# RATER EFFECTS IN CLINICAL PERFORMANCE RATINGS OF SURGERY RESIDENTS

Cherdsak Iramaneerat

Carol M. Myford

Department of Educational Psychology, College of Education

University of Illinois at Chicago

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April 2006.

Correspondence concerning this article should be addressed to Cherdsak Iramaneerat, 1407 S. Indiana Ave. Chicago, IL 60605 USA, Email: cirama1@uic.edu

# RATER EFFECTS IN CLINICAL PERFORMANCE RATINGS OF SURGERY RESIDENTS


## Abstract

A multi-faceted Rasch measurement (MFRM) approach was used to analyze clinical performance ratings of 24 first-year residents in one surgery residency program in Thailand to investigate three types of rater effects: leniency, rater inconsistency, and restriction of range. Faculty from 14 surgical services rated the clinical performance of residents using an 11-item rating instrument. We analyzed the ratings using a three-faceted Rasch model that defines the probability of a particular resident receiving a particular rating as a function of the level of clinical performance of a resident, the difficulty of an item, and the severity of a clinical service. Faculty in 14 clinical services showed significant differences in severity. Faculty in the intensive care unit gave inconsistent ratings, while faculty in the trauma and minor operating room services showed restriction of score range. We recommended using a MFRM approach in analyzing clinical performance ratings. A MFRM analysis not only provides measures of residents' clinical performance that have been adjusted for systematic difference in clinical services severity, but also identifies specific clinical services that exhibit aberrant rating behaviors. The residency program can use this diagnostic information to help determine what changes are needed in the approach to evaluate resident performance.

(Contains 16 references, 1 table, and 2 figures)

# RATER EFFECTS IN CLINICAL PERFORMANCE RATINGS OF SURGERY RESIDENTS

Among the many methods used in evaluating surgical residents' performance, having attending faculty rate clinical performance is the most widely employed method (Kwolek et al., 1997; Sloan, Donnelly, Drake, & Schwartz, 1995). The use of ratings rests on the assumption that faculty are capable of some degree of precision and objectivity. Nevertheless, faculty can introduce various systematic and unsystematic sources of variance into performance ratings that are associated with their own rating behavior and not with the actual performance of residents. These sources of construct-irrelevant variance are collectively called *rater effects* (Scullen, Mount, & Goff, 2000). Making valid inferences from clinical performance ratings requires monitoring and controlling of these rater effects.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and multi-faceted Rasch measurement (MFRM) (Linacre, 1989) are two statistical approaches for detecting and measuring rater effects. Generalizability theory is an extension of reliability in classical test theory. Researchers investigating rater effects have used it widely to determine the amount of variance in ratings that can be attributed to raters as compared to other sources. Generalizability theory provides a useful method for estimating various sources of error in a rating operation, but it does not offer any method for adjusting ratings to eliminate rater errors (Linacre, 1996; Shavelson, Webb, & Rowley, 1989).

Multi-faceted Rasch measurement is an extension of the basic Rasch model (i.e., a measurement model for simultaneous calibration of person ability and item difficulty on the same scale). The basic Rasch model for dichotomous data uses only one parameter, item difficulty, to estimate person ability. The multi-faceted Rasch measurement model extends the

basic Rasch model to also measure the severity of the raters and/or the difficulty of the tasks involved in the measurement process. MFRM not only allows us to monitor the effects of raters, but it also offers a way to effectively adjust ratings for systematic rater severity error (Downing, 2005). Within the context of assessing residents' clinical performance, MFRM considers each resident and faculty at the individual level and attempts to liberate each resident's clinical performance measure from the effects of differences in faculty severity (Linacre & Wright, 2004). MFRM can also help determine whether individual faculty are exhibiting various types of rater effects besides severity (Myford & Wolfe, 2003, 2004). In addition, MFRM can provide bias analyses to uncover interactions involving a rater and other aspects of a rating operation (McNamara, 1996).

Despite its demonstrated utility for investigating rater effects, researchers have not used MFRM to study rater effects in clinical performance ratings of surgery residents. This study applied a MFRM approach to analyze clinical performance ratings of residents in the Department of Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand. The purpose of this study was to investigate the applicability of a MFRM approach for investigating rater effects in clinical performance ratings. Specifically, this study addressed the following questions:

1. How well could clinical services faculty differentiate residents based on their clinical performance?
2. Did some of the clinical services faculty rate more severely than others?
3. Did each clinical service consistently employ the rating scales? Did ratings by any of the clinical services show evidence of restriction of range?
4. How did clinical services faculty use the rating scale for each item?

## Method

### *Residents*

We examined clinical performance ratings of 24 first-year residents who worked in the Department of Surgery, Faculty of Medicine Siriraj Hospital in the year 2001 to 2002. Two of them are female, and 22 are male. This group of residents is a mix of residents from various specialties: ten from general surgery, four from orthopedic surgery, three from urology, three from cardiothoracic surgery, two from neurosurgery, and two from pediatric surgery.

### *Clinical services*

Each resident worked in 13 different clinical services, each for a period of four weeks. Eleven clinical services were required for everyone: pediatric surgery, urology, minor operating room, neurosurgery, plastic surgery, intensive care unit, anesthesiology, orthopedic, cardiothoracic surgery, trauma, and head-neck-breast surgery. The other two services were general surgery. Residents were randomly assigned to two of the three services: general A, B, or C. At the end of each four-week period, faculty members within each service rated the clinical performance of residents they supervised using a standard rating form. The Department of Surgery allowed each service to use its own discretion in choosing the appropriate faculty member(s) to assign ratings (i.e., different faculty members within that clinical service might assign ratings for each rotation period). Therefore, for the purposes of this study, we considered each clinical service as one rater unit.

### *Rating form*

Each clinical service rated residents' clinical performance on 11 items: (1) knowledge, (2) clinical judgment, (3) curiosity, (4) industriousness, (5) effectiveness, (6) medical records,

(7) punctuality, (8) responsibility, (9) relationship with patients, (10) relationship with other doctors, and (11) relationship with other healthcare workers. Faculty assigned ratings using a continuous scale ranging from one to five, with one representing a poor performance, two representing a fair performance, three representing an acceptable performance, four representing a good performance, and five representing an excellent performance.

*Rating procedure*

The clinical service rotation was set up so that the 13 different clinical services rated each of the 24 residents on 11 items. A total of 3,432 responses were expected. However, 11 responses were missing; thus, only 3,421 responses (99.7%) were analyzed.

*Analyses*

We conducted our analyses of clinical performance ratings with a multi-faceted version of the Rasch measurement model using the Facets computer program (Linacre, 2005). The MFRM model used in this study takes the form:

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

where $P_{nijk}$ is the probability of resident $n$ receiving a rating of $k$ on item $i$ from clinical service $j$

$P_{nij(k\text{-}1)}$ is the probability of resident $n$ receiving a rating of $k$-1 on item $i$ from clinical service $j$

$B_n$ is the level of clinical performance of resident $n$

$D_i$ is the difficulty of item $i$

$C_j$ is the severity of clinical service $j$, and

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

The MFRM model uses a logistic transformation of the observed ratings to a logit score. This model adjusts measures of resident clinical performance for the effects of item difficulty, clinical service (rater) severity, and the manner in which clinical services faculty used the rating scales. This model allowed each item to be calibrated to have its own rating scale structure (i.e., the rating scale category thresholds for individual items could be different). Because the model studies ratings as categorical variables, we multiplied the original continuous ratings by ten to convert continuous decimal ratings of one to five to categorical integer ratings of 10 to 50 before analyzing the data.

We used a reliability of separation index and a fixed-effect chi-square statistic to determine whether there were significant differences between residents in their levels of clinical performance, and clinical services in the levels of severity they exercised. A reliability of separation index is an indicator of how well the residents are separated by their performance (or how well the clinical services are separated by their levels of severity). This index can be interpreted in the same way as traditional indices of reliability, such as Cronbach's coefficient alpha and KR-20 are interpreted (Engelhard, 2002). For resident performance, high reliability indicates that the assessment can effectively separate residents according to their levels of clinical performance with a high degree of confidence. On the other hand, for clinical services severity, high reliability indicates that the clinical services faculty do not exercise the same level of severity when rating residents. The higher the reliability, the less interchangeable the clinical services.

A fixed-effect chi-square is a chi-square test for a null hypothesis that states that there are no significant differences between residents in terms of their clinical performance, or between clinical services in terms of their levels of severity. A significant chi-square value for resident

clinical performance measures indicates that at least two residents are statistically significantly different in their levels of clinical performance. A significant chi-square value for clinical service measures indicates that at least two clinical services exercised statistically significantly different degrees of rating severity. We interpreted all the statistical tests under the assumption of a Type I error rate of 0.05.

We monitored how consistently each clinical service used the rating scales by examining clinical services' infit mean-square values. Infit mean-square values are weighted mean squared residual statistics. The weighting reduces the influence of less informative, low variance, off-target responses (Wright & Masters, 1982). The expected value is one when the model fits the data. An infit mean-square value less than one indicates too little variation in the ratings, while a value more than one indicates too much variation in the ratings. Because this is a low-stakes assessment, we set lower- and upper-control limits at 0.5 and 2.0, respectively (Linacre, 2002).

We examined category probability curves to assess how faculty raters used the rating scale for each item. We determined the number of functioning categories for each item by counting the number of separate peaks that became the most probable rating for a clearly defined region along the clinical performance continuum[1].

## Results

The MFRM analysis yielded the variable map (Figure 1) that shows the calibrations of residents' clinical performance, clinical service severity, and item difficulty on an equal interval

[1] Since our focus was in the study of rater effects, we only evaluated the functioning of the rating scales with category probability curves. There are also other indicators of rating scale functioning. For those who are interested in detailed evaluation of rating scale functioning using these indicators, please refer to Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith, Jr. and R. M. Smith (Eds.), *Introduction to Rasch Measurement: Theory, models, and applications.* Maple Grove, MN: JAM Press.

logit scale. The scale was calibrated so that higher logit values indicated higher performing residents, more severe clinical services, and more difficult items. Residents' clinical performance measures ranged from -0.08 to 0.35 logits, having only one resident who had poorer performance than the average clinical service severity and item difficulty (which were calibrated to have mean values of zero logits). The urology service was the most severe clinical service (severity measure = 0.24 logits), providing the lowest average rating. On the other hand, the general A service was the most lenient clinical service (severity measure = -0.25 logits), providing the highest average rating. The item that was the most difficult to get high ratings on was Item 1 (knowledge), which had a difficulty measure of 0.38 logits. The easiest item was Item 4 (industriousness), which had a difficulty measure of -0.13 logits.

[INSERT FIGURE 1 ABOUT HERE]

*How well could clinical services faculty differentiate residents based on their clinical performance?*

Residents' clinical performance measures ranged from -0.08 to 0.35 logits, with a mean of 0.22 logits and a standard deviation of 0.09 logits. All residents' clinical performance measures conformed well to the model, as indicated by their infit mean-square values, which ranged from 0.50 to 1.38. The resident separation reliability was 0.91, which is statistically significant, $\chi^2$ (23) = 259.1, $p < 0.05$. The resident separation index was 4.64, which suggests that faculty in these clinical services could discern about five distinct groups of clinical performance among the 24 residents.

*Did some of the clinical services faculty rate more severely than others?*

Clinical service severity levels ranged from -0.25 logits for the general A service to 0.24 logits for the urology service, with a mean of 0 logits and a standard deviation of 0.12 logits (Table 1). The clinical services separation reliability was 0.97, which is statistically significant, $\chi^2$ (13) = 495.2, $p < 0.05$. The clinical services separation index was 8.25, suggesting that the 14 clinical services that provided the ratings exercised about eight statistically distinct levels of severity when rating the residents' performance.

[INSERT TABLE 1 ABOUT HERE]

*Did each clinical service consistently employ the rating scales? Did ratings by any of the clinical services show evidence of restriction of range?*

Most clinical services used the rating scales consistently, having infit mean-square values within the control limits (between 0.5 and 2.0). The intensive care unit was the only clinical service that rated somewhat inconsistently (infit mean-square value = 2.21). Ratings by trauma and minor operating room services exhibited restriction of range (infit mean-square values = 0.42 and 0.31, respectively). Our review of raw ratings confirmed the restriction of range in these two clinical services. Trauma service gave 92% of their ratings for good performance, while only 4% were for acceptable performance, and another 4% were for excellent performance, with no ratings for poor or fair performance. Minor operating room service gave 85% of their ratings for good performance, while only 15% were for excellent performance, with no ratings for poor, fair, or acceptable performance.

*How did clinical services faculty use the rating scale for each item?*

All items, except the first item (knowledge), functioned as 4-category rating scales, as indicated by their category probability curves. That is, the graph for each item exhibited four separate category curves that become the most probable rating for some particular range of the resident clinical performance continuum (Figure 2A). By contrast, the graph for Item 1 exhibited only three separate category probability curves, suggesting that this item functioned as a 3-category rating scale (Figure 2B).

[INSERT FIGURE 2 ABOUT HERE]

## Discussion

This study demonstrated the utility of a MFRM approach for detecting and measuring rater effects in clinical performance ratings of surgery residents. We identified clinical services that displayed evidence of exhibiting a variety of rater effects in their ratings, including severity, inconsistency, and restriction of range.

The statistically significant results from the chi-square test performed on the clinical services severity measures suggest that some clinical services were more severe in assigning ratings than others. The differences in severity levels between clinical services could introduce construct-irrelevant variance into the measures of residents' clinical performance, especially in this study where different residents worked in different general surgery services. Residents who worked in general A would have an unfair advantage over residents who worked in general B or general C, if we relied on unadjusted raw ratings. To obtain more objective measures of residents' clinical performance that are fairer to every resident, we should take into consideration the differences in severity levels between clinical services. Because the MFRM model measures

residents' clinical performance on the same scale with clinical services' severity measures, the model can adjust the residents' clinical performance measures for clinical services' severity differences (i.e., providing the score each resident would receive had a clinical service of average severity rated that resident), thus producing measures that are fairer to everyone.

The intensive care unit employed the rating scales in an inconsistent manner. Intensive care unit faculty may have applied different rating standards to different residents, providing many unexpected ratings. This resulted in a poor fit of the data to the measurement model. On the other hand, we found evidence of restriction of range in the ratings of the trauma and minor operating room services. Unlike the intensive care service, these two services employed the rating scales in a too consistent manner that resulted in an overfitting of their ratings to the measurement model. When we inspected the frequency distributions of ratings assigned, we found that these two services only used the high categories on the rating scale when evaluating residents, neglecting to use low rating categories. Such restriction of the range of ratings given to the residents would consequently limit the ability of the ratings to differentiate residents with good performance from those with poor performance, which was the main purpose of the clinical performance ratings.

The study of the functioning of the rating scale helped us understand how clinical services faculty interpreted the rating categories for each item. Despite defining five rating categories, none of the 11 items had a functioning five-category scale. Ten items had four functioning categories, and one item (Item 1: knowledge) had only three functioning categories.

Our findings suggested that the rating operation was not working as well as it could. Clinical services faculty were not functioning in an interchangeable manner. Clinical services

faculty exhibited a variety of rater effects in their ratings, including severity, inconsistency, and restriction of range. Given such findings, what could be done to improve this rating operation?

One strategy might be to work with clinical services faculty to help them become aware of the rater errors they were committing. For example, faculty in the trauma and minor operating room services exhibited restriction of range in their ratings. They might benefit from feedback that helps them see that they only used a limited range of rating categories. Engaging them in discussions of behaviors they might observe in their services that would fall into the rating categories they were not using might enable them to feel more comfortable using the full range of rating scale categories.

While providing feedback might prove useful for dealing with some rater errors, there are other strategies that might be employed to improve this rating operation, as well. We may address some rater errors by working to improve the items. For example, we found that intensive care unit faculty tended to rate inconsistently. Intensive care faculty may have a different interpretation of what the items mean from faculty in other services. The lack of operational definitions of items on the rating form can lead to different services using different indicators to assess performance on each item. Faculty might rate more consistently if each item was clearly defined. Clinical services faculty might benefit from engaging in a discussion of how they interpret these items. Through discussions, they could share their individualistic interpretations with the goal of arriving at a common understanding of each item's meaning. The rating form could then be revised to include an agreed upon definition for each item to be rated.

Another strategy for improving this rating operation might be to work on revising the rating scale. In this study, we found that faculty did not use the 5-category scale as a 5-category scale. This finding suggested that it might be advisable to reduce the number of rating categories

included in the clinical performance rating form. Alternatively, if clinical services faculty feel strongly about retaining all five rating categories, then it may be worthwhile to create a better differentiated, more explicit rating scale for each item. Faculty could engage in discussions about what specific behaviors they would expect to see at each rating category of each item. Through discussions, faculty should identify specific indicators of poor, fair, acceptable, good, and excellent performance within their clinical service for each item. Faculty could then incorporate these indicators into the description of each rating category. In the process of defining the rating scale categories, faculty might decide that some items need fewer categories than five, while some may need more than five. Accordingly, the revised rating form might have items with differing numbers of categories on the scales, depending upon how many clear levels of performance the faculty could define for each item.

*Implications*

In an incompletely crossed rating design like the one in this study, in which different sets of clinical services rated each resident, the MFRM model adjusts each resident's clinical performance measure to control for the effects of systematic clinical services variance that is due to differences in the levels of severity that the clinical services exercised. In effect, the measurement model can "wash out" the effects of this particular rater error on resident measures so that the resulting measures reflect the scores that residents would have received had clinical services of average severity rated each resident.

For other rater errors that a MFRM analysis detects but for which it cannot adjust resident measures to correct for these sources of construct-irrelevant variance, the Department of Surgery can use the output to identify the clinical services that exhibit aberrant rating behaviors (e.g.,

inconsistency and restriction of range). The results from a MFRM analysis help bring problematic rating behaviors to the Department's attention. With such information, the Department of Surgery can devise ways to work with problematic clinical services to try to change these behaviors through rater feedback, rater training, refining the rating instrument, as well as close monitoring of faculty rating behaviors in subsequent resident evaluations.

The major limitation of this study is the rating design employed. Faculty only provided ratings within their clinical service; they did not evaluate residents outside their specialty. This rating design makes it difficult to determine the precise nature of rating errors we discovered. Each clinical service has unique clinical tasks that can elicit different levels of performance from different residents. Because we could not assign faculty to evaluate residents outside their specialty, we could not cross faculty with clinical tasks in the services. Thus, we do not know whether the rating errors we discovered were due to the differences in the nature of the tasks between clinical services, or to differences in the level of severity that the faculty in each clinical service exercised.

## Conclusion

We demonstrated the utility of a MFRM approach for studying rater effects in clinical performance ratings of surgery residents. The MFRM analyses revealed various types of rating errors in ratings from many clinical services, including severity, inconsistency, and restriction of range. The MFRM model provided measures of residents' clinical performance that were adjusted for the differences in clinical services severity, yielding measures that each resident would receive if clinical services faculty exercising an average severity level rated him/her. However, many types of rater errors cannot be corrected mathematically, including rating inconsistency and restriction of range. These rater errors need to be addressed through other

means, such as providing feedback, rater training, or working to improve key aspects of the rating operation (e.g., refining the items and/or rating scales). While the results from MFRM analyses did not dictate a specific course of action, they suggested potential avenues to explore to reduce rater errors. Through continued monitoring of subsequent rating operations, one can then determine whether the initiated changes are having the desired effect.

# References

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of score and profiles*. New York: Wiley.

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education, 39*, 353-355.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Lawrence Erlbaum Associates.

Kwolek, C. J., Donnelly, M. B., Sloan, D. A., Birrell, S. N., Strodel, W. E., & Schwartz, R. W. (1997). Ward evaluations: Should they be abandoned? *Journal of Surgical Research, 69*(1), 1-6.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In G. Englehard, Jr. & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 85-98). Norwood, NJ: Ablex Publishing Corporation.

Linacre, J. M. (2002). What do infit and outfit mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2005). Facets (Version 3.57). Chicago, IL: Winsteps.

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.

McNamara, T. F. (1996). Raters and ratings: Introduction to multi-faceted measurement. *Measuring second language performance* (pp. 117-148). London: Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922-932.

Sloan, D. A., Donnelly, M. B., Drake, D., & Schwartz, R. W. (1995). Faculty sensitivity in detecting medical students' clinical competence. *Medical Teacher, 95*(3), 335-342.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Table 1

*Clinical Services Measurement Report (Ordered by Severity Measures)*

| Clinical Services | Severity measure | Standard error | Infit mean-square |
|---|---|---|---|
| Urology | 0.24 | 0.02 | 0.53 |
| Intensive care | 0.15 | 0.02 | 2.21 |
| Plastic surgery | 0.11 | 0.02 | 1.75 |
| Pediatric surgery | 0.07 | 0.02 | 0.70 |
| General B | 0.04 | 0.02 | 0.80 |
| Trauma | 0.03 | 0.02 | 0.42 |
| Anesthesiology | 0.01 | 0.02 | 1.37 |
| Head-Neck-Breast | 0.01 | 0.02 | 0.76 |
| General C | -0.01 | 0.02 | 0.81 |
| Minor operating room | -0.06 | 0.02 | 0.31 |
| Neurosurgery | -0.08 | 0.02 | 1.66 |
| Cardiothoracic surgery | -0.10 | 0.02 | 0.84 |
| Orthopedics | -0.17 | 0.02 | 1.03 |
| General A | -0.25 | 0.02 | 0.62 |

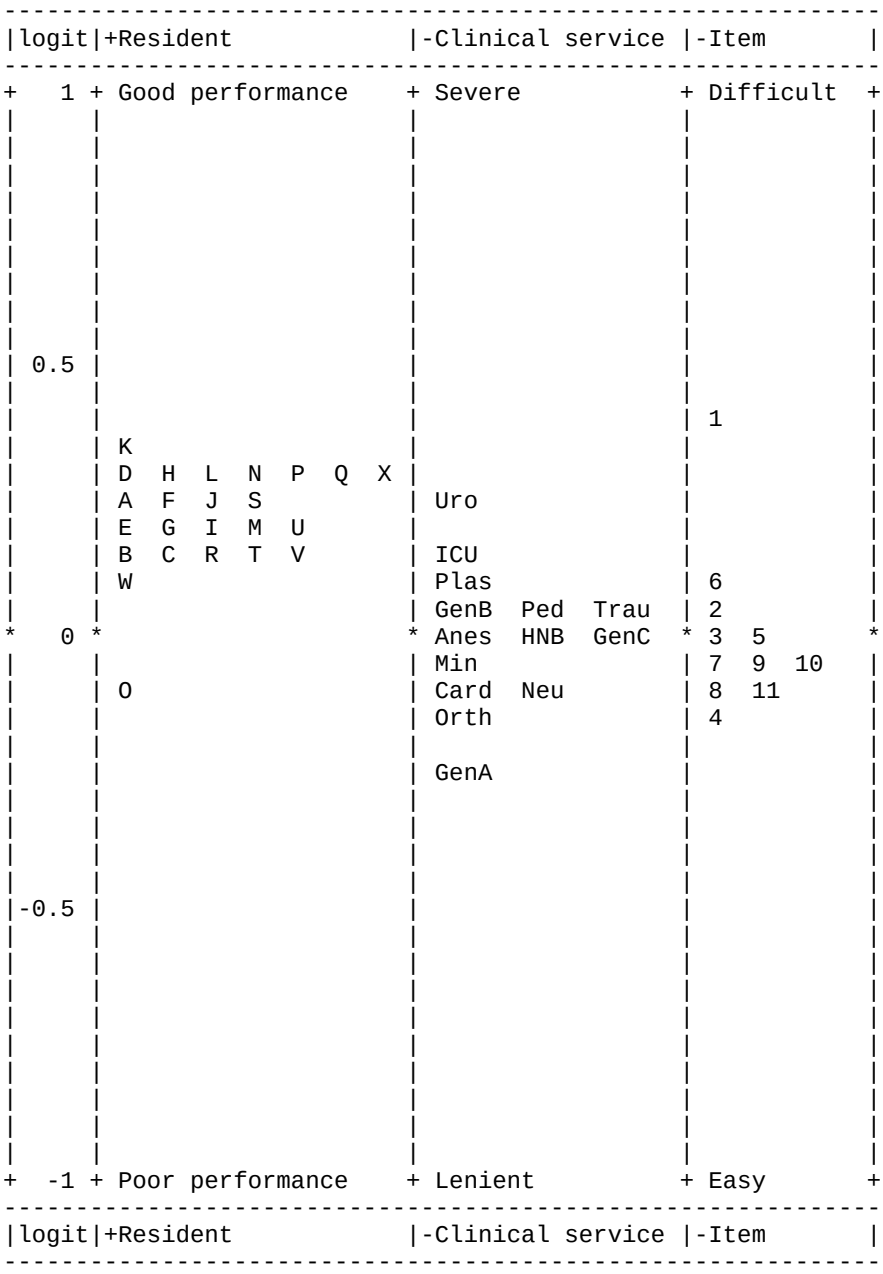
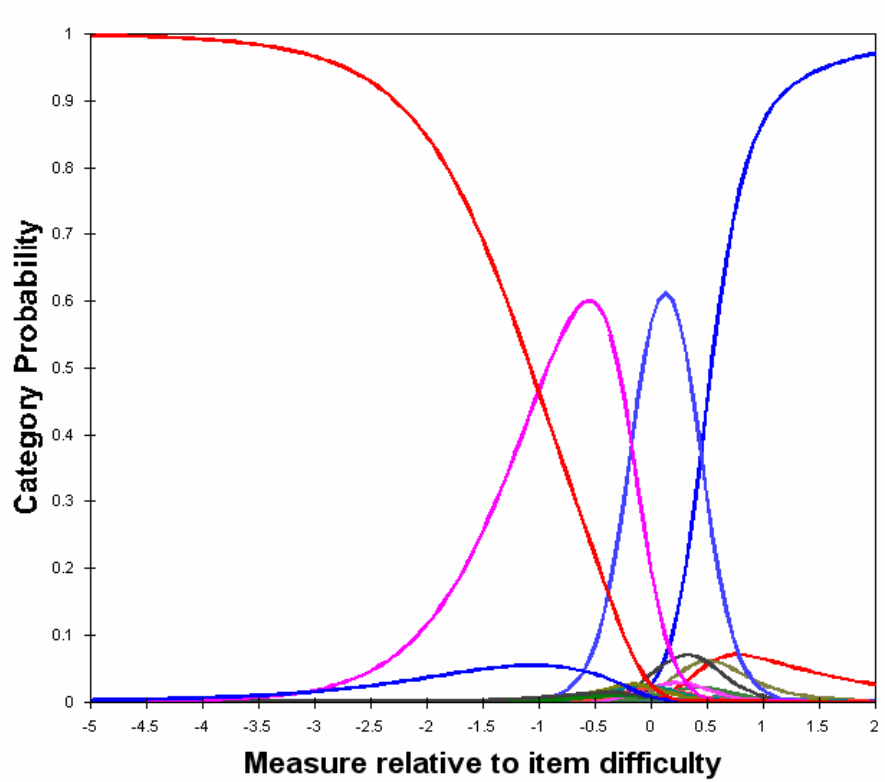
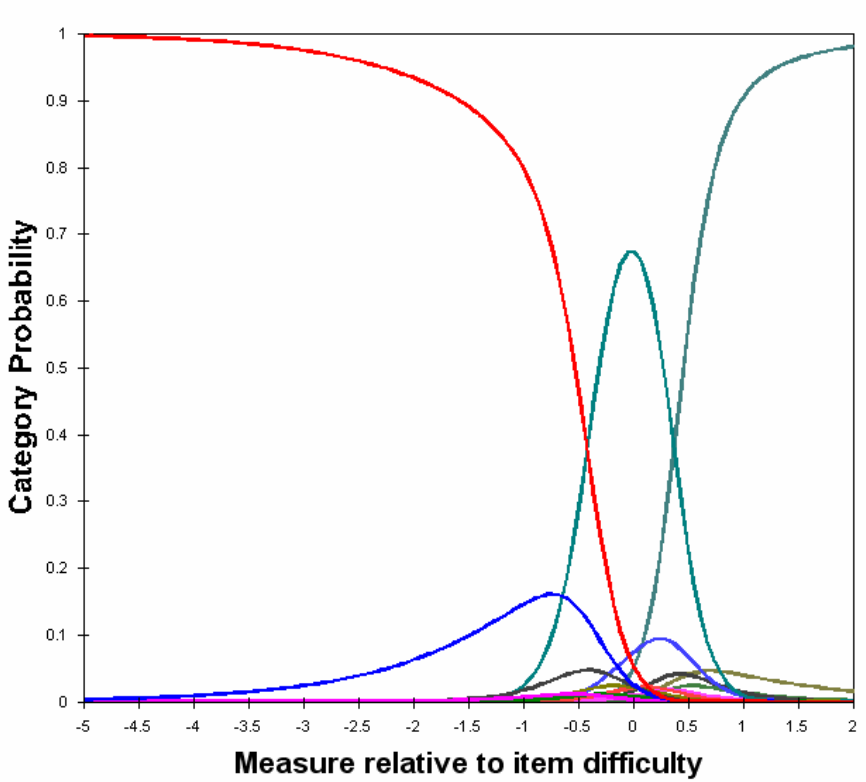



Figure 1. Variable map of clinical performance ratings of surgery residents.

Residents' clinical performance (second column), clinical service severity (third column), and item difficulty (fourth column) were measured on the same scale of logit measures (first column), allowing the comparison of the three facets. The higher the measure, the better the performance of the resident, the more severe the clinical service faculty, and the more difficult the item.




A. An item with four functioning categories

B. An item with three functioning categories (Item 1)

Figure 2 Category probability curves of rating items.

We converted an original continuous rating ranging from 1 to 5 to a categorical rating of 10 to 50, producing 41 possible categories. When Facets produced category probability curves, it produced a separate curve for each category. However, the only functioning categories are those that have specific range on clinical performance continuum in which they become the most probable ratings.