IRT Scoring and Test Blueprint Fidelity

Applied Psychological Measurement

2018, Vol. 42(5) 393–400

© The Author(s) 2018

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621618754897

journals.sagepub.com/home/apm

Gregory Camilli¹

Abstract

This article focuses on the topic of how item response theory (IRT) scoring models reflect the intended content allocation in a set of test specifications or test blueprint. Although either an adaptive or linear assessment can be built to reflect a set of design specifications, the method of scoring is also a critical step. Standard IRT models employ a set of optimal scoring weights, and these weights depend on item parameters in the two-parameter logistic (2PL) and three-parameter logistic (3PL) models. The current article is an investigation of whether the scoring models reflect an intended set of weights defined as the proportion of item falling into each cell of the test blueprint. The 3PL model is of special interest because the optimal scoring weights depend on ability. Thus, the concern arises that for examinees of low ability, the intended weights are implicitly altered.

Keywords

assessment, guessing, item response theory, psychometric theory, scoring

Multiple choice (MC) is the most common type of item format because of more economical development and scoring. Typically, only one option is correct for each item, yielding a dichotomous item score (wrong = 0, right = 1). As noted above, assessment programs may use item response theory (IRT) for conducting development, calibration, and scoring activities. A number of IRT models are available for scoring with MC items, including the one-parameter logistic (1PL; Rasch), two-parameter logistic (2PL), and three-parameter logistic (3PL) models. As shown below, each of these models has implications for how the method of scoring aligns to content representation. For simplicity, a standard test design or blueprint where items are classified into a cells corresponding to content subdomains (or strands) is assumed. For simplicity, the case in which items are cross-classified by content and cognitive process is not examined, but this does not limit the ultimate conclusions of the article. *Blueprint fidelity* in this study refers to how well the method of scoring matches the intended weights specified in an assessment design with respect to content domains.

Test blueprints are often detailed in a framework that guides both content focus and the development process for an assessment instrument. Frameworks capture both subject-specific skills as well as skills of cognitive processes. Because an assessment should reflect attainment and changes in instructional practices and educational objectives, the items constituting an

¹Rutgers, The State University of New Jersey, New Brunswick, USA

Corresponding Author:

Gregory Camilli, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901-1108, USA.

Email: greg.camilli@gse.rutgers.edu

assessment need to be adequately sampled according to framework categories and associated specifications. Given that a test is carefully assembled according to a framework, a method of scoring should be consistent with the intended coverage. The question then arises of whether an IRT model produces scores that are consistent with an intended framework.

Different IRT scoring models implicitly lead to alternative scoring weights. This gives rise to the possibility that IRT scoring formulas may reweight blueprint representation for certain examinees. This may be an innocuous effect if a test is strictly unidimensional—in which case item weights only affect scoring efficiency. However, tests are rarely unidimensional, and many assessments are not developed in a way that is consistent with a unidimensional framework. For example, a mathematics test may include a number of strands (e.g., measurement, algebra) and/or levels of complexity. Test scores, if they are to be interpreted relative to the test blueprint, should be consistent with that design rather than an implicit assumption of unidimensionality.

This issue can be thought of in terms of Bayesian item selection. As noted by Veldkamp (2010), item selection in a practical context involves at least three different types of test specification:

Some specifications deal with item content, item type, answer key, gender bias, or ethnical bias. These specifications are generally denoted as categorical criteria. Other criteria deal with quantitative specifications like word count or time limit. These criteria can be formulated as a function of the items involved. Psychometric aspects can also be described as test specifications. In fact, they are examples of quantitative test specifications. The aspects can be formulated as a function of the items involved. (p. 155)

If a set of items meets the desired specifications, then one condition is satisfied for score interpretation to be consistent with the test blueprint. For adaptive tests, the specifications are used to construct Bayesian selection rules. Linear forms of test can also be assembled according to a number of criteria. However, once a set of items or a form is constructed, a scoring model must be applied to provide a proficiency estimate. A second condition for accurate score interpretation is that the scoring model must be consistent with the desired specifications.

The standard 2PL IRT model employs discrimination parameters in the estimation of ability. An additional c parameter is used in the 3PL model to disentangle both item and ability estimates from the effects of guessing. The c parameter presents a number of challenges, and a variety of psychometric models and approaches have been proposed (Cao & Stokes, 2008; Mariani, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; Sheng, 2013; van der Linden, Klein Entink, & Fox, 2010; von Davier, 2009; Wang & Xu, 2015; Yamamoto, 1989). Nonetheless, the 1PL, 2PL, and 3PL model remain the predominant practical models for IRT scoring. A brief sketch of the work by Chiu and Camilli (2013) below illustrates how a standard 3PL IRT model may affect blueprint representation for individuals. This provides a foundation for the current article, which involves how well a scoring method aligns to a test blueprint in terms of content.

IRT Models and Optimal Weights

In IRT, examinees with different ability levels have different probabilities of answering an item correctly. In terms of the 3PL IRT model, the probability P for an examinee with ability θ is given as

$$P(u = 1 | \theta, a, b, c) = c + (1 - c)\Psi, \quad (1)$$

where

$$\Psi = \frac{\exp[Da(\theta - b)]}{1 + \exp[Da(\theta - b)]}. \quad (2)$$

Above, $D = 1.7$ is a scaling constant, and a , b , and c are an item's discrimination, difficulty, and guessing parameters, respectively.

Birnbaum (1968) first showed how *locally best weights* can be obtained from the 2PL and 3PL model and thus created the link to the observed score metric (Birnbaum, 1968, Section 19.3).¹ Chiu and Camilli (2013) represented this optimal item weight as

$$w(\theta) = \frac{a}{1 + c\Omega}. \quad (3)$$

It can be seen that as the odds $\Omega = (1 - \Psi)/\Psi$ of getting the item incorrect *increases*, the scoring weight *decreases*. These scoring weights, which are derived from the maximum likelihood estimate of ability (Chiu & Camilli, 2013; Chiu, 2011), can be used to demonstrate the empirical weight given to the cells of a test blueprint. For a 2PL model, the scoring formula simplifies to $w(\theta) = a$, and for the 1PL model, the optimal weight is $w = 1$ across persons and items. Let u represent the scored (0 or 1) response on the item. Observed item responses u_i for an examinee, when weighted sums are obtained with $w_i(\theta)$ across items i , result in a sufficient statistic for ability for the 1PL model (and the 2PL model if a is known). No corresponding sufficient statistic exists for the 3PL model.

Because $w(\theta)$ is relatively lower for lower ability examinees when scoring with the 3PL model, a higher proficiency examinee will receive more credit for a correct response on a particular item. This raises a fairness issue because equal item performance is unequally rewarded (Camilli, Briggs, Sloane, & Chiu, 2010). Another issue raised in both 2PL and 3PL scoring is whether scores obtained for examinees at different levels of proficiency accurately reflect the domains of test content. This is the main topic explored below.

Test Design

For a mathematics test, the content areas might be defined in terms of number sense, measurement, algebra, probability, and so forth, whereas processes might be defined as procedural, conceptual, and problem solving. These two facets are typically crossed in a matrix, and a test is initially designed by determining what proportion of total items will fall into each cell. Below, the focus is solely on content classifications to illustrate the effects of scoring. For standard IRT models, optimal (or locally best) weights can be used to quantify the degree of match with the test blueprint. Assume there are $k = 1, 2, \dots, K$ cells in the test blueprint, and assume there is a set of framework-based weights corresponding to each cell. The question then arises of how closely these two sets of weights match.

With 1PL, 2PL, and 3PL scoring, the optimal weights determine the relative emphasis for each content area k . For a given cell k of the test blueprint, the normalized scoring weight for an individual with ability θ can be written as

$$p_k(\theta) = \frac{\sum_i^{n_k} w_i(\theta)}{\sum_i^N w_i(\theta)}, \quad (4)$$

where i is the item index, N represents the total number of items in the blueprint, and n_k is the number of items in cell k . With a 1PL model, $p_k(\theta) = n_k/N$ because $w_i(\theta) = 1$ for all i . Under the

2PL model, examinees receive the same scoring weights $w_i(\theta) = a_i$ for an item. Under the 3PL model, weights vary with estimated ability, as shown in Equation 3. Note that classifying items into a set of framework cells to some degree is a subjective process, and items may have fuzzy membership in more than one content cell. For this reason, the overall weight of a cell averaged across items is more meaningful than the weight of any one particular item.

An important question is how the intended set of cell weights is identified. In many cases, weights are specified in the test blueprint or framework as the percentage of items devoted to each content area. For example, for the National Assessment of Educational Progress (NAEP),

When there are several content-related subscales defined for a subject area, the subscales are combined into a composite scale used to describe the overall attainment of students in that subject area. The weight of each subscale in the composite reflects the relative importance of the scale as specified in the framework developed by the National Assessment Governing Board. The weights are approximately proportional to the number of items in each scale at a given grade level.²

In 2011, the weights for Grade 4 were given as Number Properties and Operations (.40); Measurement (.20); Geometry (.15); Data Analysis, Statistics, and Probability (.10); and Algebra.

The framework weights can be used to represent the baseline for determining blueprint fidelity. For the purposes of this article, it is assumed that a test is intentionally assembled so that the proportion of items in each cell of the test blueprint matches the cell's desired scoring weight. Automatic assembly for linear test forms (Muckle, 2015) may also involve parameters similar to framework weights, but even in this case, the scoring model may implicitly alter these values. In Bayesian adaptive testing (Veldkamp, 2010), the situation is similar: Items may be selected with respect to design constraints such as content weights, but 2PL and 3PL scoring may still implicitly alter the design. With a 1PL model, scoring fidelity is accomplished in a straightforward manner—though this does preclude the validity of other scoring models.

Illustration

For an empirical demonstration of how scoring affects test design, a set of item parameters from the NAEP was used from Grade 4 in 2009. There were 108 items with nonzero values of c , and this set of item parameters was chosen to mimic potential test designs with primarily MC items. The present purpose is to investigate the effects of scoring. This “design” is limited in the sense that the full NAEP item sets include a number of response formats, and there is no attempt to generalize the results below to the full NAEP assessment. Rather, the point is to base the current investigation on realistic item parameters.

In Figure 1, normalized weights implied by the optimal scoring formula are graphed across the range of ability for NAEP mathematics strands. The weights for Number (.40), Measurement (.20), Geometry (.15), Data Analysis (.10), and Algebra (.15) corresponded fairly well to the actual percentages of selected items in these categories at 35%, 24%, 16%, 13%, and 12%, respectively. It can be seen in Figure 1 that in the neighborhood of average ability and above ($\theta > 0$), the normalized optimal weights match the framework weights closely. However, for lower proficiency examinees ($\theta < -1$), the optimal weights begin to shift. This can be seen in the weight for the Number, which decreases, and in the weight for the Geometry, which increases.

To explore conditions that might lead to stronger blueprint-scoring mismatches with respect to guessing, two extreme-case scenarios are examined. The first scenario, as shown in Figure 2, is obtained by sorting items on c parameters from high to low and assigning this ordered list to

Figure 1. Distribution of weights across ability levels for different content strands.

content categories ordered 1 (Number) through 5 (Algebra), with the original number of items in each category. This models a scenario in which items from the Number strand have the most guessing, items from Measurement have the next highest level of guessing, items from Geometry the third, and so on. The second extreme-case, as shown in Figure 3, is obtained by sorting items from low to high on c and assigning this sorted list to the categories order by content. This models a scenario in which items from the Number strand have the least guessing, items from Measurement have the next lowest level of guessing, and so on.

Upon inspecting Figures 1 and 2, the effects of guessing can be readily seen for the Number strand. With a high level of guessing, the weight attributed to this strand is notably in the lower range of ability. The opposite effect can be seen in Figure 3 in which guessing is the lowest for the Number strand. In Figure 2, the Algebra strand has the lowest level of guessing, and its weight shifts upward as ability decreases beginning at $\theta < -1$. However, the shift (from about 15% to 20%) is less than that for Number, mainly because the number of Algebra items is relatively low. As shown in Figure 3, when Algebra has a high degree of guessing, the weight across ability changes very little. Overall, it appears that content weights only begin to shift in the presence of guessing at relatively low levels of ability. These extreme-case scenarios show that it is possible, though perhaps not likely, to have scoring mismatches with content frameworks. This would occur in situations in which there are large differences in c parameters across content categories.

Discussion

The question of individual psychometric fairness is whether it makes sense that not all students receive the same credit for the same correct answer (Chiu & Camilli, 2013). Certainly if a 2PL or 3PL IRT model fits the data, use of this model for scoring may provide a technical advantage in terms of calibration and equating. In turn, this may *benefit* individual test takers in terms of increased reliability. The argument for the 3PL model is that an implicit correction for guessing is applied, which may reduce the impact of unexpected item responses. However, the main topic of this article concerns blueprint fidelity. It was shown that although the 1PL model provided de

Figure 2. Extreme-case scenario with items from the Number strand having the highest degree of guessing.

Figure 3. Extreme-case scenario with items from the Number strand having the lowest degree of guessing.

facto alignment to the blueprint, applications of the 2PL and 3PL model for scoring did not have strong effects on blueprint fidelity with NAEP-like items, but there was some cause for concern at $\theta < -1$. The use of 2PL and 3PL scoring may require examination for each particular use of assessment. The extreme-case scenarios examined suggest scoring mismatch is a possibility.

Other factors can also lead to how items are weighed in the more general sense of item information. At low levels of ability, very difficult items contribute little information, and the opposite is true with easy items for high-ability examinees. This is true for all IRT models considered in this article. Locally best scoring weights are related to sufficient statistics for ability

in the 1PL and 2PL models, but also to item information. For the 3PL model, the information for an item can be factored as,

$$I(\theta) = \frac{a^2(P-c)^2}{(1-c)^2} \frac{Q}{P} \\ = (1-c)w(\theta)[a\Psi(1-\Psi)], \quad (5)$$

where P is defined in Equation 1, Ψ is defined in Equation 2, and item subscripts have been dropped for convenience. Note that the optimal weight $w(\theta)$ is defined in Equation 3, and $a\Psi(1-\Psi)$ is the theoretical 2PL information function. Total information for the 3PL model can therefore be represented as a product of $(1-c)$, the optimal weight, and the theoretical 2PL item information. In other words, the optimal weight contributes to overall item information as well as scoring. For these reasons, the optimal score weight appears to be a useful tool for studying IRT scoring alignment.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The *locally best weight* was defined by Birnbaum was developed for two-point classification problems. More formally, given a value of θ , the problem is to classify $\theta > \theta_1$ or $\theta < \theta_2$ where $\theta_2 > \theta_1$. For the logistic test model, the same weighted-sum formula is obtained in the determination of admissible rules across the range of θ . Note that for the one-parameter logistic (1PL) and two-parameter logistic (2PL) models, w obtains the sufficient statistic for proficiency.
2. See the National Assessment of Educational Progress (NAEP) Technical Documentation, Definition of Composite Scales at https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_determination_composite.aspx

ORCID iD

Gregory Camilli <https://orcid.org/0000-0003-2430-0342>

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.
- Camilli, G., Briggs, D. C., Sloane, F. C., & Chiu, T.-W. (2013). Psychometric perspectives on test fairness: Shrinkage estimation. In K.F. Geisinger (Ed.-in-Chief), B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Assoc. Eds.), *APA handbooks in psychology: APA handbook of testing and assessment in psychology: Volume 3. Testing and assessment in school psychology and education* (pp. 571-589). Washington, DC: American Psychological Association.
- Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, 73, 209-230.

- Chiu, T.-W. (2011). *Correction for guessing in the framework of the 3PL item response theory* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Accession Order No. AAI 3418414)
- Chiu, T.-W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, 37, 58-68.
- Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39, 426-451.
- Muckle, T. (2015). Web-based item development and banking. In S. Lane, M. R. Raymond, M. R., & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 241-258). New York, NY: Routledge.
- Sheng, Y. (2013). An empirical investigation of Bayesian hierarchical modeling with unidimensional IRT models. *Behaviormetrika*, 40, 19-40.
- Van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34, 327-347.
- Veldkamp, B. P. (2010). Bayesian item selection in constrained adaptive testing using shadow tests. *Psicológica*, 31, 149-169.
- Von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives*, 7, 110-114.
- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68, 456-477.
- Yamamoto, K. (1989). *A hybrid model of IRT and latent class models* (ETS Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.