Eurasian Journal of Educational Research, Issue 54, 2014, 117-134

An Explanatory Item Response Theory Approach for a Computer-Based Case Simulation Test

Nilüfer KAHRAMAN*

Suggested Citation:

Kahraman, N. (2014). An explanatory item response theory approach for a computer-based case simulation test, *Eurasian Journal of Educational Research*, 54, 117-134. http://dx.doi.org/10.14689/ejer.2014.54.7

Abstract

Problem: Practitioners working with multiple-choice tests have long utilized Item Response Theory (IRT) models to evaluate the performance of test items for quality assurance. The use of similar applications for performance tests, however, is often encumbered due to the challenges encountered in working with complicated data sets in which local calibrations alone provide a poor model fit.

Purpose: The purpose of this study was to investigate whether the item calibration process for a performance test, computer-based case simulations (CCS), taken from the United States Medical Licensing Examination® (USMLE®) Step 3® examination may be improved through explanatory IRT models. It was hypothesized that explanatory IRT may help improve data modeling for performance assessment tests by allowing important predictors to be added to a conventional IRT model, which are limited to item predictors alone.

Methods: The responses of 767 examinees from a six-item CCS test were modeled using the Partial Credit Model (PCM) and four explanatory model extensions, each incorporating one predictor variable of interest. Predictor variables were the examinees' gender, the order in which examinees encountered an individual item (item sequence), the time it took each examinee to respond to each item (response time), and examinees' ability score on the multiple-choice part of the examination.

Results: Results demonstrate a superior model fit for the explanatory PCM with examinee ability score from the multiple-choice portion of Step 3. Explanatory IRT model extensions might prove useful in complex performance assessment test settings where item calibrations are often problematic due to short tests and small samples.

Recommendations: Findings of this study have great value in practice and implications for researchers working with small or complicated response data. Explanatory IRT methodology not only provides a way to improve data modeling for performance assessment tests but also enhances the

* Dr. Başkent University, Ankara, Turkey, e-mail: nkahraman6@gmail.com

inferences made by allowing important person predictors to be incorporated into a conventional IRT model.
Keywords: Explanatory Item Response Theory, Partial Credit Model, Item Response Theory, Performance Tests, Item calibration, Ability estimation, Small tests

Introduction

Over the past few decades, Item Response Theory (IRT) applications have become a vital part of the scoring processes in many large-scale test settings. IRT encompasses a family of nonlinear models that provide an estimate of the probability of a correct response on a test item as a function of the characteristics of the item (e.g., difficulty, discrimination) and the ability level of test takers on the trait being measured (e.g., Hambleton, Swaminathan & Rogers, 1991; McDonald, 1999; Skrondal & Rabe-Hesketh, 2004). IRT models are particularly appealing in that if the IRT model fits the data set, the resulting item and ability parameters can be assumed to be sample independent (item and ability parameter invariance property). Practitioners working with *multiple-choice tests* have long utilized IRT models to link observable examinee performance on test items to an overall unobservable ability, as wells as to evaluate performance of test items and test forms for quality assurance (See Hambleton & Van der Linden, 1982, for an overview).

Applications of IRT models to *performance tests*, however, have long been encumbered by the challenges encountered in modeling novel performance test data. Historically, one issue was that the IRT models were developed for dichotomous items (Spearman, 1904; Novick, 1966). This made them unsuited for performance tests that often had items with ordinal categorical scales (to allow scoring partially correct answers). However, extensions for polytomous items (Bock, 1972; Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996; Samejima, 1969) soon emerged and solved this particular issue. Another issue that remains to date is goodness of model fit. Although performance tests with novel item formats are believed to be more suited for measuring higher-level examinee abilities (Kane & Mitchell, 1996; Nitko, 1996), they are also typically very difficult to model (e.g., Masters, 1982; Yen, 1983). One reason is that performance tests are almost always drastically shorter than their multiple-choice counterparts. This makes it very challenging for many performance tests to satisfy the demand for large numbers of items for IRT models because it is often very expensive to develop and administer performance tests that are as lengthy as their multiple-choice counterparts. Another reason is the contextual effects introduced by the novelty of test. The influence of various person and test design variables is often amplified for performance tests, undermining the goodness of fit for the estimated IRT models. To this end, the current study investigates whether an alternative IRT modeling approach with added covariates from the generalized linear and non-linear mixed modeling framework (Embretson, 1998; De Boeck & Wilson, 2004; Wang, Wilson & Shih, 2006) can be used to help improve model estimation for a novel performance tests, namely, for computer-based case simulations (CCS).

Purpose

The current study was undertaken to investigate whether the item calibration process for the CCS examination could be improved using an explanatory IRT model. The CCS is a part of the United States Medical Licensing Examination® (USMLE®) Step 3 and was introduced in 1999, when the examination transitioned from paper-and-pencil administration to computer-based administration. This examination uses a small series of computer-based case simulations (CCS items) to expose examinees to interactive patient-care simulations; for each simulation they must initiate and manage patient care, while receiving patient status feedback and managing the simulated time in which the case unfolds (Margolis, Clauser, & Harik, 2004; Clauser, Harik, & Clyman, 2000).

The explanatory IRT model application presented in this paper explores the usefulness of four different predictor variables in improving the item calibration process of the CCS examination: examinees' gender, the order in which each individual CCS item was presented during the examination (item sequence), the time it took each examinee to respond to each item (response time), and examinees' ability score on the multiple-choice part of Step 3. Although only the latter covariate was hypothesized to be an important predictor of examinee performance on the CCS, as it is the only construct-relevant covariate, the importance of the other variables were also tested as potential predictors. The usefulness of item sequence and response time were explored, relying on the recent literature that suggests their usefulness as predictors of examinee performance (e.g., Ramineni, Harik, Margolis, Clauser, Swanson & Dillon, 2007; Lu & Sireci, 2007; Leary & Dorans, 1985; Yen, 1980). The importance of the gender variable was tested mainly to investigate whether CCS items were easier for one of the gender subgroups.

A series of alternative explanatory IRT models were estimated using the Partial Credit Model (PCM) and with one predictor variable at a time. This resulted in the following models: (1) a base model (no covariates), (2) an explanatory model with gender effect, (3) an explanatory model with item response time effects, (4) an explanatory model with item sequence effects, and (5) an explanatory model with examinees' scores on the multiple-choice part of the Step 3 examination (MCT score). Table 1 gives an overview of the estimated models. The PCM model with no covariates was used as a base model to evaluate the hypothesized improvement in model fit for each explanatory model with one added covariate. This "one covariate" at a time approach was to ensure that, if and when observed, any improvement in model fit is due to the added covariate alone.

Table 1.

Estimated Models and Covariates

Models	Covariates			
	Gender	Item response times	Item sequence	Step 3 MCQ Test Score
1. PCM				
2. PCM with gender	X			
3. PCM with response times		X		
4. PCM with item sequence			X	
5. PCM with MCT score				X

Method

Data

Study data included the responses of 767 examinees to a six-item CCS test, each of which was administered in random order under standard testing conditions with a maximum of 25 minutes of testing time per item. For this analysis, examinee responses were coded using a 3-point category scale from 0 to 2, with 2 representing maximum credit for a given CCS item.

Model Estimation

For dichotomous items, under the Rasch model (Rasch, 1960; Wright, 1997), the probability of a positive response (or a correct answer) to item i for person j with latent trait θ is

$$P\left(Y_{ji}=1 \mid \theta_j, \beta_i\right)=\frac{\exp\left(\theta_j-\beta_i\right)}{1+\exp\left(\theta_j-\beta_i\right)} \quad (1)$$

where β_i is the difficulty of item i. The probability of a person's answering an item correctly is, therefore, a function of the difference between the person's ability and the difficulty of the item. The person parameters are assumed to be independently and normally distributed with a mean of zero and a variance of σ^2. In other words, the person parameter is a random effect while the item parameter is a fixed effect.

The partial credit model (PCM, Masters, 1982) extends the Rasch model for binary responses to pairs of adjacent categories in a sequence of ordered responses. For an item on an m-point scale, there are m-1 step parameters to estimate. Step parameters, β_{im}, refer to the value of θ_j where the probabilities of responding in category m and m-1 are equal. For an item with a 3-point scale, the probabilities of responding to each of the categories are given by

$$P(Y_{ji}=0\mid\theta_j,\beta_{im})=\frac{1}{1+\exp(\theta_j-\beta_i)+\exp(2\theta_j-\beta_{i1}-\beta_{i2})},$$

$$P(Y_{ji}=1\mid\theta_j,\beta_{im})=\frac{\exp(\theta_j-\beta_{i1})}{1+\exp(\theta_j-\beta_i)+\exp(2\theta_j-\beta_{i1}-\beta_{i2})},$$

(2)

$$P(Y_{ji}=2\mid\theta_j,\beta_{im})=\frac{\exp(2\theta_j-\beta_{i1}-\beta_{i2})}{1+\exp(\theta_j-\beta_i)+\exp(2\theta_j-\beta_{i1}-\beta_{i2})}.$$

Figure 1 plots category response functions for an illustrative CCS item with a 3-point scale with β_{i1}=-1 and β_{i1}=1. In the figure, it can be seen that the category response functions for categories 0 and 1 intersect at β_{i1} while the category response functions for categories 1 and 2 intersect at β_{i2}.

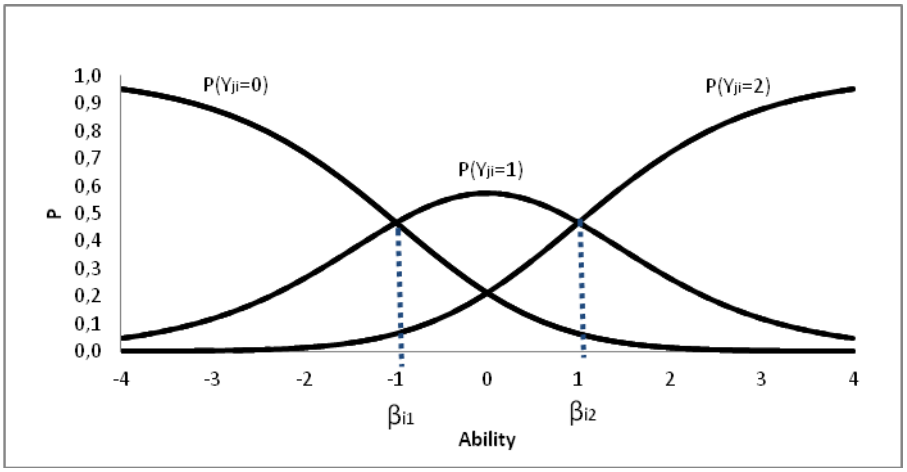

Figure 1. Category Response Functions for an illustrative CCS item with a three-point scale

The Partial Credit Model with random effects

The linear random effects PCM with a person covariate Z_{ji} is given by

$$P(Y_{ji}=0\mid Z_{jim}\theta_j,\beta_{im})=\frac{1}{1+\exp(Z_{ji1}\theta_j-\beta_i)+\exp(2Z_{ji2}\theta_j-\beta_{i1}-\beta_{i2})}$$

$$P\left(Y_{ji}=1 \mid Z_{jim}\theta_j, \beta_{im}\right)=\frac{\exp\left(Z_{jji1}\theta_j-\beta_{i1}\right)}{1+\exp\left(Z_{ji1}\theta_j-\beta_i\right)+\exp\left(2Z_{ji2}\theta_j-\beta_{i1}-\beta_{i2}\right)},$$

(3)

$$P\left(Y_{ji}=2 \mid Z_{jim}\theta_j, \beta_{im}\right)=\frac{\exp\left(2Z_{ji2}\theta_j-\beta_{i1}-\beta_{i2}\right)}{1+\exp\left(Z_{ji1}\theta_j-\beta_i\right)+\exp\left(2Z_{ji2}\theta_j-\beta_{i1}-\beta_{i2}\right)}$$

The PCM and the explanatory PCMs were estimated using the PROC NLMIXED routine of the Statistical Analysis System (SAS version 9.1.3). For the analysis we used a quasi-Newton-Raphson optimization technique and a non-adaptive Gauss-Hermite approximation with 10 quadrature points for each dimension (the SAS code used in calculations is given at the end of this paper as an appendix). Goodness of model fit was evaluated using the -2 log likelihood, the *Akaike information criterion* (AIC) (Akaike, 1974) and the *Bayesian information criterion* (BIC; Schwarz, 1978), with lower values indicating better fit.

Results

The results summarized in Table 2 show that by adding the MCT score of examinees to the model as a random effect, the PCM model fit was improved, producing the lowest -2 log likelihood, AIC, and BIC. There was no improvement over the base PCM for the remaining explanatory models with gender, item sequence, or item response time as a covariate. Table 3 lists category threshold and variance parameter estimates produced by these models. As revealed by the observed improvement in the corresponding model fit statistics, MCT score was the only significant predictor (0.50) among the four considered. Item sequence, response times, and gender effects were all approximately zero (0.01, smaller than 0.001, and 0.04, respectively).

Table 2.

Model Fit Comparisons

Model	Number of Parameters	-2 Log Likelihood	AIC	BIC
PCM	13	8909	8935	8996
PCM with gender	14	8908	8936	9002
PCM with response times	14	8896	8924	8989
PCM with item sequence	14	8906	8934	8999
PCM with MCT score	14	8862	8889	8955

* Models with multiple predictors were not feasible for this data set since only the MCT score was useful as a predictor among the four considered.

Table 3.

Parameter Estimates

Parameter	Models				
	PCM	PCM with gender	PCM with response times	PCM with item sequence	PCM with MCT score
$b1_{cat1}$	-0.74	-0.72	-1.06	-0.66	-1.06
$b2_{cat1}$	-0.74	-0.72	-1.03	-0.66	-0.17
$b3_{cat1}$	-1.35	-1.33	-1.62	-1.28	-0.78
$b4_{cat1}$	-0.33	-0.31	-0.59	-0.25	0.24
$b5_{cat1}$	-0.47	-0.45	-0.78	-0.40	0.10
$b6_{cat1}$	-0.97	-0.95	-1.25	-0.90	-0.40
$b1_{cat2}$	-0.86	-0.84	-1.16	-0.79	-0.29
$b2_{cat2}$	-1.14	-1.12	-1.41	-1.07	-0.57
$b3_{cat2}$	0.31	0.32	0.05	0.37	0.87
$b4_{cat2}$	-0.61	-0.59	-0.85	-0.54	-0.04
$b5_{cat2}$	-0.71	-0.69	-1.00	-0.64	-0.14
$b6_{cat2}$	0.51	0.52	0.25	0.58	1.07
Effect of the predictor variable	-	0.04	0.00	0.01	0.50
σ^2	0.21	0.21	0.23	0.21	0.17

* Standard Error of the estimates ranged between 0.08 and 0.16.

Figure 2 and Figure 3 plot category response functions for the six CCS items using threshold parameters estimated by the base PCM and the best fitting explanatory PCM with MCT Score predictor, respectively. Comparing the graphical displays of probabilities computed for each response category given in Figure 1 with Figure 2 reveals that aiding the base PCM model with MCT score greatly improves the functional form of CCS items.

Discussion

Explanatory IRT models incorporating item or person covariates are increasingly used in many test settings to learn more about predictors of examinee performance (e.g., Fischer, 1983; De Boeck & Wilson, 2004; Embretson, 1984; Embretson, 1997) and to help improve item calibration and scoring procedures (e.g., Fox, 2005; Harting, Frey, Nold & Klieme, 2012; Zinderman, 1991). The premise of the current paper is that they may also be useful in the context of authentic performance assessment tests with small tests. This paper demonstrates that explanatory PCMs with meaningful

predictors might prove useful in calibrating complex performance tests similar to the USMLE CCS, which otherwise could not be calibrated.

For the CCS application presented in this paper, the meaningfulness of four individual predictor variables was tested: examinees' gender, the order in which each individual CCS was presented during the examination (item sequence), the time it took each examinee to respond to each case (response time) and examinees' ability score on the multiple-choice part of Step 3. While only the latter predictor variable was found to be of statistical and practical significance, the results nicely illustrate how an explanatory approach can be used to investigate the usefulness of individual predictor variables in model estimation. Although it was not feasible for the present application, as only one of the covariates was found to be of statistical importance, it is recommended that researchers explore multivariate model extensions to further assess if a more complex model with multiple predictors may further improve model fit.

The findings of this study have great value for researchers and practitioners working with small performance tests and complex response data in which local calibrations alone provide a poor model fit. Explanatory model extensions of PCM not only provide a way to improve data modeling for short performance assessment tests but also open other possibilities by allowing various person predictors to be added to conventional item response models, which are limited to item predictors alone. Future research should investigate the influence of other item and person predictors on CCS performance to determine if any can lead to a stronger model fit, more stable parameter estimates, or a more precise measure of CCS proficiency. One predictor of future interest, for example, could be the examinees' postgraduate medical training (Dillon, Henzel & Walsh, 1997; Feinberg, 2012). Examinees who are exposed to a broad range of training during their residency or clinical experience might perform better on the MC items of Step 3 as compared to examinees who have a narrow training focus (Sawhill, Dillon, Ripkey, Hawkins, & Swanson, 2003).

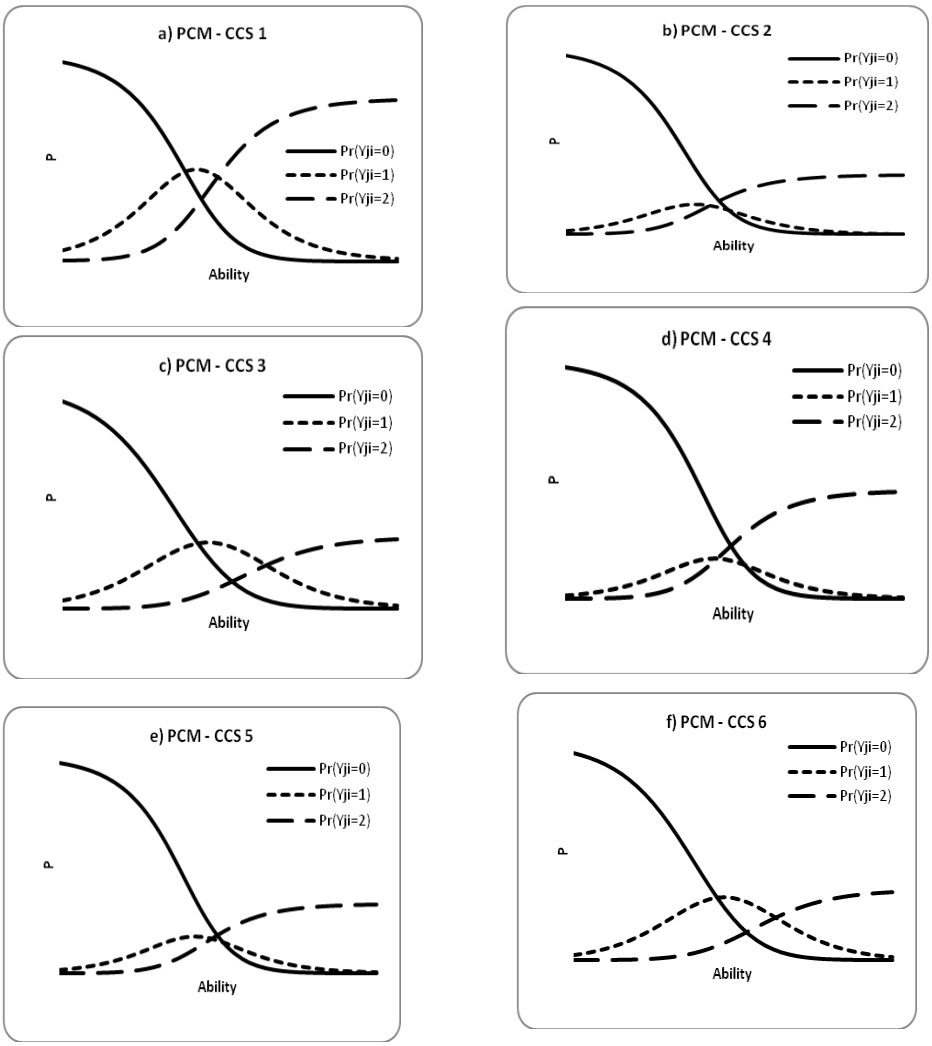

Figure 2. Item Characteristics Curves for the six CCS items estimated by the base PCM

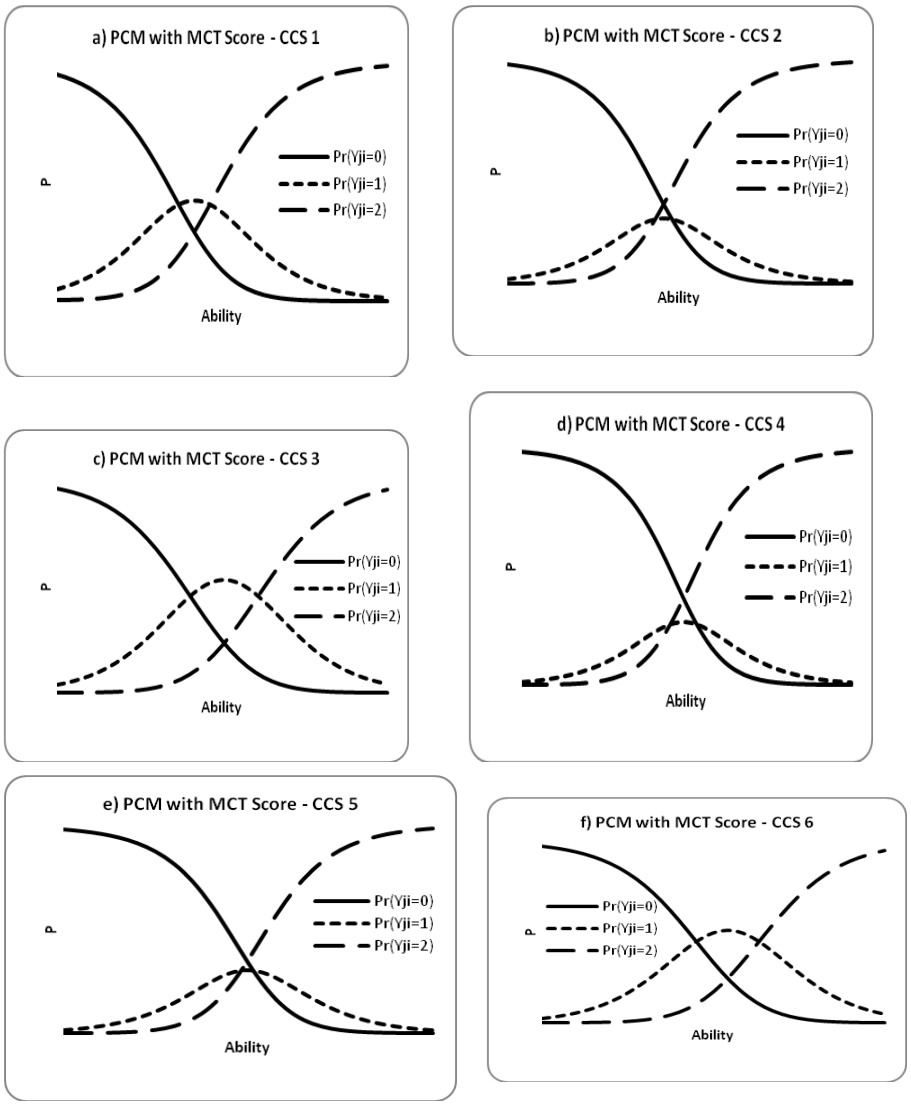

Figure 3. Item Characteristics Curves for the six CCS items estimated by the explanatory PCM with MCT Scores

References

Akaike, M. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.

Clauser, B. E., Harik, P., Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37,* 245-262.

De Boeck, P. & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York, NY: Springer.

Dillon, G. F., Henzel, T. R., & Walsh, W. P. (1997). The impact of postgraduate training on an examination for medical licensure. In Advances in Medical Education (pp. 146-148). Springer Netherlands.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, 49(2), 175-186.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380-396.

Embretson, S. E. (1997). Multicomponent response models. In *Handbook of modern item response theory* (pp. 305-321). Springer New York.

Feinberg, R. A. (2012). The impact of postgraduate training on USMLE® step 3® and its computer-based case simulation component. *Journal of general internal medicine,* 27 (1), 65-70.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, 48 (1),3-26.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K. and Sykes, R. C. (1996). Scaling Performance Assessments: A Comparison of One-Parameter and Two-Parameter Partial Credit Models. *Journal of Educational Measurement*, 33: 291–314.

Fox, J.P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, 58: 145–172.

Hambleton, R. K., Swaminathan H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hambleton, R. K., & Van der Linden, W. J. (1982). Advances in Item Response Theory and Applications: An Introduction, *Applied Psychological Measurement*, 6 (4), 373-378

Harting, J., Frey, A., Nold, G. & Klieme, E. (2012). An Application of Explanatory Item Response Modeling for Model-Based Proficiency Scaling, *Educational and Psychological Measurement*, 72 (4), 665-686

Kane, M. B. & Mitchell, R. (1996). *Implementing Performance Assessment*. Mahwah, NJ: Lawrence Erlbaum Ass.

Leary, L. F & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern, Review of Educational research, 55, 387-413.

Lu, Y & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*, 29-37.

Margolis, M.J., Clauser B. E., and Harik P. (2004). Scoring the computer-based case simulation component of USMLE Step 3: A comparison of preoperational and operational data. *Academic Medicine, 79*, 62 - 64.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47 (2), 149-174.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah NJ: Erlbaum

Nitko, A. J. (1996). *Educational assessment of students* (2nd. Ed.). Englewood Cliffs NJ: Prentice-Hall.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1-18.

Ramineni, C., Harik, P., Margolis, M.J., Clauser, B.E., Swanson, D.B. & Dillon, G.F. (2007) Sequence Effects in the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills (CS) Examination. *Academic Medicine*, 10, S101-S104.

Rasch, G. (1960), An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19: 49–57.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika,* Monograph No. 17.

Sawhill, A. J., Dillon, G. F., Ripkey, D. R., Hawkins, R. E., & Swanson, D. B. (2003). The impact of postgraduate training and timing on USMLE Step3 performance. *AcademicMedicine, 78*, 10-12.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Skrondal, A.& Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. CRC Press

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201-292.

United States Medical Licensing Examination® (USMLE®). (2010). Step 3® Content Description Online. Retrieved June 6, 2010, from the World Wide Web: http://www.usmle.org/examinations/step3/step3_content.html.

Wang, W.-C., Wilson, M. and Shih, C.L. (2006), Modeling Randomness in Judging Rating Scales with a Random-Effects Rating Scale Model. *Journal of Educational Measurement*, 43: 335–353.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14:97–116.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, 17: 297–311.

Yen, W. M. (1983). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30: 187–213.

Zinderman, A. H. (1991). A generalized Rasch model for manifest predictors. Psychometrika, 56 (4), 589-600.

Açıklayıcı Madde Tepki Kuramının İnteraktif Bir Bilgisayar Simülasyon Testine Uygulanması

Atıf:

Kahraman, N. (2014). An explanatory item response theory approach for a computer-based case simulation test, *Eurasian Journal of Educational Research*, 54, 117-134.

Özet

Problem: Test geliştirme ve geliştirilen testlerin güvenirlik ve geçerliğini araştırmada sıkça kullanılan Madde Tepki Modelleri çoktan seçmeli testlerde uzun zamandır madde ve test kalitesini kontrol amacıyla kullanılmaktadır. Bu modellerin aynı amaçla uygulamalı testlerde kullanımı ise birçok zorluk ile karşılaşmıştır. Bu zorluklardan ilki ilk geliştirilen Madde Tepki Modellerinin sadece ikili puanlanan test maddeleri için uygun olmasıydı. Oysa uygulamalı test maddeleri çoğunlukla kısmı puanlama gerektirecek şekilde geliştirilir. Kısmi puanlamaya uygun Madde Tepki Modellerinin geliştirilmesiyle bu sorun kısa zaman içerisinde çözümlendi. Bir diğer zorluk ki hala güncelliğini korumaktadır, uygulamalı test verilerinin Madde Tepki Modelleri ile modellenmeye daha az uygun oluşlarıdır. Bir başka deyişle, uygulamalı testlerde kullanıldığında Madde Tepki Modelleri uygulamaları güvenirliği çok iyi olmayan madde ve kişi istatistikleri ile sonuçlanabilmektedir. Bunun iki önemli nedeni uygulamalı testlerin çoktan seçmeli testlere göre daha kısa oluşları ve de uygulamalı test sorularının ölçülmesi istenen becerilerle direk olarak ilgili olmayan birçok faktörlerin etkisine çoktan seçmeli sorulardan daha açık oluşlarıdır. Uygulama testleri ile çalışan psikometristler de diğer testlerle çalışan meslektaşları gibi Madde Tepki Modellerinin sağlayacağı örneklem bağımlılığı oldukça düşük olan madde ve kişi istatistiklerine ihtiyaç duymakta ve yukarıda sayılan zorlukları aşabilecek yeni modellerin geliştirilmesini beklemektedir.

Amaç: İkincil değişkenleri model hesaplamalarına yordayıcı olarak dahil etmeye izin veren Açıklayıcı Madde Tepki Modelleri birçok farklı ortamda uygulanan bir çok testin madde ve kişi istatistiklerinin kalitesinin arttırılmasında kullanılmaktadır. Ancak bu modellerin uygulamalı testlerde kullanıldıklarında sıkça karşılaşılan düşük model uygunluğu ve düşük güvenirlik problemlerini çözmede kullanılması ile ilgili bir çalışma henüz yapılmamıştır. Bu çalışmanın amacı Madde Tepki Modelleri kullanıldığında veriye uygunluk indeksleri düşük çıkan altı adet interaktif

uygulamalı madde içeren bir uygulama testi için Açıklayıcı Madde Tepki Modellerinin iyi bir alternatif olup olmadığını değerlendirmekti.

Yöntem: Bu çalışmanın örneklemi araştırmaya konu olan uygulamalı CCS (Computer Case Simulations) testini alan 767 kişinin altı uygulama sorusuna verdiği cevaplardan oluşmaktadır. CCS Amerika'da çalışma lisansı almaya hak kazanabilmek için hekim adaylarının aldıkları üç aşamalı bir testin, üçüncü ve son aşmasında verilen bir uygulama testidir. Hekim adayları bu son aşamada çoktan seçmeli bir testin yanı sıra bu uygulama testini de alırlar. Sınav sırasında, her bir CCS uygulaması için hekim adaylarına bilgisayar ortamında bir hasta profili verilir. Hekim adayları uygun olduklarını düşündükleri teşhis ve takipleri interaktif bir ortamda yapabilmektedir. Her bir CCS için hekim adayları maksimum 25 dakika harcayabilir. Bu çalışmada örneklemdeki kişiler her uygulama sorusundaki performansları için yanlış uygulamaya 0, kısmi doğru uygulamaya 1 ve doğru uygulamaya 2 puanla puanlanmıştır.

Kısmi puanlama kullanıldığı için, Kısmi Puanlama Madde Tepki Modelleri (Partial Credit Modeling) ile hesaplanan beş ayrı model kullanılmıştır. İlk model hiçbir yordayıcı değişken olmadan, yani geleneksel kısmi puanlama Madde Tepki Modelleri ile hesaplanmıştır. İkinci model uygulama sorusunun sırası, üçüncü model uygulama sorusuna ne kadar zaman harcandığı, dördüncü model hekim adayının cinsiyeti ve beşinci model hekim adayının son aşama sınavının çoktan seçmeli sorulardan oluşan kısmından aldığı puanı yordayıcı olarak kullanarak hesaplanmıştır. Her yordayıcının faydalılığını test etmek için her bir Açıklayıcı Madde Tepki Modeli için hesaplanan veriye uygunluk indeksleri geleneksel Madde Tepki Modeli için hesaplanan indeksleri ile karşılaştırılmıştır.

Bulgular: Model uygunluk indeksleri çoktan seçmeli bölümden alınan test puanının iyi bir yordayıcı olduğunu göstermektedir. Uygulama sorusunun hangi sırayla cevaplandığı, uygulama sorusuna harcanan toplam zaman ve hekim adayının cinsiyeti yordayıcı olarak faydalı bulunmamıştır. Karşılaştırıldığında Madde Tepki Modeli ve çoktan seçmeli test puanı ile hesaplanan Açıklayıcı Madde Tepki Modelli ile hesaplanan madde eşik değerlerini kullanarak elde edilen figürler açıkça göstermektedir ki iyi bir yordayıcı ile kurulan bir Açıklayıcı Madde Tepki Modeli madde istatistikleri ile kişilerin beceri düzeyleri arasındaki fonksiyonel ilişkiyi iyi yönde değiştirebilecektir.

Öneriler: Uzmanlar kişilerin bilgi ve becerilerini ortaya koyabilecekleri uygulama sınavlarının, çoktan seçmeli sınavlara birçok bakımdan üstün olduğunu düşünürler. Ancak uygulama sınavları ile elde edilen test puanlarının güvenirliği çoktan seçmeli sınavlarla karşılaştırıldığında genellikle düşüktür. Test güvenirliğini arttırmanın en olağan yolu olan madde sayısını arttırma uygulama sınavları için çok kolay olmamaktadır. Uygulama sorularını geliştirmek, uygulamak ve puanlamak oldukça emek yoğun ve pahalı olabilmektedir. Test maddeleri artırılamıyorsa, bir alternatif uygulama elde bulunan ek verilerin yapılan model tahminlerinde kullanılması olabilir. Bu çalışma böylesi bir yaklaşımla yapılmıştır.

Bulgular göstermektedir ki geleneksel Madde Tepki Modeli uygulandığında kabul edilebilir veriye uygunluk indeksleri ve güvenilir madde istatistikleri elde etmede güçlük çeken uygulama testleri Açıklayıcı Madde Tepki Modellerinin uygulamalarından yararlanabilir. Bu araştırmaya konu olan CCS uygulama testi için alınan sonuçlar göstermektedir ki ikincil değişkenlerin sağlayacağı ek bilgi, bu bilgi olmadan elde edilecek tahminleri iyi yönde değiştirecektir. Elbette Açıklayıcı Madde Tepki Model'inin başarılı olması için ikincil verilerin elde bulunması ve modele eklenmesi başlı başına yeterli olmayacaktır. Bu ikincil değişkenlerin katkısının ne olacağı bu araştırmada da kullanılan aşamalı bir yaklaşım ile ayrı ayrı değerlendirilmelidir. Açıklayıcı Madde Tepki Model uygulamaları kullanıcılara farklı model geliştirme imkanı da sunmaktadır. Örneğin, araştırmacılar, eldeki veriler uygun olduğunda, birden fazla ikincil değişkenin de dahil edilebileceği alternatif modeller ile interaksiyon ihtimallerini de kolayca çalışabilirler.

Anahtar Sözcükler: Kısmı Puan Modeli, Madde Tepki Modeli, uygulama testleri, madde istatistikleri, başarı tahmini

APPENDIX A. SAS SYNTAX

```
/* Read in*/
data CCS;
infile "H:\CCS\DATA\SASdataIN.dat";
INPUT per index y3 I1 I2 I3 I4 I5 I6 niseq time MC male;
RUN;

/* Estimate*/
/* Model 1 - PCM no covariates, CCS data - PCM three categories 0-2, */
PROC NLMIXED data=CCS method=gauss technique=quanew noad qpoints=10;
PARMS  b1_1-b1_6=0 b2_1-b2_6=0 sd=0.5;
beta1=b1_1*I1+b1_2*I2+b1_3*I3+b1_4*I4+b1_5*I5+b1_6*I6;
beta2=b2_1*I1+b2_2*I2+b2_3*I3+b2_4*I4+b2_5*I5+b2_6*I6;
        exp1=exp(theta-beta1);
        exp2=exp(2*theta-beta1-beta2);
denom=1+exp1+exp2;
if (y3=0) then p=1/denom;
else if (y3=1) then p=exp1/denom;
else if (y3=2) then p=exp2/denom;
if (p>1e-8) then ll=log(p);
else ll=-1e100;
Model y3~general(ll);
```

```
RANDOM theta~normal(0,sd**2)subject=per;
ESTIMATE 'sd**2' sd**2;
RUN;

/* Model 2 - PCM with item sequence covariate, CCS data - PCM three categories: 0-2 */
PROC NLMIXED data=CCS method=gauss technique=quanew noad qpoints=10;
PARMS  b1_1-b1_6=0 b2_1-b2_6=0 ts=0 sd=0.5;
              theta=eps+ts*niseq;
beta1=b1_1*I1+b1_2*I2+b1_3*I3+b1_4*I4+b1_5*I5+b1_6*I6;
beta2=b2_1*I1+b2_2*I2+b2_3*I3+b2_4*I4+b2_5*I5+b2_6*I6;
            exp1=exp(theta-beta1);
            exp2=exp(2*theta-beta1-beta2);
            denom=1+exp1+exp2;
if (y3=0) then p=1/denom;
else if (y3=1) then p=exp1/denom;
else if (y3=2) then p=exp2/denom;
if (p>1e-8) then ll=log(p);
else ll=-1e100;
Model y3~general(ll);
RANDOM eps~normal(0,sd**2)subject=per;
ESTIMATE 'sd**2' sd**2;
RUN;

/* Model 3 - PCM with response time covariate, CCS data - PCM three categories: 0-2 */
PROC NLMIXED data=CCS method=gauss technique=quanew noad qpoints=10;
PARMS  b1_1-b1_6=0 b2_1-b2_6=0 ti=0 sd=0.5;
              theta=eps+ti*time;
beta1=b1_1*I1+b1_2*I2+b1_3*I3+b1_4*I4+b1_5*I5+b1_6*I6;
beta2=b2_1*I1+b2_2*I2+b2_3*I3+b2_4*I4+b2_5*I5+b2_6*I6;
            exp1=exp(theta-beta1);
            exp2=exp(2*theta-beta1-beta2);
            denom=1+exp1+exp2;
if (y3=0) then p=1/denom;
else if (y3=1) then p=exp1/denom;
else if (y3=2) then p=exp2/denom;
```

```
if (p>1e-8) then ll=log(p);
else ll=-1e100;
Model y3~general(ll);
RANDOM eps~normal(0,sd**2)subject=per;
ESTIMATE 'sd**2' sd**2;
RUN;
```

/* Model 4 - PCM with gender covariate: male coded as 1, CCS data, PCM three categories: 0-2*/

```
PROC NLMIXED data=CCS method=gauss technique=quanew noad qpoints=10;
PARMS  b1_1-b1_6=0 b2_1-b2_6=0 g=0 sd=0.5;
          theta=eps+g*male;
beta1=b1_1*I1+b1_2*I2+b1_3*I3+b1_4*I4+b1_5*I5+b1_6*I6;
beta2=b2_1*I1+b2_2*I2+b2_3*I3+b2_4*I4+b2_5*I5+b2_6*I6;
          exp1=exp(theta-beta1);
          exp2=exp(2*theta-beta1-beta2);
          denom=1+exp1+exp2;
if (y3=0) then p=1/denom;
else if (y3=1) then p=exp1/denom;
else if (y3=2) then p=exp2/denom;
if (p>1e-8) then ll=log(p);
else ll=-1e100;
Model y3~general(ll);
RANDOM eps~normal(0,sd**2)subject=per;
ESTIMATE 'sd**2' sd**2;
RUN;
```

/* Model 5 - PCM with MCT Scores as a person covariate, CCS data - PCM three categories: 0-2, */

```
PROC NLMIXED data=CCS method=gauss technique=quanew noad qpoints=10;
PARMS  b1_1-b1_6=0 b2_1-b2_6=0 t=0 sd=0.5;
          theta=eps+t*MC;
beta1=b1_1*I1+b1_2*I2+b1_3*I3+b1_4*I4+b1_5*I5+b1_6*I6;
beta2=b2_1*I1+b2_2*I2+b2_3*I3+b2_4*I4+b2_5*I5+b2_6*I6;
          exp1=exp(theta-beta1);
          exp2=exp(2*theta-beta1-beta2);
```

```
        denom=1+exp1+exp2;
if (y3=0) then p=1/denom;
else if (y3=1) then p=exp1/denom;
else if (y3=2) then p=exp2/denom;
if (p>1e-8) then ll=log(p);
else ll=-1e100;
Model y3~general(ll);
RANDOM eps~normal(0,sd**2)subject=per;
ESTIMATE 'sd**2' sd**2;
RUN;
```