UnderStandingAmericaStudy

UAS85 - SCORING EXPLANATION

USC Dornsife Center for Economic and Social Research

9/21/2018

SCORING EXPLANATION

The cognitive measures (number series, picture vocabulary, verbal analogies) were taken from the Woodcock–Johnson Tests of Cognitive Abilities® (Mather and Jaffe, 2016). The tests were designed to measure the respondent's quantitative reasoning (number series) and lexical knowledge (picture vocabulary, verbal analogies). Each measure consists of 15 items, which are scored dichotomously as correctly solved or incorrect.

Cognitive test scores for UAS panel respondents are derived using a two-parameter logistic Item Response Theory (IRT) model. In this IRT model, the probability of correctly solving a test item is viewed as a function of a test taker's ability level and the difficulty and discrimination parameters of the test item. The difficulty parameter measures the ability level at which there is a 50% chance of answering the item correctly, whereas the discrimination parameter measures how sensitive this probability is to differences in the ability level. The two-parameter logistic model allows both the difficulty and discrimination parameters to differ across test items.

IRT scoring requires sufficient unidimensionality of the test items, which was evaluated with confirmatory factor analysis for binary outcome variables. Common criteria for adequate model fit include a root mean squared error of approximation (RMSEA) less than .06, Tucker-Lewis index (TLI) greater than .90, and comparative fit index (CFI) greater than .90. A one factor model provided a good fit to the data for the number series and verbal analogies tests [number series: RMSEA = .041 (90% CI = .038/.044), TLI = .90, CFI = .91; verbal analogies: RMSEA = .022 (90% CI = .018/.026), TLI = .98, CFI = .98]. Adequacy of model fit was somewhat less consistent for the picture vocabulary test [RMSEA = .042 (90% CI = .039/.046), TLI = .80, CFI = .83], but was considered sufficient to support unidimensionality.

Whereas the verbal analogies test items in UAS44 consisted of 15 items from Form A, the UAS85 administered 15 items from the alternate Form B (together with 4 selected items from Form A). This was done to mitigate practice effects for repeated test-takers. A critical issue when using multiple test forms over time is that how to calibrate the data onto a common scale so that they can be directly compared (e.g., in longitudinal analyses). To ensure that the verbal analogies test scores in UAS85 were calibrated on the same matric as those in UAS44, IRT linking methods were used. Specifically, when calibrating the verbal analogies test in UAS85, the item parameters of the 4 Form A items (which were used in both UAS44 and UAS85) were prefixed at the values determined in the UAS44 calibration sample. Item parameters of the 15 Form B items (as well as the mean and variance of the latent cognitive proficiency scores) were freely estimated in the

UAS85 calibration sample. Such development of a common scale is theoretically justified by the invariance property of IRT modeling (Lord, 1980).

Item difficulty and discrimination parameters were calibrated based on weighted samples of 2,579 UAS respondents, with weights ensuring that the demographic variables race, sex, age, education, and household income in the survey sample match their population counterparts. The estimated item parameters are shown in Table 1. Because weights are unavailable for the LA County subsample, these were not included in the estimation. However, we compute scores for all respondents, regardless of whether they have a weight or not, provided that they answer at least one item. The only exception is that respondents who completed the picture vocabulary test in Spanish do not receive scores on this measure. The reason is that the parameters for items of this lexical knowledge test may differ between languages, but the sample size of Spanish test takers is currently too small to evaluate this.

The final IRT-based scaled scores are converted into T-scores, where 50 is the mean and 10 is the SD of a census-weighted sample of the general United States population. The T-score metric has widespread use in psychological testing and has been adapted, for example, by the Patient-Reported Outcomes Measurement Information System (PROMIS®, Cella et al., 2010). A score of 50 means that the person's cognitive ability is equal to that of the average person in the general population, a score of 60 means that the person's ability is one standard deviation above average, and a score of 40 means that the person's ability is one standard deviation below average.

Table 1: Item parameters of the verbal analogies test administered in UAS85 (N=2,579)

Item	Difficulty	Discrimination
vea_31 [a]	-3.537144	1.374603
vea_32 [a]	-1.024851	2.383876
vea_41 [a]	-0.6252622	3.456446
vea_51 [a]	-0.2907981	2.751877
veb_11	-2.01947	1.167854
veb_12	-2.8253	1.684404
veb_13	-2.75946	1.012478
veb_21	-2.20259	1.325372
veb_22	-1.90122	1.844262
veb_23	-0.81129	2.12974
veb_31	-3.64363	0.74165
veb_32	-1.58716	1.04458
veb_33	0.536478	1.661343
veb_41	-1.32511	1.465988
veb_42	-1.01314	3.249794
veb_43	-0.65488	1.330163
veb_51	0.501523	1.472374
veb_52	1.824073	0.873192
veb_53	1.853467	1.328828

Note: a difficulty and discrimination parameters for these test items were prefixed at the values determined in the UAS44 calibration sample.

REFERENCES

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology, 63, 1179-1194

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mather, N, Jaffe, L.E. (2016). Woodcock-Johnson IV: Reports, Recommendations, and Strategies. Jossey-Bass: Hoboken, NJ.