UnderStandingAmericaStudy

UAS484 - SCORING EXPLANATION

USC Dornsife Center for Economic and Social Research

11/1/2022

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 picture vocabulary test items in UAS43 and UAS293 consisted of 15 items from Form A, the UAS484 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 picture vocabulary test scores in UAS484 were calibrated on the same matric as those in UAS43 and UAS293, IRT linking methods were used. Specifically, when calibrating the picture vocabulary test in UAS83, the item parameters of the 4 Form A items (which were used in UAS43, UAS84 and UAS293) were prefixed at the values determined in the UAS43 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 UAS484 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 2,672 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 picture vocabulary test in UAS484 (N=2,672)

Item	Difficulty	Discrimination
pva_12 [a]	-2.37670	1.501935
pva_23 [a]	-1.24801	1.178884
pva_42 [a]	-0.293597	1.809553
pva_51 [a]	0.578610	1.050929
pvb_11	-1.56834	1.422314
pvb_12	-2.76452	1.659332
pvb_13	4.01591	0.267922
pvb_21	2.07605	0.229604
pvb_22	-0.94619	1.26501
pvb_23	-0.01121	0.714774
pvb_31	-2.15409	1.765476
pvb_32	-0.54762	1.302921
pvb_33	0.54544	2.479717
pvb_41	0.31272	1.533608
pvb_42	0.73104	1.385429

Note: [a] difficulty and discrimination parameters for these test items were prefixed at the values determined in the UAS43 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.