

Assignment: Statistical model and your example assessment.

Choice: NAEP Mathematics

In this paper, I will discuss the statistical model that is employed in the NAEP assessment and how it relates to the psychological model discussed in the previous paper.

Nationally representative probability samples of eighth-grade students are selected using a complex multi-stage sampling design that involves sampling schools within selected geographic areas across the country and adding sampling weights to ensure that valid inferences can be drawn between the student samples and the respective populations from which they are drawn. These weights are added to account for the disproportionate concentration of Black and/or Hispanic students in some schools; for students who attend non-public schools; and to account for the lower sampling rates for small schools. Minority students are also over-sampled, to help meet this end. To help ensure adequate sample representations for each state, NAEP provides substitutions for non-participating public and non-public schools, so that the representation of the student population corresponds to figures from the US Census.

Analyses are conducted to determine the percentages of students who give various responses to each of the cognitive and background questions administered in the test. When computing these percentages, missing responses prior to the last intentional response (on a student's block) are treated as intentional omissions whereas missing responses at the end of the blocks administered are considered "not reached." Therefore, the percentages of responses calculated for each question are based on the number of completed questions, using this rule. For multiple-choice and short constructed-response questions, a lack of response was treated as if the student had not been administered the final question. However, if the last item of the test was an extended constructed-response question, it was treated as if the student intentionally did not answer the question.

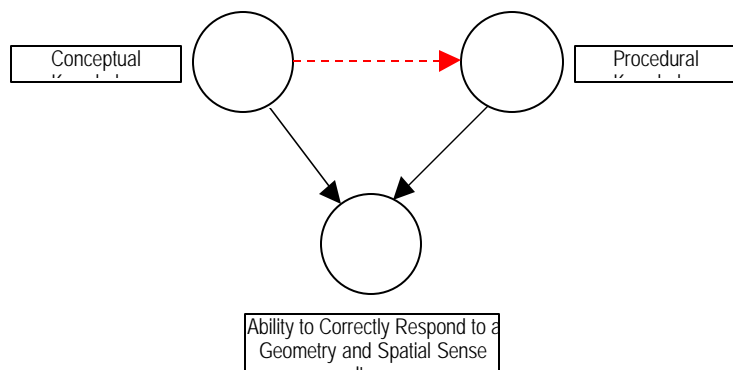
Item response theory is used to estimate the average mathematical scale scores. These scores are computed for the nation, for various subgroups of interest within the nation, and for states and territories. The main purpose for this analysis is to provide a common scale on which performance characteristics can be compared for the various subgroups. Because a BIB-spiraling design is utilized these composite scores are designed to represent the distribution of the performances of the population (note: students are administered blocks of items from any given test and are not administered an adequate number of items to make reliable inferences of their individual performances). A scale ranging from 0 to 500 was created to report population performances for each content strand, summarizing student performance across all three distinct assessment types (multiple-choice, short constructed-response, and extended constructed response). The scales have a mean of 250 and a standard deviation of 50. A composite scale was also created to render a weighted average across the content strands; where the middle of the scale was targeted for the eighth grade performances (whose scores in 1996 ranged between 231 and 375). It should be noted that the lower range of the overall scale pertains to the fourth-grade assessment and the upper range of the scale targets the twelfth-grade assessment.

Three different IRT models are used, based on the different item types in the assessment. Multiple-choice items are scaled using the three-parameter logistics model (3PL); short constructed-response items are scaled using the two-parameter logistics model (2PL) for right/wrong responses; and the

extended constructed-response questions are rated using the Muraki partial credit model (GPL), thus allowing raters to award partial credit for student answers that may be incorrect with regard to their final solution, but are correct in the cognitive processes employed to answer the question.

What is striking about this statistical model is that it demonstrates a departure from the student and evidence models found in the psychological model as set forth by the frameworks committee. The IRT models employed in the statistical model are, by design, only able to capture students' knowledge within the content strands. The frameworks committee's intention was to objectively measure students "Mathematical Abilities" (a.k.a. cognitive abilities) as well as their "Mathematical Power" (a.k.a. reasoning and communicative skills). The problem presented here is that the NAEP items may be clear in their assessment of content domain being assessed, yet at the same time, although a student may need to have the conceptual or procedural knowledge to complete an item, there is no independent way of isolating those attributes as a separate score.

It seems that even in the presence of the partial credit model (for extended constructed-response questions), the cognitive processes cannot be independent of each other, e.g., a student must have the conceptual knowledge as well as the procedural knowledge to understand how to solve a complex item. In the case of this item type, partial credit can be given, but only to the extent that the raters have identified that the correct overall cognitive processes have been utilized. In addition, the test developers have taken great effort to not only identify which content strand each item intends to measure, but have also identified the cognitive and communicative attributes the item is intended to measure as well. To this end, it appears that in the statistical model, the best probabilities that one could hope for would be conditional probability distributions where the cognitive factors being considered would be dependent upon each other, thus confounding any clear analysis of these factors. Following is a diagram illustrating an example of this thought:



It seems to me (at best) that any conditional probabilities that could be calculated, would be calculated to the extent that the overall cognitive abilities could only be measured and, as such, a classification can be made about students in rating their performance as either adequate or inadequate.

In thinking in terms of a Bayes nets, it seems that these cognitive variables may be considered conditionally independent. However, this may not hold true in all situations. If the item type were multiple-choice, a student might be able to provide the correct response because he/she

understands it conceptually; but the procedural knowledge variable may not provide additional information about students in this item type. Finally, it appears that it would be difficult to illicit a performance indicative of these behaviors alone. The scoring model for this assessment allows only for correct/incorrect ratings, and the only allowances made to document these constructs would be in the partial-credit scoring model of extended constructed-response items. If all of the “Mathematical Abilities” and “Mathematical Power” can, at best, be identified as overall concepts, their individual component behaviors cannot be isolated and rated independently by making those inferences inductively. Another problem with trying to isolate these behaviors is that it seems possible that each of these performances (children) may have multiple parents. Throughout all descriptions of the different disclosed items in the different content strands, the test developers have indicated that the items require these behaviors to produce correct responses. Again, the results are confounded.

F. has indicated that the only solution might be to document the evidence through the use of a join tree, by clustering the variables into overlapping cliques. However, I’m not certain that this methodology would be easily applied to the current paper-and-pencil assessment. It seems that it may work better in a computer-delivered environment where information about every keyboard or mouse action can be captured and analyzed to evidence the different cognitive constructs (e.g., procedural knowledge) in an unconfounded manner. However, I wonder shifting to the new “Technically Rich Environments” (i.e., computer test administrations for NAEP) would produce other confounding variables such as expertise of students’ computer skills. In this vein of thought, it seems to me that the proficiency that students possess would vary nationally as well as within each state, based on such factors as the economics of the school district or the financial situation of individual families. Would this added variable now confound the content strand variables? Does the question now become, “Are we measuring a students’ content knowledge or their ability to convey their content knowledge through computer-administered tests?”

A final thought. B. recently completed an investigation of NAEP variables using factor analysis as well as cluster analysis. She ran her analyses on the NAEP fourth-grade mathematics assessment data. She was trying to ascertain if these types of analyses would reveal the cognitive dimensions underlying the framework of the test. The results of her cluster analysis led her to the conclusion that, “...the cluster method did not identify the underlying content or cognitive processes as specified by the test developers.” The factor analysis led her to the same conclusion. She indicated that this was because the items were not grouped according to content or cognitive processes, but were instead grouped by the difficulty of the mathematical concept and the complexity of the procedure needed to solve the item. Since the difference between the fourth-grade assessment differs from the eighth-grade assessment only in terms of task complexity and deeper content knowledge, her analysis may also apply to the eighth-grade assessment.