HOW GOOD ARE OUR RATERS?

RATER ERRORS IN CLINICAL SKILLS ASSESSMENT

Cherdsak Iramaneerat

Rachel Yudkowsky

University of Illinois at Chicago

Paper presented at the 2006 Graduate Educational Conference:

Education and the public good: Interdisciplinary trends in graduate scholarship

Chicago, Illinois USA

February 2006

Correspondence concerning this article should be addressed to Cherdsak Iramaneerat,

1407 S. Indiana Ave. Chicago, IL 60605 USA, Email: cirama1@uic.edu

HOW GOOD ARE OUR RATERS?

RATER ERRORS IN CLINICAL SKILLS ASSESSMENT

Abstract

A multi-faceted Rasch measurement (MFRM) model was used to analyze a clinical skills assessment of 173 fourth-year medical students in a Midwestern medical school to investigate four types of rater errors: leniency, inconsistency, halo, and restriction of range. Each student performed six clinical tasks with six standardized patients (SPs), who were selected from a pool of 17 SPs. SPs provided ratings of the performance of each student in six skills: history taking, physical examination, interpersonal skills, communication technique, counseling skills, and physical examination manner. Participating SPs showed statistically significant differences in their rating severity, indicating rater leniency error. Four SPs exhibited rating inconsistency, as evidenced from their high infit mean-square statistics. None of the SPs had low infit mean-square statistics, indicating that they did a good job in avoiding halo and restriction of range errors. These findings provided diagnostic information which was useful for a medical school to guide the ways to improve the quality of subsequent clinical skills assessments by eliminating construct-irrelevant variance caused by rater errors.

(Contains 24 references, 2 tables, and 1 figure)

HOW GOOD ARE OUR RATERS?

RATER ERRORS IN CLINICAL SKILLS ASSESSMENT

We assess clinical skills of medical students in four levels: know, know how, show how, and do. Performance in the levels of 'show how' and 'do' cannot be assessed with traditional paper-based tests. These higher levels of performance are usually assessed with standardized patients (SPs), who are lay persons trained to portray a scripted patient presentation in a standardized and consistent fashion (Van der Vleuten & Swanson, 1990; Yudkowsky, Alseidi, & Cintron, 2004). Often, SPs are also trained to provide ratings to students' performance.

The use of ratings provided by SPs rests on the assumption that SPs (raters) can objectively provide scores with some degree of precision (Myford & Wolfe, 2003). Nevertheless, when raters participate in the act of rating, they are engaging in a complex and error-prone cognitive process (Cronbach, 1990). Raters can introduce various sources of variance into performance ratings that are associated with their own rating behavior and not with the actual performance of the ratee. These sources of variance are collectively called *rater effects* (Scullen, Mount, & Goff, 2000) or *rater errors* (Myford & Wolfe, 2003). These errors are important sources of construct-irrelevant variance in the assessment results. Construct-irrelevant variance in performance ratings is error introduced into performance ratings by variables extraneous to the performance of ratees, causing a threat to validity of inferences made from the assessment results (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Downing & Haladyna, 2004). Thus, to ensure the objectivity in clinical skills assessment using SPs, we should monitor and control these rater errors.

There are many causes of rater errors. Different causes result in different patterns of ratings and require different remedies. It is important that researchers understand the different nature of these rater errors to develop effective strategies to control them. Researchers have classified rater errors into many types, according to their causes and rating patterns. The four most studied rater errors are: (1) leniency, (2) inconsistency, (3) halo, and (4) restriction of range. There are other types of rater effects as well, such as contrast error, logical error, proximity error, recency error, order effects, and rater bias. However, these errors are not quite commonly studied, perhaps because they are more difficult to detect (Myford & Wolfe, 2003). In addition, these uncommonly studied rater effects can also be considered as special cases of rater inconsistency. In this study, we focused on the four common rater errors.

1. Leniency

Leniency is a constant tendency of a rater to give out ratings to ratees higher than they should receive (Guilford, 1954). This causes a problem in interpreting assessment results. When a student receive high ratings, we cannot know that those are high ratings because of a student's good performance or because of the rater's tendency to give out high ratings. On the other end of the extreme is a rater who constantly gives out low ratings. We call such rater a severe rater or a hard rater. Severe raters can cause problems in making valid inferences from assessment results as much as lenient raters can. Because both lenient and severe raters introduce a similar kind of problems to the rating operation, researchers generally use the term 'leniency error' to apply to a general, constant tendency of a rater to rate too high or too low for whatever reasons (Myford & Wolfe, 2003). In a rating operation where raters are different in their levels of severity, some are lenient while others are severe, and ratees are evaluated by different sets of raters, this leniency error can produce unfair assessment results. In other words, the fact that one student received

higher scores than another can either be due to his/her better performance or due to his/her better luck in getting more lenient raters.

2. Inconsistency

Rater inconsistency is a rater's tendency to apply one or more rating scales in a manner inconsistent with the way in which the other raters apply the same scales. In other words, a rater exhibits more random variability than expected in his or her ratings. Because rating inconsistency shows up as increasing randomness in ratings, researchers also call it a *randomness effect* (Myford & Wolfe, 2004). The presence of rater inconsistency indicates the rater's lack of understanding of rating criteria, making the proposed interpretation of ratings difficult and less meaningful (Downing & Haladyna, 2004). A rater who rates inconsistently increases the randomness in ratees' scores by assign high ratings to those who deserve low ratings and assign low ratings to those who deserve high ratings, reducing the ability of the scores to reliably differentiate between competent and incompetent ratees.

3. Halo effect

Halo effect is by far the most studied type of rater effect. It was first observed by Wells in 1907 when he noted a tendency of a rater to allow an individual's performance in general or performance in one trait to influence the performance in other traits. Thorndike later labeled this rating phenomenon a *halo error*, and noted its characteristic as having intercategory correlations higher than reality. Newcomb later defined this halo effect in 1931 as a rater behavior in rating similarly logically related behaviors as those classified under a single trait. The close relation between ratings of two behaviors, therefore, springs from logical presumptions in the minds of the raters, rather than from actual behaviors (Cooper, 1981).

Since the early conception of the halo effect, many researchers attempted to explain this abnormally high intercorrelation between rating dimensions, resulting in multiple versions of conceptual definitions of halo effect. Fisicaro and Lance (1990) reviewed these conceptual definitions and summarized them into three classes of causal models:

1. General impression model: Halo effect occurs because a rater's general impression of a ratee has a common causal effect on many dimensional evaluations. The conceptual definition of halo effect according to this model is provided by King, Hunter, and Schmidt (1980) as "the tendency of a rater to allow overall impressions of an individual to influence the judgment of that person's performance along several quasi-independent dimensions of job performance" (p. 507), by Nisbett and Wilson (1977) as "the influence of a global evaluation on evaluations of individual attributes of a person" (p. 250), and by Linn and Gronlund (2000) as "an error that occurs when a rater's general impression of a person influences the rating of individual characteristics" (p. 279).

2. Salient dimension model: Halo effect occurs because a rater's evaluation on one salient dimension directly influences the rater's evaluation on a second, less salient dimension. The conceptual definition of halo effect according to this model is provided by Anastasi (1988) as the influence of a rater's evaluation of ratee behavior on one or more salient dimensions on evaluations of behavior on other dimensions, and also by Robbins (1989) as "the tendency for an evaluator to let the assessment of an individual on one trait influence his or her evaluation of that person on other traits" (p. 444).

3. Inadequate discrimination model: Halo effect occurs because a rater's misinterpretation of which ratee behaviors belong to which dimensions. The conceptual definition of halo effect according to this model is provided by Saal, Downey, and Lahey (1980)

as “a rater’s failure to discriminate among conceptually distinct and potentially independent aspects of a ratee’s behavior” (p. 415).

Halo effect is a threat to the validity of inferences made from performance ratings because it produces repetitious ratings, inflating the reliability estimates (Downing & Haladyna, 2004).

4. Restriction of range

Restriction of range is the clustering of ratings around a particular portion of the rating scale. Raters who have a restriction-of-range effect arbitrarily restrict their ratings to one particular polarity of the scale (either lenient or severe ratings) or the midpoint of the scale (central tendency), failing to employ ratings in other portions of the scale (Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980). The most studied pattern of restriction of range is when raters avoid both extremes of the scale and tend to rate everyone as average. This special case of restriction-of-range effect is called *central tendency error* (Linn & Gronlund, 2000). Central tendency error suggests raters’ hesitation in giving out extreme judgments, resulting in the displacement of individuals in the direction of the mean of the total group (Guilford, 1954). Restriction of range can cause two undesirable results. First, it puts in doubt a single rating of an individual. A high or low rating might reflect the personal outlook of the rater rather than the actual performance or personal characteristics of ratees. Second, it limits the variability of any individual’s ratings, thus failing to provide reliable discrimination between ratees (Anastasi, 1988; Linn & Gronlund, 2000).

This paper demonstrated an approach in detecting these rater errors using the multi-faceted Rasch measurement (MFRM) model (Linacre, 1989). The MFRM model is an extension of the basic Rasch one-parameter item response theory (IRT) model. IRT is a class of

psychometric models used to estimate examinees' ability and the difficulty of test items on the same scale (Downing, 2003; Hambleton, Swaminathan, & Rogers, 1991). The basic Rasch model uses only one parameter, item difficulty, to estimate examinees' ability. MFRM extends the basic Rasch model by adding parameters describing facets of measurement interest other than item difficulty (such as rater severity or task difficulty) to the model (Linacre & Wright, 2004). MFRM concerns itself with obtaining from each examinee's raw ratings a linear measure corrected for the particular raters or tasks that the examinee encountered. MFRM attempts to free each examinee's measure from the effects of differences in rater severity or task difficulty (Linacre, 1989; Linacre & Wright, 2004).

The purpose of this study was to assess the degree to which rater errors occurred in clinical skills assessment of medical students. This served as one source of validity evidence of this assessment. The study also indicated the sources of rater errors, guiding us to the ways to reduce and eliminate these errors.

Method

Participants

We examined a clinical skills assessment of one class of fourth-year medical students in a Midwestern medical school in the United States of America. This class had 183 students in total. Their gender distribution is 56% male and 44% female. Their ethnicity includes White (38%), Asian (38%), Hispanic (10%), Black (9%), and others (5%). Their age distribution was shown in Table 1.

[INSERT TABLE 1 ABOUT HERE]

From this group of 183 students, only 173 students (95%) participated in this clinical skills OSCE. Each student was asked to performed six clinical tasks. Each clinical task involved interacting with one SP under a specified clinical setting. The six SPs that each student encountered were selected from a pool of 17 SPs. This pool had 65% male and 35% female. Their ethnicity includes White (88%), and Black (12%).

Clinical tasks

This assessment measures clinical skills competence using six clinical tasks. Each task was a simulated clinical scenario asking a student to interact with a SP in a certain clinical situation that requires some types of clinical skills. These tasks generally required students to take medical history, perform certain types of physical examination, and discuss investigations and treatment options with a SP or his/her family member. For the purpose of test security, detailed content of these tasks could not be provided.

Rating procedure

The SPs rated the performance of each student at the end of each simulated clinical encounter using a rating form specifically designed for the clinical scenario. Different clinical scenarios employed different sets of items in the rating form. Nevertheless, all the items used in these rating forms were crafted to address six clinical skills: (1) history taking, (2) physical examination, (3) interpersonal skills, (4) communication technique, (5) counseling skills, and (6) physical examination manners.

Items that targeted history taking skills assessed students' ability in asking appropriate questions to collect relevant medical information from a patient and/or his or her family member that should lead to a correct medical diagnosis and a proper treatment. These items were presented in the form of a checklist where SPs marked whether a specific question was asked or not by a student during the encounter.

Items that targeted physical examination skills assessed students' ability to perform a specific type of physical examination on a patient to check for physical signs that are important for making clinical diagnosis. These items were presented in a checklist format asking SPs to check whether a specific type of physical examination was performed correctly by a student.

Items that targeted interpersonal skills assessed students' competency in communicating with a patient and/or his/her family member verbally (such as greeting, asking questions, and the use of medical jargons) and non-verbally (such as showing interest, allowing a patient to tell a story, and being a good listener). These items were generally presented on a five-point rating scale asking the level of agreement with the statement of good communication practice.

Items that targeted communication techniques probed the techniques that students used in communicating with a patient and/or his/her family member. Examples of these items are maintaining eye contact, logical sequence of an interview, clarification of key clinical points. These items were generally presented on a four-point rating scale asking the level of agreement with the statement of good communication techniques.

Items that targeted counseling skills focused on how students provided medical information and medical advice to a patient and/or his/her family members. These items checked for various aspects of counseling skills such as clarity of explanation, providing of alternatives, openness to questions, and confirmation of patient's understanding. The items were presented on

a five-point rating scale asking the level of agreement with the statement of good counseling techniques.

Finally, items that targeted physical examination manners checked students' approach during the physical examination to ensure that a patient were comfortable and did not feel embarrassed while allowing a student to examine various parts of his/her body. These items were presented in a five-point rating scale asking the level of agreement with the statement of good physical examination manners.

Because different clinical skills employed different scoring format and used different number of items in each clinical task, raw ratings from each clinical skills were converted into a skill score on a five-point rating scale ranging from 1 for unacceptable performance to 5 for excellent performance to ensure that all skills were weighted equally in the final scoring.

Analyses

We conducted a MFRM analysis utilizing the Facets computer program (Linacre, 2005), using a four-faceted model, in which the probability of receiving any given rating was determined by 1) students' clinical skills competence, 2) case difficulty, 3) SP severity, and 4) the difficulty of the clinical skill. This four-faceted model can be expressed mathematically as:

$$P_{mnijk}(X \mid \theta) = \frac{e^{\Sigma(B_n - C_j - D_i - F_m - E_{ik})}}{\Sigma e^{\Sigma(B_n - C_j - D_i - F_m - E_{ik})}} \quad , \qquad (1)$$

where

P_{mnijk} is the probability of student n being rated by SP j on skill i of case m with rating category k

B_n is the clinical skills competence of student n

C_j is the severity level of SP j

D_i is the difficulty level of skill i

F_m is the difficulty level of case m, and

E_{ik} is the difficulty of rating k relative to (k-1) for skill i

The MFRM analyses provided SPs' measures of severity in logit units. Mean SP severity was fixed at 0 logit. SPs who were more lenient than average had negative logit measures, while those who were more severe than average had positive logit measures. We determined whether these SPs were significantly different in their levels of severity with a chi-square statistic that tests for a null hypothesis which states that there were no significant differences in SP severity measures. A significant chi-square value ($p < 0.05$) indicates that at least two SPs were significantly different in their severity levels (Myford & Wolfe, 2004).

We examined infit mean-square values of SP severity measures to check for rater inconsistency, halo, and restriction of range. Infit mean-square values are weighted mean-square residual statistics that measure how different the observed ratings are from the ratings that the measurement model expects. The weighting reduces the influence of less informative, low variance, off-target responses (Wright & Masters, 1982). The more observed ratings deviate from the expected ratings, the higher their mean-square values. The expected infit mean-square value equals 1.0 when the model fits the data. An infit mean-square value less than 0.4 indicates that the SP gave ratings with too little variation, suggesting the presence of halo and restriction of range. On the other hand, an infit mean-square value greater than 1.2 indicates the SP's lack of consistency in applying the rating criteria (Wright & Linacre, 1994).

Results

The calibration of students' clinical skills competence, case difficulty, SP severity, and the difficulty of the clinical skills using the MFRM model can be summarized with a variable map (Figure 1). The MFRM model provided measures of all four facets on to a common logit scale. A logit scale is the logarithmic scale of odds of receiving a high rating over a low rating. Higher logits indicated students who were more competent, cases that were more difficult, SPs who were more severe, and clinical skills that were more difficult.

[INSERT FIGURE 1 ABOUT HERE]

Students' clinical skills competence measures were normally distributed, ranging from 0.45 to 2.47 logits, with a mean of 1.41 logits and a standard deviation of 0.41 logits. Cases varied in their difficulty measures from -1.47 logits for the easiest case (Case 6) to 0.79 logits for the most difficult case (Case 5), with a mean of 0 logit and a standard deviation of 0.74 logits. Clinical skills also varied in their difficulty measures from -0.81 logits for the easiest skill (interpersonal skills) to 0.79 logits for the most difficult skill (communication technique), with a mean of 0 logit and a standard deviation of 0.55 logits.

SP severity measures ranged from -0.76 logits (SP 17) to 0.55 logits (SP 1) with a mean of 0 logit and a standard deviation of 0.37 logits (Table 2). A chi-square test indicated that these SPs were significantly different in their severity measures (χ^2 (16) = 203.4, $p < 0.05$). This indicated that these SPs exhibited rater leniency errors.

We checked infit mean-square statistics of SP measures to assess rater inconsistency, halo, and restriction of range in their ratings. Infit mean-square values of SP measures ranged

from 0.56 (SP 6) to 1.35 (SP 4), with a mean of 1.02 and a standard deviation of 0.21. Four SPs (24%) had their infit mean-square values greater than 1.2, indicating their rating inconsistency. These four problematic SPs are SPs 3, 4, 7, and 8. None of the 17 SPs had their infit mean-square values lower than 0.4, suggesting that these SPs were free from halo and restriction of range (Table 2).

[INSERT TABLE 2 ABOUT HERE]

Discussion

This study demonstrated an approach to study various types of rater errors in an assessment of clinical skills of medical students using the MFRM model. We focused on four types of rater errors: leniency, inconsistency, halo, and restriction of range.

The statistically significant chi-square test of SP severity measures suggested that some SPs rated more severely than others. The differences in severity levels between SPs could introduce construct-irrelevant variance into the ratings of medical students' clinical skills competence, especially in this study, where different students encountered different sets of SPs. Students who were rated by lenient raters would have an unfair advantage over those who were rated by severe raters, if we relied on unadjusted raw ratings of these SPs. To obtain more objective measures of students' clinical skills competence that are fairer to every student, we should take into consideration the differences in severity levels between SPs. Because the MFRM model measures students' clinical skills competence on the same scale with SP severity measures, the model can make appropriate adjustment to students' clinical skills competence measures for the differences in SP severity measures (i.e., providing the measure that each

student should receive if he/she was rated by a SP with average severity measure), producing fair measures of clinical skills competence.

There were four SPs who rated inconsistently. These four SPs 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. We should work with these four SPs to help them become aware of their errors. They might benefit from feedback on their errors and how to avoid these errors. They might need some clarification on how to interpret some items in the rating form. Engaging them in discussions about the criteria used to justify ratings in each category would allow them to reflect on their misconceptions and lead to a common understanding of each rating category of each item. In some instances where rater errors occur so widespread among many raters, we can consider providing appropriate rater training to these raters.

Our study demonstrated that none of these 17 SPs exhibited halo and restriction of range errors. Although some of them provided more variance in their ratings than others, all of them contributed adequate variance in their ratings, producing infit mean-square statistics that were higher than the lower control criterion. Thus, there seemed to be no need to provide additional rater training on halo and restriction of range to these SPs. Nevertheless, we should still keep monitoring their rating patterns to see if these types of rater errors occur in other assessments.

Conclusion

We demonstrated the utility of a MFRM model in the study of rater errors in the context of clinical skills assessment of medical students. Like other types of ratings, clinical skills assessment using SPs is also subjected to rater errors. Leniency error and rating inconsistency showed up as a significant problem in this assessment. On the other hand, these SPs did a good

job in avoiding halo and restriction of range rating errors. We can use the diagnostic information obtained from the analysis to improve the quality of subsequent clinical skills assessments by eliminating construct-irrelevant variance from rater errors.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218-244.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper and Row.

Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. *Med Educ, 37*(8), 739-745.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(4), 419-429.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw Hill.

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

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65*(5), 507-516.

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

Linacre, J. M. (2005). Facets [computer program]. version 3.57 (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.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

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.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250-256.

Robbins, S. P. (1989). *Organizational behavior* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

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

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.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch measurement Transactions, 8*(3), 370. Available from : URL: http://www.rasch.org/rmt/rmt383b.htm.

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

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.

Table 1.

The Distribution of Medical Students by Age and Gender

Age group	Male	Female	Total
20 - 25 years	8 (8%)	6 (7%)	14 (8%)
26 - 30 years	76 (75%)	61 (75%)	137 (75%)
31 - 35 years	15 (15%)	11 (14%)	26 (14%)
more than 35 years	3 (3%)	3 (4%)	6 (3%)
Total	102 (100%)	81 (100%)	183 (100%)

Table 2.

SP Measurement Report (Sorted by measures from the highest to the lowest)

SP	Measures (logits)	Standard Error	Infit Mean-square
1	0.55	0.09	1.12
6	0.47	0.14	0.56
18	0.43	0.11	0.98
8	0.34	0.08	1.21
19	0.33	0.13	1.07
15	0.20	0.06	1.12
3	0.16	0.14	1.25
11	0.07	0.10	0.94
13	0.01	0.07	1.03
5	-0.07	0.06	0.80
12	-0.08	0.08	1.06
9	-0.09	0.07	1.02
14	-0.20	0.07	1.01
7	-0.25	0.09	1.25
4	-0.41	0.07	1.35
2	-0.71	0.14	0.61
17	-0.76	0.17	1.00

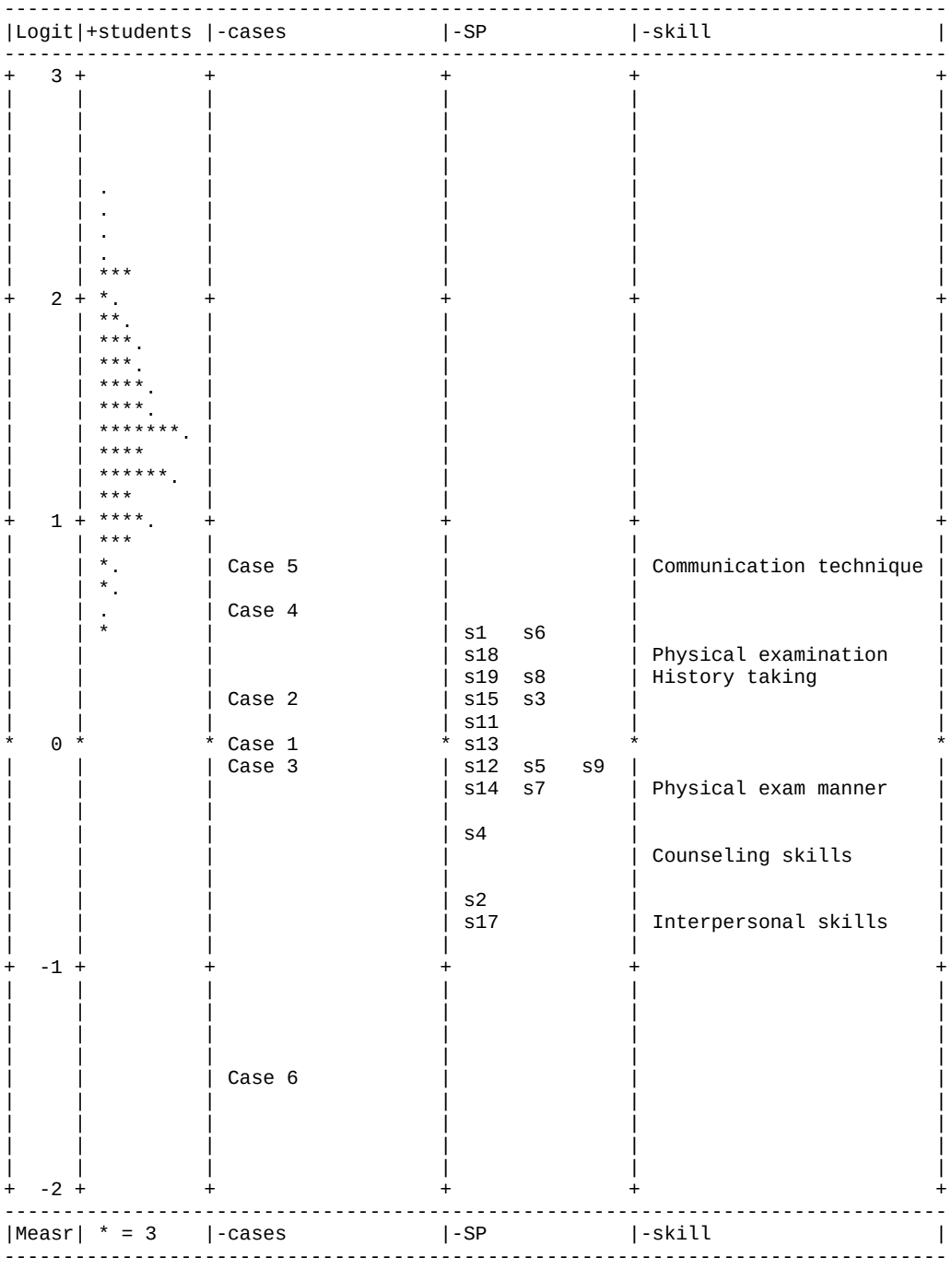

Figure 1. A variable map of a clinical skills assessment.

Student clinical skills competency (second column), case difficulty (third column), SP severity (fourth column), and clinical skills difficulty (fifth column) were calibrated on to the same scale of logit measure (first column), allowing the comparison of all the facets of measurement.