

ITEM DEPENDENCY IN AN  
OBJECTIVE STRUCTURED CLINICAL EXAMINATION

Cherdsak Iramaneerat

Carol M Myford

Rachel Yudkowsky

University of Illinois at Chicago

Paper presented at the 25<sup>th</sup> International Objective Measurement Workshop,

Berkeley, CA, April 2006

ITEM DEPENDENCY IN AN  
OBJECTIVE STRUCTURED CLINICAL EXAMINATION

Abstract

An Objective Structured Clinical Examination (OSCE) is an assessment approach employed in medical education, in which residents rotate through multiple stations of standardized clinical tasks to evaluate their clinical competence. Because items used to evaluate residents' performance in each OSCE station are linked to the same task and are rated by the same rater, their ratings may be dependent on one another, violating the assumption of conditional item independence that underlies the multi-faceted Rasch measurement (MFRM) model. We employed a MFRM model to analyze a communication skills assessment of 79 residents, using 6 OSCE stations, each scored on 18 five-point rating scale items. When we treated the rating on each item as a separate scoring unit, MFRM analyses showed item dependency in 65% of item pairs within an OSCE station according to Fisher's  $Z$  statistic, a modification of Yen's  $Q_3$  index of item dependency. This resulted in overestimation of resident separation reliability and inaccurate parameter estimation. Combining item scores in each OSCE station into a station score and using station as a scoring unit reduced the amount of item dependency to 27%. This approach produced more realistic reliability estimates and helped improve the fit of the data to the model.

## ITEM DEPENDENCY IN AN OBJECTIVE STRUCTURED CLINICAL EXAMINATION

An Objective Structured Clinical Examination (OSCE) is an assessment approach used in medical education, in which clinical competence is evaluated in a comprehensive, consistent, and structured manner, using an examination format that instructs examinees to rotate through a circuit of stations of clinical tasks (Harden, 1988; Harden, Stevenson, Downie, & Wilson, 1975; Van der Vleuten & Swanson, 1990). Standardized patients (SPs), who are lay persons trained to portray a scripted patient presentation in a standardized and consistent fashion, are commonly used in OSCE stations (Van der Vleuten & Swanson, 1990; Yudkowsky, Alseidi, & Cintron, 2004). At the end of each encounter with an examinee, the SP provides ratings of an examinee's performance on a rating form.

Ideally, items included in a rating form should function independently of one another (i.e., an examinee's rating on each item should be independent of the ratings the SP assigned the examinee on other items). However, items used to evaluate examinee performance in each OSCE station are linked to the same clinical task and are rated by the same SP. An examinee's level of performance on one item may be dependent on his/her level of performance on one or more items in the same rating form. That is, inadequate performance on one item may cue inadequate performance on one or more items in the same station. Such item dependency violates the assumption of conditional item independence that underlies both classical true score and item response theories. Item dependence can lead to inaccurate estimation of item parameters, test statistics, and examinee competency (Sireci, Thissen, & Wainer, 1991; Zenisky, Hambleton, & Sireci, 2003). When dependent items are improperly treated as independent items, reliability and test information functions are overestimated (Sireci et al., 1991; Thissen, Steinberg, & Mooney,

1989). In addition, local item dependence introduces an additional dimension into the test at the expense of the construct of interest (Wainer & Thissen, 1996).

A multi-faceted Rasch measurement (MFRM) approach (Linacre, 1989) is a psychometric approach suitable for the analysis of rater-mediated assessment data. It is an extension of the basic Rasch model that allows simultaneous calibration of multiple facets that may exert an influence on the measurement of examinee performance (e.g., rater severity, task difficulty) by adding to the model parameters describing these facets of measurement interest. A MFRM approach attempts to free each examinee's performance measure from the effects of differences in rater severity or task difficulty (Linacre & Wright, 2004). MFRM is an appropriate psychometric approach for analyzing data from an OSCE to estimate examinees' performance measures that are free from systematic rater severity error. However, when analyzing rating data with a MFRM model, researchers generally focus on the study of rater effects and often overlook the issue of local dependence of items that are linked to the same rater or to the same task. The purposes of this study were: (1) to check for the existence of item dependency in an OSCE, (2) to outline an alternative approach for analyzing rating data using a MFRM model to ameliorate the problem of item dependency, and (3) to compare separation reliability estimates, parameter estimates, and fit statistics obtained from MFRM analyses when the assumption of local independence was violated and when the assumption was met.

## Method

### *Participants*

We examined an OSCE for communication skills assessment of 79 residents from one Midwestern medical school in the United States. Sixty-eight of them were internal medicine

residents, and 11 of them were family medicine residents. From 70 records with gender identification, there were 39 (66%) male and 20 (34%) female internal medicine residents, and five (45%) male and six (55%) female family medicine residents.

### *Tasks (OSCE stations)*

This assessment measures communication skills competence using six OSCE stations. Each station was a simulated clinical scenario asking a resident to perform a specific communication task. The six tasks were: (1) providing patient education, (2) obtaining informed consent, (3) dealing with a patient who refuses treatment, (4) counseling an elderly patient who has been abused, (5) giving bad news to a patient, and (6) conducting a physical examination. Only one SP portrayed each case throughout the assessment.

### *Rating scale*

The SPs rated the performance of each resident at the end of each encounter using a standard rating form which was composed of 18 items, each asking the SP for his/her level of agreement with a statement about the resident's communication skills, using a five-point scale ranging from one (*strongly disagree*) to five (*strongly agree*). All items were phrased positively; thus, higher ratings indicated better performance.

### *Analyses*

We first analyzed the data using a three-faceted model, which took the form:

$$\ln[P_{nij} / P_{nij(k-1)}] = B_n - D_i - C_j - F_{ik} \quad (1)$$

where  $P_{nij}$  is the probability of resident  $n$  receiving a rating of  $k$  on item  $i$  in OSCE station  $j$ ,

$P_{nij(k-1)}$  is the probability of resident  $n$  receiving a rating of  $k-1$  on item  $i$  in OSCE station  $j$ ,

$B_n$  is the level of communication competence of resident  $n$ ,

$D_i$  is the difficulty of item  $i$ ,

$C_j$  is the difficulty of OSCE station  $j$ , and

$F_{ik}$  is the difficulty of receiving a rating of  $k$  relative to a rating of  $k-1$  for each item.

This MFRM model provided measures of resident communication competence adjusted for the systematic differences in OSCE station difficulty and item difficulty. This model operated under the assumption that items were independent of one another (i.e., when a SP rated resident  $n$  on item  $i$ , the SP was not influenced by the ratings he/she assigned that same resident on any other items appearing on the rating form for that OSCE station). In item response theory, when a pair of items is locally independent, they will have zero residual correlations, conditioning on (after accounting for) the latent trait of interest (Sitjima & Molenaar, 2002; Yen, 1993).

A commonly used method for assessing item dependence is Yen's  $Q_3$  statistic (Yen, 1984, 1993), which is the correlation of the residuals for a pair of items after partialling out the latent trait estimate. In this study, we assessed local item dependency using the standardized Fisher's  $Z$  approach, a modification of Yen's  $Q_3$  index of local item dependence (Shen, 1996). Fisher's  $Z$  approach modified  $Q_3$  in two aspects: (1) it adjusted the residuals by the accuracy of the resident communication competence measure, and (2) it established a practical significance level for item dependency (Shen & Yen, 1997). The computation of Fisher's  $Z$  involved three steps:

- (1) Calculate the standardized residuals for each rating of resident  $n$  on item  $i$ :

$$d_{ni} = (\text{observed rating} - \text{expected rating})/SE_n \quad (2)$$

- (2) Correlate the standardized residuals,  $d_{ni}$ , for all pairs of items  $i, j$  in each OSCE station across all residents.

(3) Compute Fisher's  $Z$  statistic to normalize the Pearson correlation, using the formula:

$$Z_{ij} = \frac{1}{2} \log \frac{(1 + r_{ij})}{(1 - r_{ij})} \quad (3)$$

We considered a pair of items to be significantly dependent if their Fisher's  $Z$  statistic was more than 2 standard deviations above the mean of Fisher's  $Z$  for practically independent items or less than 2 standard deviations below the mean of Fisher's  $Z$  for practically independent items. Practically independent items were items on the same examination that were independently developed and did not link to a common task or SP. In this study, we randomly selected 195 pairs of different items from different OSCE stations to serve as a sample of practically independent items. We determined the extent of item dependency by checking the mean Fisher's  $Z$  statistic and the percentage of item pairs with significant Fisher's  $Z$  statistics.

After exploring the extent of item dependency in an OSCE with a traditional MFRM analysis, we suggested an alternative approach for analyzing the data to ameliorate the item dependency problem. One method that researchers have used successfully in the analysis of multiple-choice dichotomous items was to group a series of dichotomously scored dependent responses into a single polytomous response and then analyze the data using polytomous IRT models, treating each group of dependent items as one scoring unit (Thissen et al., 1989; Yen, 1993). Our study elaborated this notion by treating each OSCE station as a scoring unit. In other words, for each resident, we averaged the ratings the SP assigned to all items for a given OSCE station to produce a *station score*, which we considered as one rating in the MFRM analysis. Thus, each resident would have only six ratings, instead of 108 ratings, for the analysis. To avoid having station scores with decimal places, we multiplied the averaged ratings of each OSCE station by ten to produce station scores in integers ranging from 10 (*poor performance*) to 50 (*excellent performance*). We analyzed station scores using the following MFRM model:

$$\ln[P_{nj(k)} / P_{nj(k-1)}] = B_n - C_j - F_{jk} \quad (4)$$

where  $P_{nj(k)}$  is the probability of resident  $n$  receiving a station score of  $k$  in OSCE station  $j$ ,

$P_{nj(k-1)}$  is the probability of resident  $n$  receiving a station score of  $k-1$  in OSCE station  $j$ ,

$B_n$  is the level of communication competence of resident  $n$ ,

$C_j$  is the difficulty of OSCE station  $j$ , and

$F_{jk}$  is the difficulty of receiving a station score of  $k$  relative to a score of  $k-1$  for each OSCE station.

We determined whether this approach helped resolve the item dependency problem by checking the mean Fisher's  $Z$  statistic, as well as the percentage of item pairs with significant Fisher's  $Z$  statistics. We then compared resident separation reliability estimates, parameter estimates, and fit statistics obtained from the two MFRM analyses. We considered residents or stations to be overfitting when their infit or outfit mean-square values were below 0.4 or their standardized infit or outfit values were below -2.0. We considered residents or stations to be underfitting when their infit or outfit mean-square values were above 1.2 or their standardized infit or outfit values were above 2.0 (Linacre, 2002; Wright & Linacre, 1994).

## Results

### *Item dependency*

Practically independent items had a mean Fisher's  $Z$  item dependency index of -0.03 with a standard deviation of 0.05. Three percent of practically independent item pairs had significant Fisher's  $Z$  statistics. Our first MFRM analysis in which we considered each item as a scoring unit yielded a mean Fisher's  $Z$  statistic of 0.12, with a standard deviation of 0.10. Sixty-five percent of the item pairs had significant Fisher's  $Z$  statistics. By contrast, our second MFRM analysis in

which we used station scores as scoring units yielded a mean Fisher's  $Z$  statistic of  $-0.09$ , with a standard deviation of  $0.08$ . The number of item pairs with significant Fisher's  $Z$  statistics was reduced to 27% (Table 1).

[INSERT TABLE 1 ABOUT HERE]

#### *Resident separation reliability*

Our first MFRM analysis in which we used items as scoring units yielded a resident separation reliability of  $0.94$ . On the other hand, when we considered stations as scoring units to ameliorate the problems of item dependency, our MFRM analysis yielded a resident separation reliability of  $0.74$ . Our improper treatment of dependent items as independent items in the initial MFRM analysis resulted in an overestimation of reliability by 27%.

#### *Resident communication competency estimates*

Our first MFRM analysis in which we used items as scoring units provided resident communication competency measures ranging from  $0.20$  to  $2.65$  logits with a mean of  $1.39$  and a standard deviation of  $0.52$ . On the other hand, when we considered stations as scoring units, our MFRM analysis provided resident communication competency measures ranging from  $-0.37$  to  $0.68$  logits with a mean of  $0.13$  and a standard deviation of  $0.23$ . Although resident communication competency measures we obtained from the second analysis were lower and exhibited less spread than those we obtained from our first analysis, the two sets of resident communication competence measures were highly correlated with a Pearson correlation of  $1.00$ .

When we used items as scoring units, 15 and 19 residents were underfitting according to their infit and outfit mean-square values, respectively. On the other hand, when we used station

scores as scoring units, we detected 27 underfitting residents and 13 overfitting residents using infit mean-square values, and 25 underfitting residents and 13 overfitting residents using outfit mean-square values. However, according to their standardized fit statistics, the first analysis yielded 8 overfitting and 8 underfitting residents (from infit) and 10 overfitting and 14 underfitting residents (from outfit). By contrast, the second analysis yielded 4 overfitting and 2 underfitting residents (from infit) and 1 overfitting and 5 underfitting residents (from outfit) (Table 2). The first analysis provided resident communication competency measures that better fit the measurement model according to unstandardized fit statistics, but the second analysis provided resident communication competency measures that better fit the measurement model according to standardized fit statistics.

[INSERT TABLE 2 ABOUT HERE]

#### *Station difficulty estimates*

Our first MFRM analysis in which we used items as scoring units provided station difficulty measures ranging from -1.46 to 0.66 logits with a mean of 0 and a standard deviation of 0.86. When we considered stations as scoring units, the station difficulty measures ranged from -0.22 to 0.14 logits with a mean of 0 and a standard deviation of 0.14. Despite the differences in the station difficulty measures, the two analyses resulted in very similar station difficulty ordering. Station 2 (informed consent) became relatively more difficult, while Station 1 (patient education) became relatively easier in the second analysis. All the other four stations maintained their same order of difficulty.

When we used items as scoring units, 2 and 3 stations were underfitting according to their infit and outfit mean-square values, respectively. On the other hand, when we used station scores

as scoring units, we found no misfit according to infit mean-square values, but 2 stations were underfitting according to their outfit mean-square values. From standardized infit and outfit statistics, the first analysis yielded 3 overfitting and 3 underfitting stations. On the other hand, all six stations fit well in the second analysis. The second analysis provided station difficulty estimates that better fit the measurement model according to both unstandardized and standardized fit statistics (Table 2).

### Discussion

This study demonstrated the importance of checking for violation of the item independence assumption when conducting MFRM analyses of rating data that have many items linked to the same task or the same rater. Conducting MFRM analyses using dependent items as separate scoring units is a violation of a basic psychometric assumption of the model that may result in overestimation of separation reliability estimates. In addition, in our case, conducting such analyses also resulted in resident communication competency measures exhibiting poorer fit to the measurement model according to standardized fit statistics, as well as station difficulty measures that exhibited poorer fit to the model according to both standardized and unstandardized fit statistics. Although our use of station scores in MFRM analysis helped alleviate an item dependency problem, one should not automatically use station scores without first ascertaining whether item dependency is present. Combining ratings from multiple items into a station score resulted in loss of information and a decrease in resident separation reliability. We recommend that one first evaluate the extent of item dependency in rating data. If significant item dependency is present, combining item scores to produce station scores and using them as scoring units should help alleviate the problem.

## REFERENCES

- Harden, R. M. (1988). What is an OSCE? *Medical Teacher*, 10(1), 19-22.
- Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1(5955), 447-451.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J. M., & Wright, B. D. (2004). Construction of measures from many-facet data. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 296-321). Maple Grove, MN: JAM Press.
- Shen, L. (1996). Quantifying item dependency. *Rasch Measurement Transactions*, 10, 485.
- Shen, L., & Yen, J. (1997). Item dependency in medical licensing examinations. *Academic Medicine*, 22(10) (Suppl. 1), S19 - S21.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237-247.
- Sitjsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage Publications.
- Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, 26(3), 247-260.
- Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, 2(2), 58-76.

- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15(1), 22-29.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187-213.
- Yudkowsky, R., Alseidi, A., & Cintron, J. (2004). Beyond fulfilling the core competencies: An objective structured clinical examination to assess communication and interpersonal skills in a surgical residency. *Current Surgery*, 61(5), 499-503.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2003). *Effects of local item dependence on the validity of IRT item, test, and ability statistics*. Washington, DC: Association of American Medical Colleges.

Table 1

*Fisher's Z Index of Item Dependency*

Type of items	Mean	SD	Items with significant
			Fisher's Z statistics (%)
Independent items	-0.03	0.05	3
item scores	0.12*	0.10	65
station scores	-0.09*	0.08	27

\* Significantly different from the mean Fisher's Z of independent items ( $p < 0.05$ )

Table 2

*Numbers of Misfitting Residents and Stations*

Scoring unit Criteria	Item scores		Station scores	
	Underfitting	Overfitting	Underfitting	Overfitting
<b>Resident (N=79)</b>				
Infit MnSq	15	0	27	13
Infit Zstd	8	8	2	4
Outfit MnSq	19	0	25	13
Outfit Zstd	14	10	5	1
<b>Station (N=6)</b>				
Infit MnSq	2	0	0	0
Infit Zstd	3	3	0	0
Outfit MnSq	3	0	2	0
Outfit Zstd	3	3	0	0