intended and unintended consequences of the scale score use.
- Primary focus should be on a sound and reasonable argumentation with respect to the potential consequences of using the test for this purpose and in this context.

d) (8 points)

- Select items with respect to some statistical criterion such as the standardized factor loadings, the item information or the item-to-sum score correlations or with respect to a well-motivated content-oriented approach (possibly supported by a quantitative analysis).
- Present the five items in question and describe them (the students will select different items – the properties were randomly generated).
- Evaluate the reliability of the sum scores of the shortened scale by considering coefficient omega and discuss the impact of shortening the scale in terms of the scale score reliability.
- Reflect on content coverage with the shortened test.

Table 1: Model statistics for each candidate number (ML estimation).

ID	GFI	SRMR	RMSEA	ω
1	0.937	0.048	0.089	0.853
2	0.941	0.045	0.085	0.851
3	0.940	0.047	0.085	0.846
4	0.949	0.041	0.081	0.854
5	0.950	0.041	0.079	0.848
6	0.953	0.040	0.076	0.848
7	0.951	0.046	0.076	0.845
8	0.941	0.046	0.085	0.850
9	0.961	0.039	0.068	0.839
10	0.941	0.049	0.085	0.838
11	0.935	0.050	0.089	0.850
12	0.948	0.049	0.079	0.846
13	0.957	0.038	0.074	0.850
14	0.955	0.041	0.074	0.838
15	0.930	0.052	0.093	0.850
16	0.966	0.034	0.063	0.845
17	0.954	0.039	0.074	0.849
18	0.957	0.039	0.073	0.847
19	0.948	0.046	0.078	0.841
20	0.961	0.039	0.070	0.845
21	0.955	0.040	0.074	0.850
22	0.937	0.048	0.088	0.847
23	0.962	0.036	0.068	0.847
24	0.953	0.039	0.076	0.848
25	0.950	0.042	0.080	0.846
26	0.947	0.047	0.079	0.847