A Clarification on the Response Probability Criterion RP67 for Standard Settings Based on Bookmark and Item Mapping

Huynh Huynh, *University of South Carolina*

By analyzing the Fisher information allotted to the correct response of a Rasch binary item, Huynh (1994) established the response probability criterion .67 (RP67) for standard settings based on bookmarks and item mapping. The purpose of this note is to help clarify the conceptual and psychometric framework of the RP criterion.

Keywords: bookmark standard setting, item mapping, scale anchoring, response probability criterion, RP value

Among standard setting procedures that are based on item mapping, the Bookmark method (Lewis, Mitzel, Green, & Patz, 1999) is well known and widely used in several large-scale assessment programs. Two key components of this method are creation of an ordered item booklet (OIB) and selection of a response probability (RP) value. In the OIB an item is mapped at the point on the achievement continuum (aka construct, proficiency, latent trait, and θ-scale) where the probability of a correct response equals the RP value. In addition, judges in the standard setting meeting are instructed to put a bookmark at the item in the OIB where an examinee at the cut score can answer the item correctly with a probability equal to RP. Other standard setting procedures that are based on item mapping include the item-descriptor (ID) procedure (Ferrara, Perie, & Johnson, 2002) and the one described by Wang (2003) for multiple-choice licensure and certification examinations.

Across the years, RP values that range from .50 to .80 have been used in bookmark standard setting and other applications such as scale anchoring (Beaton & Allen, 1992). At the lower end, for example, CTB/McGraw-Hill carried out the 1993 bookmark standard setting for its Terra-Nova using the RP value of .50. However, for the 1995 and subsequent editions, CTB/McGraw-Hill shifted up the RP value to .67 (or 2/3) (Lewis et al., 1999). On the other, higher end, the National Center for Education Statistics used the RP value of .80 to set the cut scores for the 1992 National Adult Literacy Survey (NALS). In the 2003 and for the National Assessment of Adult Literacy (NAAL), the RP was shifted down to .67 (National Academies of Sciences, 2005). An extensive review of research on RP values and issues on bookmark standard setting and other similar procedures is provided by Karantonis and Sireci (2006) in a previous issue of *Educational Measurement: Issues and Practice*.

With the RP value of .67 (2/3) being widely used in the field, it is important that its conceptual and psychometric bases are clearly explained and articulated. This value was derived for binary items (Huynh, 1994) by maximizing the (psychometric) information carried in the *correct response*. Writers including Karantonis and Sireci (2006) and Cizek, Bunch, and Koons (2004) often use the term "item information" or "test information" rather than the more specific term "*information of the correct response*." The seeming confusion about IRT terminology might make it difficult for psychometricians and practitioners alike to attach an appropriate interpretation to the various RP values. The purpose of this note is to help clarify the conceptual and psychometric basis of the RP value of .67.

Item response theory (IRT; Hambleton & Swaminathan, 1985) models are typically used in creating OIBs. These models include the Rasch, 2PL, and 3PL models. The bookmark procedure typically uses an RP of .67 (or 2/3) for Rasch or 2PL items. Let c represent the pseudo-chance (lower asymptote) parameter of the 3PL model. The resulting corrected-for-chance RP value for a 3PL item is typically chosen to be $(2 + c)/3$. Historically, the RP value of .67 for a Rasch binary item was first proposed by Huynh (1994) at the Joint Conference on Standard Setting for Large-scale Assessments that was jointly sponsored by National Assessment Governing Board and National Center for Education Statistics. Subsequently, Huynh (1995, 1998, 2000a, 2000b) extended his work on RP to 2PL, 3PL, and for partial credit items. Other writers including Kolstad, Cohen, Baldi, Chan, et al. (1998) also conducted studies in this area.

For binary items, the focus of Huynh's work on the RP value has been on *information of the correct response* and not on the (total) item information, which is the typical focus for item information.

Huynh Huynh, Department of Educational Studies, College of Education, University of South Carolina, Columbia, SC 29208; HHuynh@gwm.sc.edu.

The crucial difference between these two types of information apparently has been overlooked by several writers on bookmark standard setting, including Cizek et al. (2004) and Karantonis and Sireci (2006). Cizek et al. (page 38, line 13 from the top) reported that Huynh used "*test information*" and Karantonis and Sireci (page 7, second paragraph) referred to "*item information*." In addition, the last two writers also quoted the writing of Wang (2003) on the RP value of .50 for Rasch items and left an impression that there is a psychometric contradiction between the .67 result in Huynh's and Wang's argument for the RP value of .50. Certainly there are differences among *(total) item information, test information*, and *information of the correct response*. The apparent contradiction is readily resolved by noting that Wang's argument is based on the (total) item information whereas Huynh's work relies on information of the correct response.

The (total) item information for a Rasch and 2PL item is proportional to $p(1-p)$ where p is the probability of the correct response. The item information is maximized when $p = .5$. This occurs at the θ value that is equal to the item difficulty parameter (b). This well-known result has been used as a psychometric justification for mapping the item at the parameter b and for the use of the RP value of .50. Although such mapping has been very useful in test form construction and computer-adaptive testing, there are situations where other types of item information are more appropriate. Huynh (1998) argues that the item location concept does not account for expectations about examinee performance on the item because the item formation encompasses both the incorrect and correct response. In a number of situations, it may be more informative to focus on the *location of the correct response*. This location might serve as a signal that examinees whose estimated proficiency is located at this place would be "expected" to have the skills underlying the item. Huynh further argues that this type of item response interpretation appears to be more *assertive* or more reflective of positive student performance than a neutral statement that an item is located at a given place. For a Rasch or 2PL item, the total item information of $p(1-p)$ is partitioned into two components, one for the incorrect response and the other for the correct response. Because the probability of the correct response is p, the portion of the total item information partitioned to the correct response is taken as $p(1-p)p$. This information is maximized when $p = 2/3$ or .67.

Subsequent work was carried out by Huynh (1995, 1998) for 2PL, 3PL, and partial credit items. A Bayesian analysis of RP values was provided by Huynh (2000a) and a decision-theoretic approach to these values was reported in Huynh (2000b). For the 3PL with c as the pseudo-chance parameter (lower asymptote), for example, the total item information is proportional to $(1-p)\times(p-c)^2/p$ and the information of the correct response is proportional to $(1-p)(p-c)^2$. Maximizing this information yields $p = (2+c)/3$. This result provides the psychometric basis for the corrected-for-chance RP value of $(2+c)/3$, often used in item-mapping standard setting for OIBs with 3PL items.

References

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*, 191–204.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50.

Ferrara, S., Perie, M., & Johnson, E. (2002, April). *Matching the judgmental task with standard setting panelists expertise: The item-descriptor (ID) matching procedure*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer.

Huynh, H. (1994, October). *Some technical aspects in standard setting*. In *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessment Programs* (co-sponsored by National Assessment Governing Board and National Center for Education Statistics), Washington, DC, October 5–7, 1994, pp. 75–91.

Huynh, H. (1995, June). *On score locations of binary and partial credit items and their applications to scale anchoring or criterion-referenced interpretation*. Paper presented at the meeting of the Design and Analysis Committee of National Assessment of Educational Progress, Washington, DC.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23*, 35–56.

Huynh, H. (2000a, April). *On Bayesian rules for selecting 3PL binary items for criterion-referenced interpretation and creating ordered item booklets for bookmark standard setting*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Huynh, H. (2000b, April). *On item mapping and statistical rules for selecting binary items for criterion-referenced interpretation and bookmark standard setting*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting-method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998, May). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard*? Washington, DC: American Institutes for Research.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

National Academies of Sciences (2005). *Measuring literacy: Performance levels for adults, interim report*. Retrieved May 2, 2006 from http://www.nap.edu/books/0309096529/html/.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*, 231–252.