Estimating an Observed Score Distribution From a Cognitive Diagnostic Model

Gary Skaggs¹, Jesse L. M. Wilkins¹, and Serge F. Hein¹

Keywords

diagnostic testing, estimated distributions, attribute profiles

Cognitive diagnostic models (CDMs) have been a focus of research in recent years as a way to provide diagnostic information instead of a single test score on educational and psychological tests. CDM-based tests are designed to provide attribute profiles, or classifications of examinees that usually involve mastery or nonmastery, on each of the attributes measured by the test (Rupp, Templin, & Henson, 2010). As our knowledge of CDMs expands, there is likely to be a need for a way to link attribute profiles and observed test scores. The purpose of this brief report is to introduce and demonstrate a method for estimating an observed score distribution from an attribute profile.

Although any CDM can be used with the proposed method, the authors chose the Log-Linear Cognitive Diagnostic Model (LCDM; Henson, Templin, & Willse, 2009) because many popular CDMs can be derived as special cases. The probability of a correct response to item i for an individual with an attribute profile α_c is specified as

$$P(x_{ic} = 1 | \alpha_c) = \frac{\exp(\lambda_i + \lambda_i^T h(\alpha_c, q_i))}{1 + \exp(\lambda_i + \lambda_i^T h(\alpha_c, q_i))}, \quad (1)$$

where the q_i 's are the entries of the Q-matrix, and $\lambda_i^T h(\alpha_c, q_i)$ is a vector of main and interaction effects weighted by their slopes. The entries of the Q-matrix are 1 if the item measures a particular attribute or 0 if it does not. The profile α_c is a vector of 0s and 1s in which 1 indicates mastery of an attribute and 0 indicates nonmastery. Thus, $\lambda_i^T h(\alpha_c, q_i)$ will be nonzero if the item measures an attribute and the examinee has mastered that attribute. This means that the probability of a correct response increases with more attributes being mastered. The item parameter, λ_i , is the intercept of the model and indicates a baseline difficulty for the item. In this demonstration, only main effects in the LCDM were used, which makes the model equivalent to the Compensatory Reparameterized Unified Model (CRUM; Hartz, 2002).

Method

Observed score distributions for each attribute profile were estimated by adapting a recursive formula for a compound binomial, introduced by Lord and Wingersky (1984) and elaborated

¹Virginia Tech, Blacksburg, VA, USA

Corresponding Author:

Gary Skaggs, Associate Professor, Educational Research and Evaluation Program, School of Education, Virginia Tech, 1750 Kraft Drive (CRC), Rm. 2106, Blacksburg, VA 24061, USA.

Email: gskaggs@vt.edu

upon by Kolen and Brennan (2004), that was used originally for item response theory observed score equating. For a test composed of two items, the proportion of examinees with a score of 0 is $(1 - p_1)(1 - p_2)$, the proportion with a score of 2 is $(p_1)(p_2)$, and the proportion with a score of 1 is $(p_1)(1 - p_2) + (1 - p_1)(p_2)$, where p_1 and p_2 are the probabilities of a correct response based on the latent trait model. More generally, for a test composed of r items, the estimated proportion of examinees from attribute profile α_c that have an observed score of x , or $f_{rc}(x|\alpha_c)$, is as follows:

$$f_{rc}(x|\alpha_c) = f_{r-1}(x|\alpha_c)(1 - p_{ic}) + f_{r-1}(x - 1|\alpha_c)p_{ic}, \text{ for } 0 < x < r \quad (2)$$

where p_{ic} is the probability of a correct response to item i for an examinee with attribute profile α_c , according to the CDM. For scores of 0 and r , the above formula reduces to the following simpler versions:

$$f_{rc}(x=0|\alpha_c) = \prod_{i=1}^r (1 - p_{ic}). \quad (3)$$

$$f_{rc}(x=r|\alpha_c) = \prod_{i=1}^r (p_{ic}). \quad (4)$$

Results

We demonstrate the estimation procedure using two sets of items taken from the *Trends in International Mathematics and Science Study (TIMSS) Assessment 2003 for Grade 8 Mathematics*. These items and data were used in a previous study (Skaggs, Hein, & Wilkins, in press) in which Q-matrices with varying numbers of attributes were fit to the data. Here, a Q-matrix for four attributes is used: whole numbers, geometry, fractions, and algebra. For the first set of items, estimates from 20 dichotomously scored items that appeared across six TIMSS booklets were used. Ten items measured one attribute while the other 10 items measured two attributes. The items were selected because their parameters were precisely estimated, and the item response data fit the CDM. In addition, there was fairly equal coverage of the four attributes, thus simulating a situation in which a test was generated from an item bank by selecting high quality items. Because of the TIMSS booklet design, no examinees responded to all 20 items, and so observed score distributions could not be directly calculated.

In addition, a second set of 30 dichotomously scored items from TIMSS Booklet 1 was used to compare observed versus estimated raw score distributions. In this booklet, coverage of the four attributes ranged from six to 13 items per attribute. Nineteen items measured one attribute while 11 items measured two attributes. This comparison is not ideal since not all items fit the CDM equally well, nor was there equal coverage of the attributes, but it does serve as a rough validation of the procedure. The mdltm software program (von Davier, 2005) was used to estimate item parameters and examinee profiles. The estimated distributions were easily calculated in an EXCEL worksheet.

Table 1 shows the estimated observed score distributions for five attribute profiles in which different numbers of attributes were mastered. As expected, attribute profiles in which more attributes have been mastered resulted in higher estimated observed scores. It should be noted here that the location and shape of the estimated distributions depend on the item parameter estimates. As model intercepts decrease, the estimated distribution shifts lower, and vice versa. As attribute slopes increase, the distributions for attribute profiles become more peaked and therefore distinct from each other.

Table 1. Estimated Observed Score Distributions for Five Attribute Profiles.

Observation score	0000	0100	1001	0111	1111
0	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.01	0.00	0.00	0.00	0.00
6	0.03	0.01	0.00	0.00	0.00
7	0.07	0.03	0.01	0.00	0.00
8	0.12	0.07	0.03	0.01	0.00
9	0.17	0.12	0.07	0.02	0.01
10	0.19	0.17	0.12	0.06	0.03
11	0.17	0.19	0.17	0.11	0.06
12	0.12	0.17	0.19	0.16	0.11
13	0.07	0.12	0.17	0.20	0.17
14	0.03	0.07	0.12	0.19	0.21
15	0.01	0.03	0.07	0.14	0.19
16	0.00	0.01	0.03	0.07	0.13
17	0.00	0.00	0.01	0.03	0.07
18	0.00	0.00	0.00	0.01	0.02
19	0.00	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.00

If the intercepts and slopes differ systematically between attributes, it would be possible to have estimated score distributions for attribute profiles with fewer attributes mastered that are higher than profiles with more attributes mastered. This inversion of attribute profile distributions, however, is not the most likely scenario. In Figure 1, the estimated distributions for all 16 attribute profiles were plotted. These form five distinct groups of attribute profiles, each based on the number of attributes mastered. Thus, for example, the five attribute profiles in which two attributes were mastered all had very similar distributions. The distributions for attribute profiles consisting of three to 10 attributes were estimated using TIMSS item parameter estimates for the CRUM model. In all cases, it was found that estimated distributions for profiles with the same number of attributes mastered were very similar. This means that at extreme score ranges, the number of attributes mastered can be predicted quite well. For example, examinees who score at the extremes (say greater than 17 or less than 6) are likely to have mastered or not mastered most of the attributes, respectively. For these examinees, further diagnostic testing might not be necessary. It is important to note that this relationship may not generalize to noncompensatory CDMs because, in such models, mastery of all attributes is necessary to increase the probability of a correct response.

To help validate this procedure, the TIMSS items and examinee responses from Booklet 1 were used. There were 757 examinees, 309 of which were estimated to have mastered all four attributes. The raw score distribution of these examinees was compared with the distribution estimated using the above procedure. This comparison is shown below in Figure 2. The observed distribution is not smooth, due to the relatively small sample size, but it does roughly correspond to the expected distribution. This finding, however, must be viewed as tentative. Four items had parameter estimates with relatively large standard errors and fit statistics, and this booklet did not cover all four attributes equally well. Item parameter estimation error undoubtedly impacts the precision of the estimated score distributions. In addition, comparisons

Figure 1. Estimated attribute profile distributions.

Figure 2. Expected versus observed score distribution for profile 1111.

involving other profiles may show different results. Still, these results do show some promise for the proposed estimation method.

This method of score distribution estimation has been used in at least two studies so far: by Skaggs, Hein, and Wilkins (in press) for standard setting and by Xin and Zhang (2015) for test equating. However, it must be pointed out that research on CDMs is relatively new, and there are as yet few actual applications of CDMs in large-scale testing. Nevertheless, one use of the proposed method that can be predicted is its use as a screening device. Depending on the number of attributes, a test that accurately estimates attribute profiles could be quite lengthy or require a computer adaptive administration. One likely purpose of a CDM-based test will be to determine whether examinees are in need of an educational or psychological intervention.

Based on the relationship shown in Figure 1, a shorter version of such a test could serve as a screening device to quickly classify the most extreme cases and thus save significant testing time and resources. Along this vein, another potential application would be the equating of the shorter screening form with its full-length counterpart.

It is important to point out several limitations and directions for future research. This method of estimating score distributions depends on known item parameters, which instead are estimated from data. The accuracy of these estimates is in turn dependent on correct model and Q-matrix specification and sample size. The effect of these factors should be investigated. In addition, the results in this brief report are based on a compensatory model. The method should be replicated for noncompensatory CDMs.

Acknowledgment

The authors would like to thank the editor and two anonymous reviewers for their helpful comments on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Science Foundation REESE program (DRL-1109429).

References

- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.
- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 452-461.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Skaggs, G., Hein, S. F., & Willkins, J. L. M. (in press). Diagnostic profiles: A standard setting method for use with a cognitive diagnostic model. *Journal of Educational Measurement*.
- von Davier, M. (2005). Mdltn: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: Educational Testing Service.
- Xin, T., & Zhang, J. (2015). Local equating of cognitively diagnostic modeled observed scores. *Applied Psychological Measurement*, 39, 44-61.