Psychological Test and Assessment Modeling, Volume 53, 2011 (4), 486-498

Investigating the saltus model as a tool for setting standards

Karen Draney[1] *& Minjeong Jeon*[2]

Abstract

The saltus model (Wilson, 1989; Draney, Wilson, Glück, & Spiel, 2008) is a latent variable mixture model originally designed for detecting developmental discontinuities.

One difference between this model and many other latent class models is that a clear ordering of the latent classes is hypothesized, from lowest to highest proficiency. At higher classes, the relative difficulty of performance on associated sets of items decreases compared to lower sets of items and lower groups of students. Such an ordering could potentially prove useful in the process of setting standards.

Although this model has previously been used primarily for the exploration of developmental differences, this paper will investigate the possibility of using the saltus model as a tool to assist in standard setting. Illustrative analysis will be conducted with data that were used in a recent study of the Bookmark method for standard setting, and results will be compared.

Key words: saltus model, latent class models, standard settings

[1] *Correspondence concerning this article should be addressed to:* Karen Draney, PhD, Graduate School of Education, University of California, Berkeley, CA 94720, USA; email: kdraney@berkeley.edu

[2] University of California, Berkeley

Introduction

This study investigates the use of the saltus model (Wilson, 1989; Draney, Wilson, Glück, & Spiel, 2008) as a tool for setting performance standards. This model is a mixture latent variable model originally designed for detecting developmental discontinuities. One difference between this model and many latent class models, or mixture models such as that developed by Rost (1990) is that a clear ordering of the latent classes is hypothesized from lowest to highest proficiency. Higher developmental levels are represented in the saltus model as latent classes with a numerically higher positive saltus parameter. Students in this developmentally higher groups or latent classes are characterized by a lower relative difficulty of correct responses with respect to sets of items in the assessment compared to groups of students at a developmentally lower level. The performance ordering induced and parameterized by numerical quantities in the saltus model could potentially prove useful in the process of setting standards.

The saltus model

The saltus model is based on the assumption that there are H ordered stages in the population of interest. Different sets of items represent each one of these (developmental) stages, such that only persons at or above a stage are fully equipped to answer the items associated with that stage correctly. The saltus model assumes that all persons in stage h answer all items in a manner consistent with membership in that stage. However, persons within a stage may differ in terms of proficiency, much like it is the case in mixture IRT models.

To describe the model, suppose that, as in the partial credit model (Masters, 1982), the random variable $\boldsymbol{X}_{ni}$ indicates the response to item i. Items have $J_i + 1$ possible response alternatives indexed $j = 0,1,...,J_i$. The parameter indicating step j for item i will be indicated by β_{ij}; the vector of all β_{ij} by $\boldsymbol{\beta}$.

In the saltus model, a person is characterized by a proficiency parameter θ_n and an indicator vector for stage membership $\boldsymbol{\phi}_n$. If there are H potential stages, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nH})$, where ϕ_{nh} takes the value of 1 if the examinee n is in stage h and 0 if not. Only one of the ϕ_{nh} is theoretically nonzero. As it is true for θ_n the values of $\boldsymbol{\phi}_n$ are not observable.

Just as persons are associated with one and only one stage, items are associated with one and only one stage. Unlike person stage membership, however, which is unknown and must be estimated, item stage is known a priori, based on the theory that was used to develop the assessment and the items. It will be useful to denote item stage membership by the indicator vector $\boldsymbol{b}_i$. As with $\boldsymbol{\phi}_n$, $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iH})$, where b_{ik} takes the value of 1 if item i belongs to item stage k, and 0 otherwise. The set of all $\boldsymbol{b}_i$ across all items is denoted by $\boldsymbol{b}$.

The equation

$$P\left(X_{nij}=j \mid \theta_n, \phi_{nh}=1, \boldsymbol{\beta}_i, \tau_{hk}\right)=\frac{\exp\sum_{s=0}^{j}\left(\theta_n-\beta_{is}+\tau_{hk}\right)}{\sum_{t=0}^{J_i}\exp\sum_{s=0}^{t}\left(\theta_n-\beta_{is}+\tau_{hk}\right)}, \tag{1}$$

indicates the probability of response j to item i. The saltus parameter τ_{hk} describes the additive effect – positive or negative – for persons in stage h on the item parameters of all items in stage k. In a developmental context, this often takes the form of an increase in probability of success as the person achieves the stage at which an item is located, indicated by $\tau_{hk} > 0$ when $h \geq k$ (although this need not be the case). The saltus parameters can be represented together as an H by H matrix $\boldsymbol{T}$.

The probability that an examinee with stage membership parameter $\boldsymbol{\phi}_n$ and proficiency θ_n will respond in category j to item i is given by:

$$P\left(X_{nij}=j \mid \theta_n, \boldsymbol{\phi}_n, \boldsymbol{\beta}_i, \mathbf{b}_i, \mathbf{T}\right)=\prod_h\prod_k P(X_{nij}=j \mid \theta_n, \phi_{nh}=1, \boldsymbol{\beta}_i, \tau_{hk})^{\phi_{nh}b_{ik}}. \tag{2}$$

Assuming conditional independence, the modeled probability of a response vector is:

$$P\left(\mathbf{X}_n=\mathbf{x}_n \mid \theta_n, \boldsymbol{\phi}_n, \boldsymbol{\beta}_i, \mathbf{b}_i, \mathbf{T}\right)=\prod_h\prod_k\prod_i P(X_{nij}=x_{ij} \mid \theta_n, \phi_{nh}=1, \boldsymbol{\beta}_i, \tau_{hk})^{\phi_{nh}b_{ik}} \tag{3}$$

The model requires a number of constraints on the parameters. For item step parameters, we use two traditional constraints: first, $\beta_{i0} = 0$ for every item, and second, the sum of all the β_{ij} is set equal to zero. Some constraints are also necessary on the saltus parameters. The set of constraints we have chosen is the same as that used by Mislevy and Wilson (1996), and will allow us to interpret the saltus parameters as changes relative to the first (lowest) developmental stage. Two sets of constraints are used. First $\tau_{h1} = 0;$ thus, the difficulty of the first stage of items is held constant for all person groups; changes in the difficulty of items representing higher stages are interpreted with respect to this first stage of items for all person stages. Also $\tau_{1k} = 0;$ thus, items as seen by person stages higher than 1 will be interpreted relative to the difficulty of the items as seen by persons in the lowest developmental stage.

The test of English as a first foreign language

The use of this model will be investigated with data from a German national test of English as a first foreign language. This test was designed based on the German National Educational Standards, and items were developed to conform to the Common European Framework of Reference (CEF) for language learning.

The CEF is described as follows by the European Council:

> Developed through a process of scientific research and wide consultation, this document provides a practical tool for setting clear standards to be attained at successive

> stages of learning and for evaluating outcomes in an internationally comparable manner.
>
> The CEF provides a basis for the mutual recognition of language qualifications, thus facilitating educational and occupational mobility. It is increasingly used in the reform of national curricula and by international consortia for the comparison of language certificates.
>
> The CEF is a document which describes in a comprehensive manner i) the competences necessary for communication, ii) the related knowledge and skills and iii) the situations and domains of communication. The CEF defines levels of attainment in different aspects of its descriptive scheme with illustrative descriptors scale. (Source: www.coe.int/T/DG4/Linguistic/CADRE_EN.asp)

The CEF thus provides a competence model distinguishing relevant categories of communicative competence on six successive levels of proficiency. The levels are A1 (*Breakthrough*) and A2 (*Waystage*), characterizing "basic users", B1 (*Threshold*) and B2 (*Vantage*), characterizing "independent users", and C1 (*Effective Operational Proficiency*) and C2 (*Mastery*) characterizing "proficient users".

German secondary-level students are placed into different educational tracks at the end of a short transition period, which happens at the end of 6th grade in most federal states. The first dominant track is comprised of students in the *Hauptschule*, which is the least academic demanding and shortest track with graduation typically after grade 9. The second dominant track is comprised of students in the *Realschule*, which is the medium-length educational track with graduation typically after grade 10. The third dominant track is comprised of students in the *Gymnasium*, which is the longest educational track with graduation typically after grade 12 or 13 and which is typically the one track that is taken by college-bound students. Apart from these three dominant educational tracks, which are the most frequently occurring in Germany (Cortina et al., 2003), there are also integrative school forms (e.g., the *Gesamtschule*) as well as schools for children with special needs (e.g., *Sonderschulen* or *Förderschulen*) (Harsch & Tiffin-Richards, 2009).

The sampling frame for this study was restricted to students from classes in the 8th, 9th, and 10th grade in the *Hauptschule, Realschule*, *Gymnasium*, and *Gesamtschule;* students from vocational schools and schools for children with special needs were not included. The sampling process for this study was a two-stage stratified cluster sample with sampling proportional to size at each stage. At the beginning, the eligible population of students was divided into two strata for which different test designs and statistical models were fit. The two strata were defined based on the *National Educational Standards for English as a first foreign language* (e.g., KMK, 2003, 2004). They distinguish between students who receive a degree after completing the *Hauptschule* (*Hauptschulabschluss, HSA*) and students who receive a different degree after completing relevant coursework in either a *Realschule,* a *Gymnasium, or a Gesamtschule* after about grade 10 (*Mittlerer Schulabschluss, MSA*) (Harsch & Tiffin-Richards, 2009).

The NES for the first foreign language target the CEF-level A2 for the HSA-track, and B1/B2 for the MSA-track of the German school system (Harsch & Tiffin-Richards, 2009). The MSA stratum will be used for the analyses in this paper.

There were three subscales of the English test, including reading, listening, and writing. Multiple forms of the test were developed comprising a total of several hundred items. A *matrix-sampling design* was used to assign items to students, which is also known as a type of *balanced incomplete block design* in the literature on large-scale educational surveys.

In the summer of 2008, a standard setting study was carried out in which two forms of standard setting were investigated, including the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001; Lewis, Mitzel, & Green, 1996), and the Construct Mapping method (Draney & Wilson, 2010). As part of this study, a subset of items from each of the three subscales was selected (75 for reading, 74 for listening, and 19 for writing). These were selected to meet certain criteria, including representation of all CEF levels and item types, good fit, and a range of item difficulty. Items from one of the three subscales, reading, will be analyzed in the current paper. All items are scored dichotomously.

In the Reading subscale for the MSA stratum, there were no items designed to represent the C2 level, and only a very small percentage of students were believed to be performing even at the C1 level. In addition, as mentioned before, the test given to the MSA stratum targets the B1/B2 levels; A2 items were included in the examinations, but few A1 items were. Analyses for this paper will therefore focus on the A2 and B1, and B2 levels of the CEF. The number of each level of items selected for analysis is shown in Table 1.

Table 1a:
Items per CEF level

CEF Level	Items
A2	56
B1	36
B2	31

Since the assessment was based on a matrix sample of items, the data set contained a large number of responses missing by design: On average, each examinee showed 103 missing responses. To avoid computational difficulty, we selected the 50% of the examinees with the fewest missing responses for analysis. As a result, a total of 905 examinees were included in the final data set.

A set of saltus analyses were conducted on these data, using CEF level to indicate the level to which the item should be assigned (this is in place of what has often been considered a developmental level in past uses of the model; for example such analyses have been done with items representing preoperational, concrete, and formal operational stages of development in Piagetian terms; see, for example, Draney & Wilson, 2007). CEF levels are considered ordered, with language proficiency at B levels indicating higher performance in reading and listening than proficiency at A levels, and A2 and B2 proficiency higher than A1 and B1, respectively. As items were developed with the in-

tention to represent primarily one of these levels, and students are believed to be performing at only one level, this data is deemed appropriate for a saltus analysis.

Two models were fitted to the data. The first compared the A2 to the B items; the second compared the B1 and B2 items. Although we attempted to fit a three-level model comparing all groups of items, the resulting analysis was too unstable to converge; possibly this was because of the sparseness of the data due to the matrix sampling design.

Estimation methods

To estimate parameters for the model, we used a Markov chain Monte Carlo (MCMC) technique. MCMC methods have been applied to estimate complex IRT models (Bolt, Cohen, & Wollack, 2001; Gilks, Richardson & Spiegelhalter, 1996; Patz & Junker, 1999) and found to be particularly useful in estimating mixture distributions (Diebod & Robert, 1994). In this study, the MCMC algorithm as implemented in WinBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003) was used. To obtain full conditional distributions, the following prior distributions were assumed for each model parameter:

$\theta_{vc} \sim Normal(\mu_c, \sigma_{vc})$,

$v \sim Multinomial(1, \phi_c[1,2,...,G])$,

$\phi_c \sim Dirichlet(1_1, 1_2, ..., 1_G)$,

$\beta_i \sim Normal(0,100)$,

$\tau_{ck} \sim Normal(0,100)$.

Hyperparameters for several of the prior distributions were

$\mu_c \sim Normal(\eta_c, 1)$,

$\eta_c \sim Normal(0,100)$,

$\sigma_{vc} \sim Normal(0,1)I(0,)$.

where c is the number of saltus stages, or latent classes.

The $I(0,)$ function in the prior for σ_{vc} ensures that only positive values were drawn from a normal distribution. In addition, to define the metric of ability, μ_c and σ_{vc} were set to zero and one, respectively for Class 1.

To determine the number of initial iterations and burn-in iterations, the Gelman and Rubin statistics (1992) and graphical checking using trace plots were used. As a result, the Markov chains were run for 5,000 iterations past the first 5,000 burn-in iterations. Starting values are needed to define the initial state of the Markov chain. For the mem-

bership propensity ϕc, initial values of $\phi 1, \ldots, \phi G = 1/G$ were used. For all other parameters, starting values were randomly generated within the WINBUGS program. We used three chains with different starting values to ensure convergence of the Markov chain. Parameter values sampled after the burn-in iterations were obtained from a chain that is assumed to have converged to its stationary distribution, which is the posterior distribution. Convergence of the Markov chain was checked by trace plots, Gelman-Rubin statistics, and autocorrelation for the selected parameters.

Estimates of the model parameters were computed from the posterior distribution (Gilks, Richardson & Spiegelhalter, 1996). The mean of the posterior distribution was used as the parameter estimate. The possible problem of label switching in Bayesian estimation of mixture models was also monitored and is discussed in the next section.

Results

For the first of the two analyses, comparing the A2 to the B items (B1 and B2 combined), the average number of missing values per respondent was 103.6. Therefore, persons missing fewer than 103 of the 123 items were used, resulting in a usable N of 905. The MCMC analysis used 10,000 iterations, with a burn-in of 5,000. Three chains were run, from different starting values, with good convergence. The mean and standard deviation of the posterior samples of size 15,000 were computed, and will be reported as the estimate and standard error.

Item difficulties ranged from -2.74 to 3.60, with the mean difficulty constrained to 0. The estimate of the saltus parameter (τ_{22}) was 0.91 with a standard error of 0.13. The mean person proficiency for class A was constrained to 1, with an estimated standard deviation of 0.88 (.05); for class B the estimated mean was 1.33 (0.95) with an estimated standard deviation of 0.79 (.05). The estimated proportions of persons in class 1 and class 2 are each .50. These results are illustrated in Figure 1. The two far left columns of this exhibit show the distributions of proficiency for the two person groups; person group B has a slightly higher average proficiency (although the two groups show significant overlap). The last two sets of columns illustrate the item difficulties as seen by group 1, and as seen by group 2; the τ_{22} parameter has the effect of making the B items quite a bit easier for persons in class 2, such that the two groups of items are similar in overall difficulty for this second group. All of this is as would be predicted by the theory used to design the CEF levels.

However, there is a disappointment in this analysis as well. A probability of belonging to each class was also calculated for each person in the analysis. For these probabilities, the minimum and maximum probability of belonging to each class was .33 and .67, respectively; in other words, no person received a probability greater than .67 of being assigned to either class. This is clearly not high enough to definitively assign a student to a particular class of performance on a high stakes test.

Logits	Person distribution	A2 items		B1 and B2 items	
		Lower group	Upper group	Lower group	Upper group
4.0					
3.8					
3.6			108		
3.4					
3.2					
3.0					
2.8			106		
2.6	S2				108
2.4	|				
2.2	|		105		
2.0	|		94,109,102		
1.8	|		117,103,107,93,123,101		106
1.6	|		113,87,59		
1.4	|		73,98,118,76,72		
1.2	M2	4,1	104	4,1	105
1.0	|	28	61,91,66	28	94,109,102
0.8	S1 |	53	65,92,84,120,60,85,77,100	53	117,103,107,93,123,101
0.6	| |		58,96,88,121,70		113,87,59
0.4	| |		64,97,57		73,98,118,76,72
0.2	| |	6,5	119,86,68	6,5	104
0.0	M1 |	9,2,52	82	9,2,52	61,91,66
-0.2	| S2	12	75,62,69,67,116,90	12	65,92,84,120,60,85,77,100
-0.4	|	3,56,5	81,79,99,83,122,80	3,56,5	58,96,88,121,70
-0.6	|	7,44	115,74	7,44	64,97,57
-0.8	S1	41,31,14,50,24,6,49,30	111,114,95,63	41,31,14,50,24,6,49,30	119,86,68
-1.0		40,17,21,46,8,33,19,51,54,38	89,112	40,17,21,46,8,33,19,51,54,38	82
-1.2		43,11,37,18,25,22	78	43,11,37,18,25,22	75,62,69,67,116,90
-1.4		26,13	110	26,13	81,79,99,83,122,80
-1.6		29,36,55,48,16		29,36,55,48,16	115,74
-1.8		23,32,34,42,39	71	23,32,34,42,39	111,114,95,63
-2.0		45,35,15		45,35,15	89,112
-2.2					78
-2.4					110
-2.6		47		47	
-2.8					71
-3.0					

Figure 1:
Wright map comparing A2 items with B1+B2 items for 2 saltus groups

For the second analysis, comparing the B1 to the B2 items, the mean number of missing values was 58.0. Therefore, persons missing fewer than 58 of the 67 items were used, resulting in a usable N of 759. The MCMC analysis was carried out in the same manner as for the first analysis.

Item difficulties ranged from -2.62 to 3.45, with the mean difficulty constrained to 0. The estimate of the saltus parameter (τ_{22}) was 1.20 (0.11). The mean person proficiency for class A1 was again constrained to 1, with an estimated standard deviation of 1.21 (.06); for class A2 the estimated mean was 0.37 (0.95) with an estimated standard deviation of 1.10 (.06). The average estimated proportions of persons in class 1 and class 2 are each .50. These results are illustrated in Figure 2, which is constructed in the same way as Figure 1. In this figure, as in Figure 1, the τ_{22} parameter has the effect of making the B items quite a bit easier for persons in class 2, such that the two groups of items are similar in average difficulty for this second group, although the spread is considerably greater. In this figure, however, the average proficiency of the second group is lower than that of the first, which would not have been expected had this group been performing better overall on the examination.

This analysis is also disappointing with respect to the probability of belonging to each class each person in the analysis. For these probabilities, the minimum and maximum probability of belonging to each class was also approximately .33 and .67, respectively; in other words, no person received a probability greater than .67 of being assigned to either class. Again, this is not sufficiently accurate to definitively assign a student to a particular class of performance on a high stakes test.

For the current application, the three-class solution has not been achieved, probably due to a computational issue. Specifically, a trapping state occurred during iterations of the Markov chains. The reason may be found in the discussion by Bolt, Cohen, and Wollack (2001, p. 389-90). They state that a trapping state frequently occurs when applying MCMC to mixture models. It occurs when very few or no observations are assigned to a class and the mixing proportion for that class is very close to zero. Trapping states occur also in solutions involving a large number of classes and when very uninformative priors are used for the mixing proportion parameter. Considering the amount of missing responses in our data set, perhaps the three-class solution is too difficult to estimate.

Another problem when using MCMC algorithms with mixture models is called label switching: the classes can exchange identity over the course of the MCMC chain. It is likely to occur when the separation of the classes is low, as is the case in our example. For the two-class solutions, serious label switching did not occur in the current applications.

Logits	Person distribution		Item group 1		Item group 2	
			Class 1	Class 2	Class 1	Class 2
4.0						
3.8						
3.6						
3.4				52		
3.2						
3.0						
2.8						
2.6				50		
2.4						
2.2						52
2.0				49		
1.8				53,46		
1.6		S2		45,38		
1.4		|		67,61,51,47,37		50
1.2	S1	|		57		
1.0	|	|		48,62,42		
0.8	|	|				49
0.6	|	|	31,3,16	44	31,3,16	53,46
0.4	|	|	17,20	64	17,20	45,38
0.2	|	M2	10	65,40	10	67,61,51,47,37
0.0	M1	|	5,21,35	41	5,21,35	57
-0.2	|	|	36,9,28,4,29	63	36,9,28,4,29	48,62,42
-0.4	|	|	2,32,14		2,32,14	
-0.6	|	|	12,8,1	60	12,8,1	44
-0.8	|	|	26,30	66,43	26,30	64
-1.0	|	|	19,11,34,26,30	59	19,11,34,26,30	65,40
-1.2	S1	S2	27,24,6,13	55,58,39	27,24,6,13	41
-1.4			18,25,23	56	18,25,23	63
-1.6			7		7	
-1.8			33	54	33	60
-2.0						66,43
-2.2			22		22	59
-2.4						55,58,39
-2.6			15			56
-2.8						
-3.0						54

Figure 2:
Wright map comparing B1 items with B2 items for 2 saltus groups

Discussion

This set of analyses was carried out as an exploration of the usability of the saltus model as a tool for setting standards. It would appear from the results described in this paper that the saltus model is not particularly useful as a standard setting tool. However, this study had a number of limitations, which may have had a significant effect on the results. These are discussed in the following paragraphs.

Although the Reading items analyzed in this paper were constructed with the CEF levels in mind, it is not clear that the correspondence of items with CEF levels was always completely successful. There was some disagreement between item designers and content experts as to which levels were best represented by items in a particular set of items (Simon, personal communication). Also, expert English teachers involved in a standard setting study (Harsch, Pant, & Köller, 2009) comparing the Bookmark method (Lewis, Mitzel, & Green, 1996) with the Criterion Mapping method (Draney & Wilson, 2010), were quite concerned when examining the empirical difficulties of items. Although panel members were not told the intended CEF level of any given item, they often felt that items should be functioning at a different level of difficulty, given what they perceived the item's CEF level to be (Draney, Kennedy, Moore & Morell, 2009).

Prior research has shown that the saltus model can be applied successfully, with excellent classification results (e.g. most or all examinees receiving a probability of .8 or higher of belonging to a single one of the possible saltus classes), even in situations where there are relatively few items (as low as 3 or 4 representing each class), or relatively few persons in the sample (50 to 100). However, in the majority of these studies, the items were (a) based on a strong theory of cognitive development and (b) developed by the manipulation of a small number of factors, each with only a few possible values. Examples of such items involve such skills as predicting which side of a balance scale would go down based on total weight and distance of the weight from the fulcrum (Wilson, 1989); predicting the "juiciness" of mixtures based on number of cups of water and number of cups of juice (Draney & Wilson, 2007), or evaluating the truth of statements based on one of four types of syllogistic inference (Draney, Wilson, Glück, and Spiel, 2008).

However, the comprehension of a second language, whether reading, listening, or writing, is usually tested with items that vary somewhat in content as well as other features than items used in these previous studies. In addition, the tasks were developed to represent not only the CEF levels, but also the German National Educational Standards, as well as to meet a variety of standards for large-scale assessment development (Harsch & Tiffin-Richards, 2009). Thus, the extent to which results in our study were affected by complexity of the construct measured, the items used to represent that construct, and the degree to which those items represent qualitatively different levels of language proficiency, remains an open question to be answered by future research. Perhaps, if a model like saltus were chosen a priori to be used as the standard setting method, and the test items were developed, pilot tested, and edited with such a model in mind, the degree to which classification would be successful might be increased.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381-409.

Cortina, K. S., Baumert, J., Leschinsky, A., Mayer, K. U., & Trommer, L. (Eds.) (2003*). Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick [The educational system in Germany: A review of its structure and developments]*. Hamburg: Rohwolt.

Draney, K., & Wilson, M. (2010). Selecting cut scores with a composite of item types: The Construct Mapping procedure. In E. V. Smith & G. E. Stone (Eds.), *Criterion-referenced testing: Practice analysis to score reporting using Rasch measurement.* Chicago: JAM Press.

Draney, K., Kennedy, C., Moore, S., & Morell, L. (2009). Procedural Standard-setting Issues. In C. Harsch, H. A. Pant, & O. Köller (Eds.), Calibrating Standards-based Assessment Tasks for English as a First Foreign Language: Standard-setting Procedures in Germany, Vol. II. Final Technical Report. *Institute for Educational Progress (Institut zur Qualitätsentwicklung im Bildungswesen, IQB)*, Berlin, Germany.

Draney, K., & Wilson, M. (2007). Application of the Saltus model to stage-like data: Some applications and current developments. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Application*s. New York: Springer.

Draney, K., Wilson, M., Glück, J., & Spiel, C. (2008). Mixture models in a developmental context. In G. R. Hancock, & K. M. Samuelson (Eds.), *Latent variable mixture models* (pp. 199-216). Charlotte, NC: Information Age Publishing.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457-472.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice.* Washington, DC: Chapman & Hall.

Harsch, C., Pant, H. A., & Köller, O. (2009). Calibrating Standards-based Assessment Tasks for English as a First Foreign Language: Standard-setting Procedures in Germany, Vol. II. Final Technical Report. *Institute for Educational Progress (Institut zur Qualitätsentwicklung im Bildungswesen, IQB)*, Berlin, Germany.

Harsch, C., & Tiffin-Richards, S. P. (2009). Setting Standards in line with the Common European Framework of Reference. In C. Harsch, H. A. Pant, & O. Köller (Eds.), Calibrating Standards-based Assessment Tasks for English as a First Foreign Language: Standard-setting Procedures in Germany, Vol. II. Final Technical Report. *Institute for Educational Progress (Institut zur Qualitätsentwicklung im Bildungswesen, IQB)*, Berlin, Germany.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. Presented paper, Annual CCSSO National Conference on Large Scale Assessment, Phoenix AZ.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41-71

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

Rost, J. (1990). Rasch models in latent class analysis: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). WINBUGS version 1.4. [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.