Title of Manuscript: Structured Constructs Models Based on Change-Point Analysis

List names in a similar format shown below:

HYO JEONG SHIN is Associate Research Scientist, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541; hshin@ets.org. Her primary research interests include psychometric methods and large-scale assessments.

MARK WILSON is Professor, University of California at Berkeley, 4415 Tolman Hall, University of California at Berkeley, Berkeley, CA 94720; markw@berkeley.edu. His primary research interests include psychometrics, assessment and related statistical methods.

IN-HEE CHOI is Associate Research Fellow, Korean Educational Development Institute, 7 Gyohak-Ro, Deoksan-Myeon, Jincheon-Gun, Chungchungbuk-Do, 27873, Korea, ichoi@kedi.re.kr. Her primary research interests include psychometric methods.

Corresponding author's current address and telephone number

Name: Hyo Jeong Shin

Job Title: Associate Research Scientist

Affiliation: Educational Testing Service

Address: 660 Rosedale Road, Princeton, NJ 08541

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/jedm.12146](https://doi.org/10.1111/jedm.12146).

This article is protected by copyright. All rights reserved.

E-mail Address: hshin@ets.org

Office Phone:

FAX:

Areas of Specialization: psychometric methods, rater effects modeling, learning progression modeling, international large-scale assessments

Abstract

This study proposes a structured constructs model (SCM; Wilson, 2009, 2012) to examine measurement in the context of a multidimensional learning progression. The learning progression is assumed to have features that go beyond a typical multidimensional IRT model, in that there are hypothesized to be certain cross-dimensional linkages that correspond to requirements between the levels of the different dimensions. The new model builds on multidimensional item response theory models and change-point analysis to add cut-score and discontinuity parameters that embody these substantive requirements. This modeling strategy allows us to place the examinees in the appropriate learning progression level and simultaneously to model the hypothesized requirement relations. Results from a simulation study indicate that the proposed change-point SCM recovers the generating parameters well. When the hypothesized requirement relations are ignored, the model fit tends to become worse, and the model parameters appear to be more biased. Moreover, the proposed model can be used to find validity evidence to support or disprove initial theoretical hypothesized links in the learning progression through empirical data. We illustrate the technique

with data from an assessment system designed to measure student progress in a middle-school statistics and modeling curriculum.

Keywords: multidimensional item response model, change-point analysis, learning progression, structured constructs model

Structured Constructs Models Based on Change-Point Analysis

A learning progression describes a series of increasingly-sophisticated ways of thinking that students can progress through as they learn and investigate a topic. The learning progression (LP) concept is fundamental to the conceptualization of the Common Core State Standards and Next Generation Science Standards, which aim to bring coherence to curricula, instruction, and assessment (e.g., National Research Council, 2012, 2014; Wilson, 2009, 2012) and which are the basis for many statewide large-scale assessments. Having an LP framework enables us to illustrate and test our theories about the conceptual pathways followed by students as they are learning in a certain domain. Researchers and content experts in mathematics, reading, science, and other fields have been using various measurement models for empirical validation of the underlying LPs. Typically, they describe student achievement in terms of certain levels in the LP (e.g., “proficient,” etc.) by estimating or determining cut scores after applying various types of psychometric models. One example is the Construct Map approach (e.g., Wilson, 2005). In this approach, one continuous latent variable is assumed to underlie an LP (or several if it is multidimensional), ranging from one extreme to another, to assess student development.

One of the complications in real-life situations, however, is that there are often complex multiple constructs incorporating complicated relationships among the constructs, within or across an LP. Figure 1 illustrates an example that served as motivation for the substantive context for the current study. It shows two constructs from the Assessing Data Modeling and Statistical Reasoning Project (ADM; Lehrer, Kim, Ayers, & Wilson, 2014; Lehrer & Wilson, 2011). They are (1)

Conceptions of Statistics (CoS) and (2) Chance (Cha). For the CoS construct, students at the lowest level are capable of describing a distribution in only the most informal of ways, while students at the highest level can investigate and anticipate qualities of a sampling distribution. There are also two qualitative levels of sophistication in-between these two extremes. Similarly, the Chance construct describes students' progression through six distinct levels as they strive to understand probability as a measure of uncertainty. It is important to note that certain educational requirements are hypothesized between levels across the two constructs (for other examples see: Wilson, 2009, 2012). In Figure 1, these are represented by the two curved lines with arrows. Notably, the link from CoS Level 3 to Chance Level 3 represents the hypothesis that CoS Level 3 is a *requirement* for Cha Level 3, meaning a student is more likely to reach Level 3 on the Cha construct if he or she has reached Level 3 on the CoS construct. Likewise, the link from Cha Level 6 to CoS Level 4 represents the hypothesis that Cha Level 6 is a requirement for CoS Level 4.

Figure 1. An example of level-specific relationships in the ADM project using the Conceptions of Statistics (CoS) and Chance (Cha) constructs.

Given the complexity of hypothesized relationships between these two constructs, a “normal” multidimensional latent variable model (Adams, Wilson, & Wang, 1997; Reckase, 2009) will

likely not suffice. With a “normal” multidimensional item response theory model, the extent to which multiple dimensions—in this case, two, Cha and CoS— are correlated can be estimated. However, this approach is not able to incorporate relationships that go beyond the overall correlation of two or more dimensions. The level-specific relations, such as indicated by the arrow from CoS 3 to Cha 3, require the formulation of hypotheses regarding the levels of the LP, in addition to the overall relationship across a set of dimensions. In substantive terms, this can be formulated as a hypothesis that a student should reach a certain level of one dimension (i.e., the required dimension) before that student could reach a certain level of a second dimension (i.e., the target dimension).

The structured constructs model (SCM) was designed for exactly this kind of situation (Wilson, 2009, 2012). This study proposes one way to formulate a measurement model for complicated LPs using the SCM approach by introducing the idea of a change-point analysis into the measurement model. First, the cut scores are directly estimated—this corresponds to the change points that embody where on the latent construct students reach a certain level of the construct. Second, this study explores the application of the change-point SCM to simple two-level cases of the LP (e.g., mastery vs. non-mastery) as well as multiple levels including three-level cases (e.g., Level 1, Level 2, and Level 3), which occur quite frequently in the LP framework. Along the way, we illustrate how this modeling strategy allows us to test different hypotheses about the links between the requirement and target dimensions.

The paper briefly describes previous statistical frameworks for the SCM and presents an SCM based on a change-point analysis. The results of applying the method to simulated data and also to an empirical data set are presented and discussed. We conclude with suggestions for further research.

Structured Constructs Models (SCMs)

The SCM idea provides a theoretical framework of measurement models for complicated LPs in which multiple constructs are involved and the relations among the levels of the constructs are hypothesized in advance (Wilson, 2009). Therefore, the three most important features in SCM modeling¹ are to (1) identify the cut score(s) on the latent requirement construct continuum, (2) classify the examinees to place them into the levels of the requirement and target constructs, and (3) model the links so that the hypothesis on the specific relationships between the requirement levels and target levels can be incorporated. In this section, the SCM approach based on change-point analysis—which satisfies these features—is proposed after reviewing the previous statistical frameworks for the SCM.

The previous statistical frameworks for the SCM can be divided into two approaches: the SCM based on the ordered latent class models and the SCM for the latent continuous variables. First, the latent class analysis primarily aims to identify subgroups of examinees by relating the observed item responses to a set of *categorical* latent variables. After identifying the latent groups, each examinee is classified and placed into one of the levels (classes) according to his or her latent class membership (Maris, 1999). Considering the ascending structure of the LPs from lower levels to higher levels, the *ordered* latent class model is considered to be a way to describe LPs. Diakow, Irribara, and Wilson (2011) extended the ordered latent class models (Croon, 1990) to two latent variables and applied it as the SCM approach. Specifically, the crossed levels of the two constructs

¹ In the context of an LP, researchers and content experts describe student achievement in terms of levels in the developed LP (i.e., the level-referenced) in practice by classifying the students based on the cut scores. The selection of cut scores is also an interesting and important topic in the standard-setting literature, and estimating or determining cut scores after applying some type of psychometric models is one of the widely used approaches (e.g., Draney & Wilson, 2010; Shin, Choi, & Draney, 2012). In particular, estimating cut scores has advantages over using ad hoc or predetermined cut scores in terms of obtaining data-driven estimates. Several studies have investigated the use of psychometric models as a tool to assist in standard setting or determining cut scores (e.g., Draney & Jeon, 2011).

were considered to be classes, and the links between the constructs were expressed as the joint probabilities of class membership. In their approach, the levels were identified with a latent discrete variable, and the underlying person proficiency was not assumed to be a latent continuum. Instead, examinees within the same level (i.e., latent class) were assumed to be identical in terms of their probabilities of providing particular responses to the items.

However, it is plausible that students within a given level of the dimension may vary in their overall proficiency (e.g., compare with a between-category qualitative difference and within-category heterogeneity in De Boeck, Wilson, & Acton [2005]). This possibility leads to the consideration of a second approach, where within-category variation is modeled. Working with this second approach, Choi (2013) proposed an SCM for the latent continuous variables for a simple case in which there is a single connection between two constructs and the item responses are dichotomous. Each construct is assumed to be continuous and to be composed of two levels (e.g., mastery versus non-mastery, or proficiency versus non-proficiency) as illustrated in Figure 2 (left panel). In the figure, heterogeneity in construct level is represented with a normal distribution for each construct, although normality of the distributions is not required. The hypothesized link assumes that a person p is more likely to be classified at the proficient level on the target construct if that person is considered to be at the proficient level on the requirement construct. An arrow from the proficient level in the requirement construct to the proficient level in the target construct represents this hypothesized link between the levels. The construct from which the arrow starts is referred to as the “requirement construct,” and the construct at which the arrow terminates is referred to as the “target construct,” and ϑ_{pR} and ϑ_{pT} denote the continuous latent variable of person p ’s proficiency in the requirement and target constructs, respectively. There is no specific hypothesis for the relationship between nonproficient levels in the requirement construct and the target construct.

Figure 2. Hypothesized links between two constructs for a two-level case (left) and three-level case (right).

The link between the levels of the two constructs can be modeled by a discontinuity parameter, which is similar to the saltus parameter (Draney & Wilson, 2007; Wilson, 1989). In the saltus model, the estimated classes of the persons represent different development stages, and the groups of items are specified to allow persons at or above the developmental stage to have the advantage of being able to give better answers in that stage. The saltus parameter quantifies these discontinuities as additive effects on the item parameters of all items in item group when people in group respond to those items. Similarly, in the SCM approach formulated in Choi (2013), the hypothesized links between the two constructs are assumed to induce the discontinuities, and this study is also built on that idea.

SCM Based on Change-Point Analysis

Change-point analysis has been applied in various fields to detect the number of change (discontinuity) points and their locations (e.g., Chen & Gupta, 2011; Chib, 1998; Hall, Lipton, Sliwinski, & Stewart, 2000; Hinkley, 1970; Western & Kleykamp, 2004). For an example of time-series data, a simple regression model that embodies the change-point hypothesis for a continuous dependent variable is written as:

$$\hat{y}_t = \beta_0 + \beta_1 x_t + \beta_2 I_t(\pi) + \beta_3 I_t(\pi) x_t, \quad t = 1, \dots, T, \quad (1)$$

where for the change-point indicator $I_t(\pi) = 0$ for $t < \pi$, $I_t(\pi) = 1$ for $t \geq \pi$, y_t are the observations of the dependent variable at time point t , x is a covariate, and β s are the regression coefficients. Apart from the specified covariates, the primary purpose of this model is to estimate the time point at which the change occurs (π) as well as the mean change in the dependent variable at that change point (β_2).

In this study, the focus is quite similar to this simple form of change-point analysis: the estimation of the change point (π) and the average magnitude of abrupt change at the change point (β_2). However, rather than time t , we are interested in a change point in latent ability (ϑ). Specifically, we aim to estimate the change point that corresponds to the cut scores in the requirement construct, which leads to the classification of the examinees according to the cut scores. We also aim to estimate the magnitude of abrupt change at the change point, which corresponds to the strength of the discontinuity at the cut scores that incorporate the mean change in the latent ability distribution in the target construct according to the hypothesized links. However, our case is more complicated in terms of the relationship between the unit of analysis and the scale of the model parameters. In many applications of change-point analysis, the dependent variable (e.g., wage growth) is measured or adjusted in the unit of the change point (i.e., time, for example in years). On the other hand, our unit of analysis (e.g., item-level response data) is different from the scale of the change points, which will be in logit units. We use the item-level response data, which are commonly dichotomously or polytomously scored, and we aim to estimate both the change point and the discontinuity on the continuous scale along with the item parameters. Because of the properties of the Rasch model family we use in this study, the item parameters and the person latent abilities were scaled on the same logit scale (Rasch, 1960; von Davier & Carstensen, 2007). This allows us to estimate the cut scores along with the item parameters and to formulate the

discontinuity parameters at the estimated cut scores on the same scale as the latent abilities of persons.²

Based on the above similarities and complexities, we refer to the proposed model as a change-point structured constructs model (change-point SCM). In this section, the change-point SCM is considered in two forms (cf. Figure 2): (a) when there are two levels for each construct (the two-level case, e.g., proficient vs. non-proficient), and (b) when there are multiple levels for each construct (the multiple-level case, e.g., Level 1, Level 2, and Level 3). In both cases, each item is designed to measure only one construct as in the between-item multidimensional item response model (Adams, Wilson, & Wang, 1997).

Learning Progressions with Two Levels (the Two-Level Case)

We propose the formulation that the cut score is determined by the overall item difficulty in the requirement construct. In the original Rasch model (Rasch, 1960), item parameters (θ_{iR}), namely, item difficulties, are estimated for individual items. Alternatively, in this study, the item parameters are reparameterized as the combination of the average difficulty across all items and the deviation from the average difficulty for each item. We define the average item difficulty across all items, which corresponds to the intercept in the logistic regression, as the cut score. Thus, the reparameterized Rasch model is written as follows and used for the requirement construct:

$$\Pr(Y_{piR} = 1 | \theta_{pR}) = \frac{\exp(\theta_{pR} - (\gamma + \tau_{iR}))}{1 + \exp(\theta_{pR} - (\gamma + \tau_{iR}))}, \quad (2)$$

where γ represents the intercept that corresponds to the average difficulty of all items. This estimate serves as the directly estimated cut score, and τ_{iR} ($i = 1, \dots, I_R$) indicates the deviation from this average difficulty for each item i in the requirement construct. This average difficulty is based on

² On the other hand, this implies that the proposed model is more difficult to employ in the context of more general item response theory models, such as two- or three-parameter logistic models.

the reparameterization of the Rasch model and is not part of the constraint for removing the scale indeterminacy. Of course, using the overall item difficulty as the cut score seems too dependent on the items used in the tests. However, in LP research, items used to assess the constructs are designed and selected very carefully by an iterative validation process (e.g., Jin & Anderson, 2012; Jin, Shin, Johnson, Kim, & Anderson, 2015; Lehrer & Wilson, 2011). Given that the quality of the items meets a satisfactory level of testing, using the estimated cut score γ in the requirement construct, examinees are separated into the proficient and nonproficient levels on the requirement construct. Given that the estimated cut score (γ) corresponds to the change point (π) in Equation 1, the probability equation for the target construct is now written as:

$$\Pr(Y_{piT} = 1 | \theta_{pR}, \theta_{pT}) = \frac{\exp(\theta_{pT} + \lambda I(\theta_{pR} \geq \gamma) - \beta_{iT})}{1 + \exp(\theta_{pT} + \lambda I(\theta_{pR} \geq \gamma) - \beta_{iT})} \quad (3)$$

where I is the indicator function that returns 1 when $\theta_{pR} \geq \gamma$ and it is 0 otherwise, and λ represents the discontinuity parameter according to the level in the requirement construct. In particular, for the persons at the proficient level on the requirement construct, if $\theta_{pR} \geq \gamma$, the probability is augmented by λ , which represents the size of the jump at the estimated cut score γ . In the target construct, β_{iT} ($i = 1, \dots, I_T$) indicates individual item difficulties.

Equation 2 for the requirement construct and Equation 3 for the target construct constitute the change-point SCM for the dichotomous data and LPs with two levels (two-level case). Note that a discontinuity parameter (λ) for those who had higher proficiencies than the estimated cut score in the requirement construct was estimated. This corresponds to the magnitude of mean change (β_2) in Equation 1. In other words, λ represents the size of the jump to be estimated as the magnitude of the average increase on the logit scale for those who exceeded the cut score in the requirement construct. If the discontinuity parameter (λ) in the target construct is assumed to be zero while the cut score (γ) in the requirement construct is still estimated, we call it the reparameterized

multidimensional Rasch model (RMRM),³ and it is used as the baseline model for the two-level case to investigate the consequences of ignoring the discontinuity parameters.

Learning progressions with multiple levels (multiple-level case)

In real assessments, there is a greater likelihood of LPs having more than two levels. For example, when there are three levels in the LPs (e.g., Level 1, Level 2, Level 3), there can possibly be two hypothesized links between the two constructs as illustrated in Figure 2 (right panel). We can assume, for example, that if person p is considered to be at Level 2 in the requirement construct, he or she is more likely to reach Level 2 on the target construct, and that the equivalent holds for the persons at Level 3.⁴ Likewise, these hypothesized links between the two constructs are expressed as arrows from Level 2 in the requirement construct to Level 2 in the target construct, and the same is true for Level 3 in the requirement construct and Level 3 in the target construct.

When the LPs consist of three levels for each construct and the items are designed to have three score categories, a similar concept of reparameterization can be applied to the partial credit model (PCM), which is a family of the Rasch model for the polytomous data (Masters, 1982).

Originally, as written in Equation 4, the PCM specifies the probability of responding in the j th score category of item i for person p as a function of the person ability ϑ_p and item-specific step difficulty parameters δ_{ij} ($j > 0$).

$$\Pr(Y_{pi} = j | \theta_p) = \frac{\exp[\sum_{j=0}^j (\theta_p - \delta_{ij})]}{\sum_{h=0}^{m_i} \exp[\sum_{j=0}^h (\theta_p - \delta_{ij})]} \quad j = 0, 1, \dots, m_i \quad (4)$$

³ In the context of the multidimensional assessments, the RMRM can be useful if one is interested in estimating the cut score in one dimension without a particular hypothesis on the specific relationships between the levels. For example, one can use the RMRM to estimate the cut score based on the average of item difficulties and to classify examinees based on that cut score. In this case, one's interest lies in classifying examinees among two levels of LP instead of estimating individual item difficulties in the multidimensional context.

⁴ Of course, it might be less symmetric, with required Level 2 affecting target Level 3, etc.

Alternatively, the step parameters δ_{ij} for individual items can be reparameterized as the combination of the overall step difficulties ($\gamma_j, j>0$) and the deviations from the overall step difficulties for each item ($\tau_{ij}, j>0$). For example, if there are three score categories ($j=0,1,2$), two overall step difficulties γ_1 and γ_2 will be estimated. These overall step difficulties can be considered as the estimated cut scores between levels 1 and 2 and between levels 2 and 3, respectively. In other words, γ_1 represents the overall step difficulty to advance from Level 1 to Level 2, and γ_2 represents the overall step difficulty to progress from Level 2 to Level 3. Then, τ_{i1} ($i = 1, \dots, I$) indicates the deviation from the first cut score for each item i , and τ_{i2} ($i = 1, \dots, I$) indicates the deviation from the second cut score for each item i . A similar idea was explored in Diakow, Irribara, and Wilson (2013) to estimate the level boundaries and represent the construct levels. The probability equation for the requirement construct is then written as:

$$\Pr(Y_{pRi} = j | \theta_{pR}) = \frac{\exp[\sum_{j=0}^j (\theta_{pR} - (\gamma_j + \tau_{ijR}))]}{\sum_{h=0}^{m_i} \exp[\sum_{j=0}^h (\theta_{pR} - (\gamma_j + \tau_{ijR}))]}, \quad j = 0, 1, \dots, m_i. \quad (5)$$

Based on the estimated cut scores γ_1 and γ_2 , examinees are separated into three levels of performance on the requirement construct. Then, the target construct ϑ_{pT} as well as the requirement construct, ϑ_{pR} , determine the probability of getting score j on the items in the target construct. By incorporating the discontinuity parameter for the hypothesized link, the probability of person p receiving the score category j on item i in the target construct is expressed as:

$$\Pr(Y_{pTi} = j | \theta_{pR}, \theta_{pT}) = \frac{\exp[\sum_{j=0}^j (\theta_{pT} - \delta_{ijT} + \lambda_{.j} I(\theta_{pR} \geq \gamma_{.j}))]}{\sum_{h=0}^{m_i} \exp[\sum_{j=0}^h (\theta_{pT} - \delta_{ijT} + \lambda_{.j} I(\theta_{pR} \geq \gamma_{.j}))]}, \quad j = 0, 1, \dots, m_i. \quad (6)$$

Similar to the example for the two-level case, I is the indicator function that returns 1 when $\vartheta_{pR} \geq \gamma_j$; it is 0 otherwise. For those at Level 2 of the requirement construct, if $\gamma_2 > \vartheta_{pR} \geq \gamma_1$, the probability is augmented by λ_1 , which represents the size of the jump at the estimated cut score γ_1 . In the same manner, for those at Level 3 of the requirement construct, if $\vartheta_{pR} \geq \gamma_2$, the probability in

the target construct is increased as much as λ_2 . At each estimated cut score, the advantage doesn't have to be additive or the same but can have different magnitudes. This is shown by the two discontinuity parameters in Equation 6. However, it is possible to impose various constraints on the two discontinuity parameters, such as an equivalence assumption ($\lambda_1 = \lambda_2$), based on the theoretical expectations on the magnitude of discontinuity. This possibility will be examined with two empirical examples as a method of empirical validation of the hypothesized links. In the target construct, analogous to the standard PCM, the item parameter (δ_{ijT}) can be decomposed into the overall item difficulty for i (θ_{iT}) and item-specific step difficulties (τ_{ijT}).

Equation 5 for the requirement construct and Equation 6 for the target construct constitute the change-point SCM for the multiple-level case with polytomous data. This formulation can be extended straightforwardly to situations with more than three levels of performance in the LP by increasing the number of cut scores and the discontinuity parameters. If it is assumed that there is no discontinuity in the target construct ($\lambda_j = 0$) while estimating the cut scores (γ_j) is still of interest, we call it the reparameterized multidimensional PCM (RMPCM). Similar to the RMRM, the RMPCM is used as the baseline model for the three-level case to investigate the consequences of ignoring the discontinuity parameters.

In order to help understand the proposed model, an example of change-point SCM for the three-level case is provided in Figure 3.

Figure 3. An example of change-point SCM for the three-level case.

Figure 3 displays the proficiency distributions in the requirement construct and the target construct based on the change-point SCM for the example given in Figure 2 (right panel). There is a single normal distribution in the requirement construct, and there are three distributions in the target construct for each group of persons at Level 1 (black line), Level 2 (red line), and Level 3 (blue line).

As formulated in Equation 5, the examinees are divided into three groups based on the estimated cut scores γ_1 and γ_2 . In particular, Level 2 examinees in the requirement construct are more likely to achieve Level 2 in the target construct, and Level 3 examinees in the requirement construct are more likely to reach Level 3 in the target construct. The possibility of different advantages for each group is shown by the two discontinuity parameters in Equation 6, λ_1 and λ_2 , with each increasing the probability of getting better scores on the items in the target construct. Similarly, according to the hypothesized links, the discontinuity parameters are likely to have positive values, although negative discontinuity parameters can occur, at least in the mathematical model. In addition, according to the constraints on the discontinuity parameters, the distribution of the target construct

would show different patterns. As will be illustrated in the empirical example, if one link is not valid and the expected advantage does not exist (for example, $\lambda_1=0$), only two distributions on the target construct—a combination of the black line and the blue line—would be presented. As above, a less symmetric hypothesis may be postulated, based on the substantive background.

Identification

As in the multidimensional Rasch models and PCM, ϑ_{pR} and ϑ_{pT} are assumed to follow a bivariate normal distribution: $\begin{bmatrix} \theta_{pR} \\ \theta_{pT} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_R \\ \mu_T \end{bmatrix}, \begin{bmatrix} \sigma_R^2 & \sigma_{RT} \\ \sigma_{RT} & \sigma_T^2 \end{bmatrix}\right)$. For identification purposes, the means of the bivariate normal distributions are set to $\mu_R = \mu_T = 0$, and the sum of the item difficulties in the target construct θ_{iT} ($i = 1, \dots, I_T$) is constrained to be zero (although the difficulties of the requirement construct are not constrained). The cut-score parameter, the discontinuity parameter, and the deviations from the cut score for each item i in the requirement construct are freely estimated. For the multiple-level case, extra constraints need to be imposed: A higher cut-score parameter is constrained to be the negative of another cut-score parameter ($\gamma_2 = -\gamma_1$)⁵, the sum of the step difficulties within each item i (τ_{ijt}) in the target construct is constrained to be zero, and the sum of the deviations τ_{ijR} ($i = 1, \dots, I_R$) within each item in the requirement construct is also constrained to be zero.

Estimation

A Markov chain Monte Carlo (MCMC) estimation was selected to estimate the parameters of the proposed change-point SCMs. WinBUGS 1.4.3 (Lunn, Thomas, Best, & Spiegelhalter, 2000) software can be used for this purpose. In order to implement the MCMC algorithm using WinBUGS, the prior distributions must be specified for all parameters, which include the cut-score parameter,

⁵ This constraint on the cut-score parameter may be too strong for the three-level case where two cut score parameters exist. The constraint on cut-score parameter can be understood in line with the constraint imposed on the step difficulties (i.e., the sum of the step difficulties within each item i are constrained to be zero) in the PCM.

the discontinuity parameter, the item parameters, and the variance-covariance of the person proficiencies. Although each parameter can have different prior distributions, this study limits its scope to the use of simple and straightforward distributions, such as the conjugate prior distributions that make the posterior distribution belong to the same family. Following the conventions of Bayesian item response modeling (Curtis, 2010), a normal distribution was used for the fixed-effects parameters, including item parameters and discontinuity parameters. However, for the cut-score parameter, a mildly informative prior, which is a uniform distribution, was specified following the example of Western and Kleykamp (2004). The range was set from -6 to 6 based on the typical range of item and person estimates on the logit scale. An inverse-Wishart distribution, which is a conjugate prior of the variance and covariance of the multivariate normal distribution, was specified for the variance and covariance matrix of the latent ability variables in the two dimensions (Gelman et al., 2013). For the model selection, the Deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) was used. A sample WinBUGS code for two-level case simulation study is given in the Appendix.

Simulation Study

Two sets of simulation studies were designed and conducted to: 1) evaluate the recovery of the parameters of the proposed change-point SCM, and 2) evaluate the importance of estimating the cut scores from the data.

The two-level case data were generated using equations 2 and 3, and the three-level case data were generated using equations 5 and 6. For the data generation, the R software (R Development Core Team, 2013) was used. As illustrated in the model formulations, the requirement and target constructs were specified, and examinees were separated into one of the two levels or one of the three levels based on the fixed cut score(s) in the requirement construct.

The number of examinees was set to 1,000, and the latent abilities of the examinees in the two

constructs were assumed to follow a bivariate normal distribution: $\begin{bmatrix} \theta_{pR} \\ \theta_{pT} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$.

The number of items in each construct was 30 ($I_R = I_T = 30$), and the item parameters (i.e., item/step difficulties and deviations from the cut score(s)) were generated from a standard normal

distribution. Specifically, for the two-level case, the mean of the item difficulties in the target

construct was constrained to be zero ($\beta_{30T} = -\sum_{i=1}^{29} \beta_{iT}$), while there was no constraint for the

deviations from the cut score in the requirement construct (τ_{iR}). Similarly, for the three-level case,

the mean of the item difficulties in the target construct ($\beta_{30T} = -\sum_{i=1}^{29} \beta_{iT}$) and the mean of the

deviations from the cut scores in the requirement construct were constrained to be zero ($\tau_{I1R} =$

$-\sum_{i=1}^{29} \tau_{i1R}$). In addition, for both constructs of the three-level case, the second step difficulty in the

target construct was set to be the negative of the first step difficulty ($\tau_{i1R} = -\tau_{i2R}$, $\tau_{i1T} = -\tau_{i2T}$) so that

the step difficulties within each item were summed to be zero. More importantly, for the two-level

case, the cut-score parameter was specified as $\gamma = 0$ and the discontinuity parameter was specified

as $\lambda = 1$. The positive discontinuity parameter can be considered to indicate an advantage for

examinees who were classified as being at the proficient level in the requirement construct. For the

three-level case, the cut-score parameters were specified as $\gamma_1 = -0.5$ and $\gamma_2 = 0.5$, with

corresponding discontinuity parameters specified as $\lambda_1 = \lambda_2 = 1$.

After the data were generated, each dataset was analyzed using the proposed change-point

SCM to assess the recovery of the parameter. For all models, three chains with dispersed starting

values were run, and the convergence of the chains was determined with the $\hat{R}$ index with a critical

value of 1.1 (Gelman & Rubin, 1992). In order to avoid high autocorrelations between the sampling

distributions, the thinning interval was increased by as much as 3, and the posterior mean estimates

were obtained from 25,000 iterations of post-burn-in after 25,000 iterations of burn-in. A total of 30

replicates were made for each two-level and three-level case, and the bias and root mean square errors (RMSE) were calculated.

In addition, in an effort to validate the estimation of the cut-score parameter, eight incorrect values of the cut score including the generating value ($\gamma = -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0$) were selected and fixed during the analysis instead of being estimated. For this analysis, one randomly chosen simulated dataset among 30 replicates was used. After fitting the change-point SCM with each fixed cut score, the model fit statistics using DIC and the model parameter estimates were compared.

Results

Recovery of the Model Parameters

Table 1 provides the recovery results of the change-point SCM compared with the baseline models, which are the RMRM for the two-level case and the RMPCM for the three-level case. For the item parameters, including the overall item difficulties in the target construct (β_{iT}), the deviations from the cut scores in the requirement construct (τ_{iR}, τ_{i1R}), and the item-specific step difficulties in the target construct (τ_{i1T}), the averages and the standard deviations of the estimated bias and RMSE values of 60 items are presented.

In general, the variance of the requirement construct (σ_R^2) and the overall item difficulties in the target construct (β_{iT}) were recovered well in both the baseline models (RMRM, RMPCM) and the change-point SCMs for the two- and three-level cases. The bias and the RMSE values of those parameters did not differ significantly; however, the change-point SCM that incorporated the discontinuity parameter yielded a much smaller bias and RMSE values with respect to other parameters, suggesting that the parameters appeared to be recovered better by the generating change-point SCMs.

In a detailed look, for the two-level case, when the change-point SCM was applied, the bias ranged from -0.047 to 0.042, and the RMSE values ranged from 0.021 to 0.087. Even though it was small, among the model parameters, the negative bias was observed from the cut score (γ) and the discontinuity parameter (λ) as -0.047 and -0.011, respectively. For the three-level case change-point SCM, the bias ranged from -0.034 to 0.032, and the RMSE values ranged from 0.014 to 0.090. Overall, these results suggest that the proposed change-point SCM appeared to recover the generating values fairly well.

When the discontinuity parameters were ignored as in the baseline RMRM and RMPCM, the cut-score parameter (γ) was considerably underestimated for the two-level case but recovered quite well for the three-level case. This suggests that more examinees are likely to be classified at the proficient level, particularly for the two-level case, if the hypothesized link is not correctly specified in the model. More importantly, consistent with Choi's (2013) findings, the impacts of neglecting discontinuity parameters were most obvious on the variance of the latent ability in the target construct (σ^2_T) and in the covariance between the two latent constructs (σ_{RT}). As shown in Table 1, the magnitude of the bias of σ^2_T was very large, 0.907 for the two-level case and 0.545 for the three-level case, suggesting that the variance of the target construct (σ_{RT}) was substantially overestimated compared with the corresponding true value, and the correlation between the two constructs (ρ_{RT}) was also overestimated. These results indicate that failing to incorporate the discontinuity parameters can be misleading if there are hypothetical relations in the LP, resulting in more variability in the target construct variance and a higher association between the latent variables in the two constructs than there actually is.

Table 1

Bias and RMSE in the Baseline Models and the Change-Point SCM

	Two-Level Case	Three-Level Case
--	----------------	------------------

		Bias	RMSE			Bias	RMSE
Baseline models (RMRM, RMPCM)	σ_R^2	0.051	0.056	σ_R^2	0.007	0.026	
	σ_T^2	0.907	0.909	σ_T^2	0.545	0.546	
	σ_{RT}^2	0.540	0.540	σ_{RT}^2	0.266	0.267	
	ρ_{RT}	0.249	0.249	ρ_{RT}	0.117	0.117	
	γ	-0.476	0.477	γ_1	-0.001	0.014	
				γ_2	0.001	0.014	
	β_{iT}^*	0.000	0.071	β_{iT}^*	0.000	0.060	
		(0.013)	(0.008)		(0.026)	(0.017)	
	τ_{iR}^*	0.218	0.229	τ_{i1R}^*	0.000	0.074	
		(0.012)	(0.012)		(0.015)	(0.015)	
Change- point SCM				τ_{i1T}^*	-0.092	0.126	
					(0.037)	(0.035)	
	σ_R^2	0.005	0.024	σ_R^2	0.010	0.028	
	σ_T^2	0.013	0.043	σ_T^2	0.024	0.043	
	σ_{RT}^2	0.004	0.024	σ_{RT}^2	0.017	0.029	
	ρ_{RT}	0.000	0.021	ρ_{RT}	0.008	0.015	
	λ	-0.011	0.042	λ_1	0.032	0.086	
				λ_2	-0.034	0.062	
	γ	-0.047	0.057	γ_1	-0.002	0.014	
				γ_2	0.002	0.014	
	β_{iT}^*	0.000	0.071	β_{iT}^*	0.000	0.057	
		(0.013)	(0.008)		(0.009)	(0.014)	
	τ_{iR}^*	0.042	0.087	τ_{i1R}^*	0.000	0.074	
		(0.012)	(0.012)		(0.015)	(0.015)	
				τ_{i1T}^*	0.014	0.090	
					(0.015)	(0.025)	

* The averages of the bias and RMSE values of the estimated item parameters are presented, and the numbers in the parentheses correspond to the standard deviations of the bias and the RMSE.

Validation of the Cut Scores

In an effort to validate the estimation of the cut-score parameter, eight incorrect values of the cut score and the generating value ($\gamma = -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0$) were selected and anchored during the analysis for one randomly chosen simulated dataset. There was a clearly decreasing pattern in the DIC values as the cut score moved closer to the generating value, which is zero, and the change-point SCM showed the lowest DIC value.

Figure 4 presents the resulting MCMC estimates for the two-level case. Dotted lines indicate the generating values used for this particular dataset with an anchored value. For comparison

purposes, the empty dot, which is the result from the change-point SCM for the same dataset, is presented on the cut-score axis in the same figure. For the discontinuity parameter of the two-level case, the generating value was 1.0. As expected, the change-point SCM and the cut score of zero yielded the closest estimates to the correct value, while the discontinuity parameter estimate tended to be underestimated when the incorrectly fixed cut score was far from the generating value. A similar pattern is observed in the model parameters of the bivariate proficiency distribution.⁶

For the three-level case, a very similar pattern was observed, as illustrated in Figure 5. One difference from the two-level case is that there were two cut scores for the three-level case, and the generating values were -0.5 for the first cut score and 0.5 for the second score. Because the second cut score was constrained to be the negative of the first cut score, the generating value used in this figure is the one for the first cut score (-0.5).

Among the various incorrect cut-score values, one interesting exception is when the cut score was fixed to zero, which virtually leads to separating examinees into two groups instead of three. In this case, at the cut score of zero, there is a single discontinuity parameter instead of two discontinuity parameters. Similar to the two-level case, the estimated DIC value was the smallest for the change-point SCM and when the cut score was fixed to the generating value, which was -0.5. Regarding the discontinuity parameter of the three-level case, the dots represent the discontinuity parameter at the first cut score (λ_1) and the triangles represent the discontinuity parameter at the second cut score (λ_2). The estimates from the change-point SCM and anchoring the cut score to the

⁶ For this particular simulated dataset, the variance of the requirement construct distribution was 0.97, which is closest to the results from the change-point SCM, and when the cut score was fixed as zero or -0.5. The variance of the requirement construct tended to be underestimated when the cut score was set smaller than zero, while it was overestimated when the cut score was set larger than zero. For the variance of the target construct, estimates from the change-point SCM and the cut score of zero or 0.5 showed the nearest result to the generating value, and they were overestimated under the incorrect cut scores. The estimated correlation between the two constructs from the change-point SCM was the nearest to the generating value, while the correlation under the incorrect cut scores tended to be overestimated.

correct value provided similar results, which are the closest to the generating value of 1. It is clear that the discontinuity parameter was underestimated when the cut score was set to incorrect values, and it appears to be severely underestimated when the first cut score was as anchored at the positive value. This results in peculiar cut scores that have reversed signs from our expectation, for example, 0.5 for the first transition and -0.5 for the second transition.

There were also radical differences in the variance and covariance of the person proficiencies for the three-level case. For the requirement construct, the variance tended to be overestimated when the first cut score was smaller than the generating value, while the variance tended to be underestimated when the first cut score was larger than the generating value; however, the variance in the target construct was likely to always be overestimated under the incorrect cut scores. These patterns in the variance estimates resulted in an overestimated covariance and correlation between the two constructs under the incorrect cut scores. The overestimation became much worse when the first cut score (generating value = -0.5) was set to be positive.

Overall, for both the two- and three-level cases, the results of the validation of the cut scores showed that the results from the change-point SCM were very close to when the cut score was fixed to the generating value. When the fixed cut score became further away from the generating value, the model fit became worse, and the model parameters appeared to be more biased.

Figure 4. Estimated model parameters for each cut score (two-level case).

Figure 5. Estimated model parameters for each first cut score (three-level case).

Empirical Examples

Two empirical datasets, one with the two-level case and another with the three-level case, are used for illustration of the use of the change-point SCMs in real-world LP research. In particular, how the proposed model can provide empirical validity evidence for hypothesized links in the LP is discussed.

Data Sources

First, as illustrated in Figure 1, a subset of the ADM project data, which consisted of responses to the items of the two selected constructs, the Concept of Statistics (CoS) and the Chance

constructs, was used. The ADM project data was analyzed in the two-level case using the dichotomized responses (proficient versus non-proficient). The original polytomous responses on the same items were dichotomized (i.e., CoS 1,2 $\rightarrow$ non-proficient (score 0); CoS 3,4 $\rightarrow$ proficient (score 1); Cha 1,2 $\rightarrow$ non-proficient (score 0); Cha 3,4,5,6 $\rightarrow$ proficient (score 1)) so that being proficient in the CoS construct was considered to be the requirement to reach the proficient level in the Chance construct, as shown in Figure 1. Due to the complex test form design, the number of items responded to by each student was different, and the missing-at-random was assumed according to the test administration design. In the analysis, 489 middle school students who answered at least two items in each construct were included.

The second empirical dataset came from a subset of the Carbon-Transforming Processes in Socio-Ecological Systems data (Jin, et al., 2015). This data measured teacher's content knowledge (CK) and pedagogical content knowledge (PCK), which is referred to as the CK-PCK data. The CK construct describes how teachers develop their subject matter knowledge in science ranging from force-dynamic reasoning to matter and energy transformation. In relation to the CK construct, teachers are expected to be able to transform subject matter knowledge for the purpose of teaching, which is connected to both science content knowledge and general pedagogy knowledge. This PCK construct ranges from the content-general approaches to targeting the transition from naïve ideas to scientific big ideas. Readers can refer to Jin et al. (2015) for the details of the construct map for the CK-PCK data, but the brief relationship between CK-PCK corresponds with the two hypothesized links in Figure 2 (right panel). In each construct, six polytomous items ($j=0,1,2$) were used to measure the CK and PCK construct, respectively. The responses from 192 teachers who answered at least two items in each construct were analyzed.

Two-Level Case for the ADM

The proposed change-point SCM and the baseline model (RMRM) were analyzed using the ADM dichotomous data. Note again that the baseline model corresponds to a constrained version of the change-point SCM in which only the cut-score parameter was estimated and the discontinuity parameter was constrained to be zero. Thus, comparing the proposed change-point SCM with the baseline model allowed us to investigate the consequences of ignoring the discontinuity parameters. In addition, the change-point SCM in equations 2 and 3, which incorporated one cut-score parameter and one discontinuity parameter, can be compared with the models formulated in Choi (2013) for the same situation. We also employed Choi's (2013) models to investigate the effect of estimating the cut score and the different way of parameterization.

For the two-level ADM, the parameter estimates of the RMRM and the change-point SCM analysis along with Choi's (2013) models are listed in Table 2. The worst model fit came from the RMRM, which estimated only the cut score and ignored the discontinuity parameter. In terms of the estimated DIC, the change-point SCM showed the best model fit among the four different model formulations. This implies that the estimation of both the cut score and the discontinuity parameter is favored in this empirical data. Because the cut-score estimate from the change-point SCM is not statistically different from zero and Choi's models showed better fit than the RMRM, incorporating the discontinuity parameter appears to be more crucial for this empirical data than incorporating the cut-score parameter.

In the RMRM, the cut score was estimated to be -0.05 with a standard error of 0.27. In the change-point SCM in which a discontinuity parameter was incorporated, the cut score (γ) was estimated to be -0.20. This cut score is lower than the RMRM, and it resulted in 39.7% of examinees being at the non-proficient level and 60.3% of examinees being at the proficient level in the requirement construct. On the contrary, the cut score was arbitrarily constrained to be zero in Choi's

two models and classified approximately half the examinees as being in the non-proficient level and proficient level, respectively. The estimated negative cut score from the change-point SCM suggests that the items were somewhat easy overall for the examinees, and the classification of the proficient level and the non-proficient level needs to be determined at a lower logit than zero, although a 95% credible interval for the cut score included zero.

Table 2**Parameter Estimates and Standard Errors for the Two-Level Case ADM**

	Choi (2013)'s Models ($\gamma=0$)				Estimating Cut Score			
	SCM 1		SCM 2 ($\lambda_2=0$)		RMRM		Change-Point SCM	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
γ					-0.05	0.27	-0.20	0.16
λ_1	1.26	0.41	1.49	0.33			1.63	0.18
λ_2	0.66	0.49						
σ_{CoS}^2	1.70	0.25	1.75	0.24	1.86	0.26	1.72	0.24
σ_{Cha}^2	1.61	0.57	1.09	0.23	2.94	0.32	1.01	0.18
$\sigma_{CoS,Cha}$	0.85	0.44	0.32	0.23	1.55	0.21	0.25	0.16
$\rho_{CoS,Cha}$	0.51		0.23		0.66		0.19	
Level 1	48.0%		47.0%		33.5%		39.7%	
Level 2	52.0%		53.0%		66.5%		60.3%	
DIC	6533.21		6458.97		6596.33		6451.31	

In the change-point SCM, the discontinuity parameter, λ , was estimated to be 1.63 with a standard error of 0.18. The positive value of the estimated discontinuity parameter, in particular, suggests that the examinees at the proficient level of CoS had an advantage in reaching the proficient level in Chance as much as 1.63 logits on average. In comparison, the SCM 1 used in Choi (2013), in which two discontinuity parameters were incorporated for both the proficient and the non-proficient level, the examinees at the proficient level of CoS have a greater advantage in Chance ($\lambda_1=1.26$) compared with the examinees at the non-proficient level of CoS ($\lambda_2=0.66$) in terms of the

magnitude of the discontinuity parameter. Specifically, the advantage for the examinees at the non-proficient level of CoS (λ_2) was not significantly different from zero. This is reflected in the model fit comparison as well; when a single discontinuity parameter was incorporated, the model fit improved.

Regarding the item parameters, the estimates from the RMRM and the change-point SCM are illustrated in Figure 6 for ease of interpretation. Although the cut-score estimates were different between the RMRM and the change-point SCM, the item parameter estimates in both the requirement construct and the target construct appear to be quite consistent across models. Thus, as examined in the simulation study, we can conclude that ignoring the discontinuity parameter for the two-level case ADM had negligible impacts on the estimation of the item parameters.

Figure 6. Comparison of the item parameter estimates using the two-level ADM.

Finally, the latent person proficiency distribution estimated from the RMRM and change-point SCM is illustrated in Figure 7 with a solid line for RMRM and a dotted line for change-point SCM. The upper panel shows the requirement construct and the lower panel shows the target construct. In particular, the dotted curve in the lower panel represents the distribution of $\vartheta_{pT} + \lambda$ for examinees at the proficient level of the requirement construct. It is apparent that the latent ability distribution of the target construct was increased by the discontinuity parameter and the examinees in the target construct are differentiated more clearly into two groups (i.e., show a bimodal distribution). Furthermore, the variance of the latent ability in the target construct (σ_{Cha}^2) and the covariance between the two constructs, ($\sigma_{CoS,Cha}$) were estimated to be smaller compared with the RMRM. As examined in the simulation study, this suggests that the RMRM has overestimated the variance of the target construct, and thus, incorporating the discontinuity parameter in the change-point SCM has modified it appropriately.

Figure 7. Distribution of bivariate person proficiencies using the two-level ADM.

Three-Level Case for the CK-PCK

The proposed change-point SCM and the baseline model (RMPCM) were analyzed using the CK-PCK polytomous data. Four types of change-point SCMs with different constraints on the hypothesized links were analyzed. The first change-point SCM analysis (SCM 1) incorporated two cut-score parameters and two discontinuity parameters as described in equations 5 and 6. While SCM 1 is the most generalized version and estimates two discontinuity parameters freely at each cut score, the second change-point SCM analysis (SCM 2) constrained that the discontinuity parameters at the two cut scores were equivalent ($\lambda_1 = \lambda_2$). Two other restrictions on the discontinuity parameters were also investigated. The third change-point SCM analysis (SCM 3) constrained that the hypothesized link only existed between Level 2 of the requirement construct and Level 2 of the target construct without a link at Level 3 ($\lambda_2=0$). In contrast, the fourth change-point SCM analysis (SCM 4) constrained that the hypothesized link only existed between Level 3 of the requirement construct and Level 3 of the target construct without the link at Level 2 ($\lambda_1 = 0$). These assumptions underlying the four different change-point SCMs cover a variety of relationships between the requirement construct and the target construct for the typical three-level case in the LP. Comparing these four different models enables us to examine which assumption can be empirically verified for the dataset in terms of the model fits, and can serve as validity evidence to support or disprove the theoretically hypothesized links in the LP.

In the LP that describes the relationship between CK and PCK, it was assumed that teachers must achieve Level 3 in the CK construct in order to achieve Level 3 in the PCK construct. They also must reach Level 2 in the CK construct, in theory, in order to reach Level 2 in the PCK construct. Thus, hypothesized links in the three-level case CK-PCK correspond with the SCM 1 with two hypothesized links at Level 2 and Level 3, respectively. In theory, it is expected that the SCM 1 for the CK-PCK data would show the best model fit among the four different analyses.

The resulting parameter estimates from the RMPCM and the four change-point SCM analyses along with the estimated DIC values are listed in Table 3. Unlike the expectation of the cognitive scientists who developed the LP (Jin et al., 2015), the SCM 4, which specified the link only at the higher level between Level 3 of the CK and Level 3 of the PCK, fit better than the other types of SCMs. This implies that the hypothetical link between Level 2 of the CK-PCK LP is not yet empirically well supported for this data.

Table 3

Parameter Estimates and Standard Errors for the Three Levels CK-PCK

	RMPCM		Change-Point SCM 1		Change-Point SCM 2 ($\lambda_1 = \lambda_2$)		Change-Point SCM 3 ($\lambda_2 = 0$)		Change-Point SCM 4 ($\lambda_1 = 0$)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
λ_1			0.00	1.00	2.01	0.17	1.98	0.17		
λ_2			2.03	0.16	2.01	0.17			2.49	0.34
γ_1	-0.01	0.09	0.26	0.09	0.26	0.09	0.29	0.09	-0.39	0.06
γ_2	0.01	0.09	-0.26	0.09	-0.26	0.09	-0.29	0.09	0.39	0.06
σ_{CK}^2	0.75	0.15	0.53	0.12	0.52	0.13	0.50	0.13	0.74	0.17
σ_{PCK}^2	2.27	0.40	0.22	0.11	0.23	0.11	0.24	0.12	1.04	0.26
$\sigma_{CK,PCK}$	0.82	0.16	0.12	0.08	0.14	0.10	0.16	0.09	0.68	0.13
$\rho_{CK,PCK}$	0.63		0.36		0.40		0.47		0.78	
Level 1	32.3%		17.7%		16.6%		16.7%		11.5%	
Level 2	1.0%		47.9%		49.0%		49.5%		49.5%	
Level 3	66.7%		34.4%		34.4%		33.8%		39.0%	
DIC	3772.86		3726.45		3727.68		3734.79		3692.08	

In the RMPCM, the cut scores were estimated to be -0.01 and 0.01, respectively, which are not significantly different from zero. This implies that there are virtually only two distinct levels rather than three. As implied by the cut scores, the baseline RMPCM was not successful in separating the examinees into three groups. On the contrary, in the best-fitting model SCM 4, the cut scores were estimated at -0.39 and 0.39, respectively. These cut scores resulted in 11.5% of examinees being classified as Level 1, 49.5% as Level 2, and 39.0% as Level 3 of the CK construct.

Furthermore, it is interesting to note that for the other three types of SCMs, the cut-score estimates showed the opposite result: The second cut score was estimated to be lower than the first cut score, which does not make sense in our current case. We expect that the second cut score (which differentiates the higher levels, Level 2 and 3) is higher than the first (which differentiates the lower levels, Level 1 and 2). Although these reversed cut score estimates are different from the expectations, this reversed pattern in the cut scores can occur, at least in the mathematical model, and it is also reflected in the worse model fits to the data.

For ease of interpretation, the item parameter estimates from the RMPCM and the SCM 4 are illustrated in Figure 8. As can be seen in the left panel, the overall item difficulty parameter estimates (β_i) appear to be consistent between two different models, although there are some discrepancies in the right panel for the step parameter estimates. We suspect that the discrepancies may be due to the small number of the items and small sample size collected in this data, which is only six items for each construct from 192 teachers. In a future study, the minimum number of items and persons that are required for the reliable estimation of the item parameters in the change-point SCMs would be needed.

Figure 8. Comparison of the item parameter estimates using the three-level CK-PCK.

As with the examples above and as predicted from the simulation study, the estimated variance of the target construct (σ_{PCK}^2) and the covariance between the two constructs were smaller than the corresponding values of the baseline RMPCM. We see that the RMPCM has overestimated the variance of the target construct and that incorporating the discontinuity parameters in the change-point SCM has reduced the variance. The latent person proficiency distributions estimated from the RMPCM and change-point SCM 4 is illustrated in Figure 9 with a solid line for RMPCM and a dotted line for change-point SCM 4. In particular, the dotted curve in the lower panel combines the distribution of $\theta_{PT} + \lambda_2$ for examinees in the Level 3 on the requirement construct. It is apparent from the graph that the latent ability distribution of the requirement construct was quite similar for the two solutions. On the other hand, the distribution of the target construct was increased by the discontinuity parameter and the examinees on the target construct are differentiated more clearly. From the initial LP perspective, it is expected to be trimodal, but it looks bimodal because one of the two discontinuity parameters is zero. Note that under the constraint of SCM 4 ($\lambda_1=0$), the mean advantage for the Level 3 examinees of the CK in attaining Level 3 of the PCK was quite large ($\lambda_2=2.49$).

Figure 9. Distribution of bivariate person proficiencies using the three-level CK-PCK.

Conclusion and Discussion

This study has proposed one way, based on change-point analysis, to formulate a measurement model for complicated LPs using the SCM approach. In the change-point SCM, the hypothesized level-specific relations across multiple constructs are modeled by incorporating the cut-score parameters and the discontinuity parameters. More specifically, (a) the change point corresponds to the cut scores in the requirement construct, leading to the classification of the examinees according to the cut scores, and (b) the discontinuity corresponds to the magnitude of the average change at the cut point, leading to increase the likelihood to reach certain level of the target construct for those who exceed the cut score in the requirement construct. This model formulation allowed us to: (a) estimate the cut score and the discontinuity at the cut score, (b) identify examinees' proficiency in terms of the level of the construct (i.e., LP level-referenced performance) as well as the latent continuum (i.e., in terms of logits), and (c) empirically investigate the hypothesized link embedded in the LP by comparing the relative fit of models with and without these complications.

In this study, the change-point SCM for the two-level case (two constructs with a single connection) and the change-point SCM for the three-level case (two constructs with one or two connections) were examined via a simulation study and also for an empirical illustration. Results from the simulation study indicate that the proposed change-point SCM appears to recover the parameters well. In addition, when the cut score was incorrectly fixed at a specific location, and thus was further from the true value, the model fit became worse, and the model parameters, including the discontinuity parameter and the variance-covariance of the multidimensional constructs, were found to be more biased.

In the empirical example using the ADM project data and the CK-PCK data, we illustrated the use of the proposed change-point SCM for validating the hypothesized link in an LP. For the three-level case specifically, by comparing four change-point SCMs with different constraints on the hypothetical link, we found validity evidence to support or disprove the initial theoretical hypothesized links in the LP. Moreover, through the application of the change-point SCM to the empirical datasets, we see that when the model incorporates incorrect constraints on the links, it is difficult to estimate the model parameters. Thus, future simulation studies investigating these issues will be helpful.

Noting that the change-point SCMs proposed in this paper involve the two- and three-level cases, it would be straightforward to extend it to cases with additional levels by increasing the number of cut scores and discontinuity parameters. One limitation of the current study is that the proposed models assume only one direction in the hypothesized links: initiated from the requirement construct and terminated at the target construct. However, one could investigate a more complicated case where the requirement construct for one link becomes the target construct for other links (such as the second link in Figure 1, from Cha 6 to CoS 4). In this case, the SCMs proposed in this study would need further enhancement as there is now no longer a clear distinction

between the requirement and target constructs. Further exploration with SCMs based on other approaches beyond that used in this investigatory analysis may be helpful to formulate those more complicated links.

Another limitation of the current study is that the score categories exactly match the number of levels in each of the dimensions of the LP. For example, for the two-level case, responses on every item were scored dichotomously, and for the three-level case, responses on every item were scored as Level 1, Level 2, and Level 3. This situation is not unusual given that the SCM framework is very likely to be grounded on a validated theory, careful item design, and a scoring rubric, following, for example, the BEAR Assessment System (Wilson & Sloane, 2000) or assessments that are aligned with “Next Generation” assessment systems (National Research Council, 2012, 2014). However, there can be more complicated situations in which the score categories for all items or for a subset of items are different from the number of levels in the LP. Polytomous responses, for example, could have been used for a two-level case (i.e, there could be differentiated scores *within* a level). Future extensions for more complicated cases in terms of the response types and the relations between the constructs may enhance the application of the SCM modeling as a measurement model for complex LPs.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1–23.
- Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. New York, NY: Springer.
- Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2), 221–241.
- Choi, I. H. (2013). *Applications of multilevel and multidimensional item response modeling using the Bayesian estimation*, unpublished doctoral dissertation. Berkeley, CA: University of California, Berkeley.
- Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, 43(2), 171–192. doi:10.1111/j.2044-8317.1990.tb00934.x

- Curtis, S. (2010). Bugs code for item response theory. *Journal of Statistical Software*, 36(c01).
- De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, 112(1), 129–158.
- Diakow, R., Irribara, D., & Wilson, M. (2011). *Analyzing the complex structure of a learning progression: Structural Construct Models*. Presented at the National Council of Measurement in Education, New Orleans, LA.
- Diakow, R., Irribara, D., & Wilson, M. (2013). *Some comments on representing construct levels in psychometric models*. In R. E. Millsap, L.A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th Annual Psychometric Society Meeting* (pp. 319-334). New York, NY: Springer.
- Draney, K., & Jeon, M. (2011). Investigating the saltus model as a tool for setting standards. *Psychological Test and Assessment Modeling*, 53(4), 486-498.
- 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* (pp. 119–130). New York, NY: Springer.
- 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.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: CRC Press.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Hall, C. B., Lipton, R. B., Sliwinski, M., & Stewart, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer’s disease. *Statistics in Medicine*, 19(11-12), 1555–1566.
- Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1), 1–17.
- Jin, H., & Anderson, C. W. (2012). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching*, 49(9), 1149-1180.
- Jin, H., Shin, H., Johnson, M., Kim, J., & Anderson, C. W. (2015). Developing learning progression-based teacher knowledge measures. *Journal of Research in Science Teaching*, 52(9), 1269-1295.
- Lehrer, R., Kim, M-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey & K. Nguyen (Eds.), *Learning over Time: Learning Trajectories in Mathematics Education* (pp. 31-60). Charlotte, NC: Information Age Publishers.
- Lehrer, R., & Wilson, M. (2011). *Developing assessments of data modeling: Construct maps as boundary objects*. Presented at the National Council on Measurement in Education, New Orleans, LA.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2), 187–212.

- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.
- National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.
- National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.
- R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria.
- Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Shin, H., Choi, J., & Draney, K. (2012). Using item response theory models for classifying students onto levels of achievement. Symposium session presented at the 2012 International Objective Measurement Workshop (IOMW), Vancouver, British Columbia, Canada.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York, NY: Springer.
- Western, B., & Kleykamp, M. (2004). A Bayesian change point model for historical time series analysis. *Political Analysis*, 12(4), 354–374.
- Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, 105(2), 276–289.
- Wilson, M. (2005). Constructing measures: An item response modeling approach. Retrieved from <http://psycnet.apa.org/psycinfo/2004-22396-000>
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716–730.
- Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in Science: Current Challenges and Future Directions* (pp. 317–343). Rotterdam, Netherlands: Sense Publishers.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208.

Appendix

Sample WinBUGS code

```
## Simulation study for Two-level case ##

model {
  for (j in 1:J) {
```

```

        for (i in 1:30) {
            resp[j,i]~dbern(prob[j,i])
            logit(prob[j,i]) <- theta[j,1]+gammac+tau.ir[i]
        }
        for (i in 1:30) {
            resp[j,(i+30)]~dbern(prob[j,(i+30)])
            logit(prob[j,(i+30)]) <- theta[j,2]-
tau.it[i]+lambda*step(theta[j,1]-gammac)
        }
    }

    for (j in 1:J) {
        theta[j, 1:2] ~ dmnorm(mu[1:2], R[1:2,1:2])
    }

    mu[1] <- 0
    mu[2] <- 0

    gammac ~ dunif(-6, 6)

    R[1:2,1:2] ~ dwish(Omega[1:2,1:2], 2)
    IR[1:2,1:2] <- inverse(R[1:2,1:2])
    corr <- IR[1,2]/(sqrt(IR[1,1]*IR[2,2]))

    lambda ~ dnorm(0, 1)

    for (i in 1:30) {
        tau.ir[i] ~ dnorm(0, 1)
    }
    for (i in 1:29) {
        tau.it[i] ~ dnorm(0, 1)
    }

    tau.it[30] <- -sum(tau.it[1:29])
}

```

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Shin, HJ;Wilson, M;Choi, I-H

Title:

Structured Constructs Models Based on Change-Point Analysis

Date:

2017-09-01

Citation:

Shin, H. J., Wilson, M. & Choi, I. -H. (2017). Structured Constructs Models Based on Change-Point Analysis. JOURNAL OF EDUCATIONAL MEASUREMENT, 54 (3), pp.306-332. <https://doi.org/10.1111/jedm.12146>.

Persistent Link:

<http://hdl.handle.net/11343/293408>