

Determining Component Weights in a Communications Assessment Using Judgmental Policy Capturing

Leo M. Harvill, PhD, F. Forrest Lang, MD, Ronald S. McCord, MD (deceased)

Departments of Medical Education and Family Medicine
James H. Quillen College of Medicine
East Tennessee State University

Abstract Objectives: Tools are needed for determining appropriate weights for complex performance assessment components in medical education. The feasibility of using judgmental policy capturing (JPC), a procedure to statistically describe the information processing strategies of experts, for this purpose was investigated.

Methods: Iterative JPC was used to determine appropriate weighting for the six core communication skill scores from a communications objective structured clinical examination (OSCE) for medical students using a panel of four communication skill experts.

Results: The mean regression weights from the panel indicated they placed less importance on information management (8.5%), moderate and nearly equal importance on rapport building (15.8%), agenda setting (15.4%), and addressing feelings (14.1%), and greater importance on active listening (20.1%) and reaching common ground with the patient (25.5%).

Discussion: JPC is an effective procedure for determining appropriate weights for complex clinical assessment components. The derived weights may be very different for those assessment components.

Keywords: judgmental policy capturing, component weighting, physician-patient communication, objective structured clinical examination

Many skills measured in medical education are multi-faceted. While the various facets of a particular skill can often be assessed effectively, they are not always of equal importance in determining overall competence for that skill. Thus, a scoring formula that gives equivalent weights to the various facets will systematically generate a final assessment score that is different from the one that represents the best judgment of overall examinee performance by experts in the field.

An available method for determining appropriate category weights for multidimensional skill assessment is iterative judgmental policy capturing (JPC); it is appropriate for use with any assessment that includes complex performances. Hobson and Gibson¹ described policy capturing as “a general procedure designed to describe statistically the unique information processing strategies of individual raters.” The process produces a regression equation that defines the captured policy for each individual rater. The captured policy represents an explicit statement about the manner in which the rater combines and weights the various performance dimensions in arriving at final ratings. More recently, Roeber² indicated that JPC involves an appropriate group (an expert panel, a representative user group or a policymaker group)

reviewing the various components of an overall assessment and determining which of the components are more important than others. The process is an iterative one when the raters or judges are asked to repeat the process one or more times after some type of feedback and/or discussion concerning the outcomes from the previous trial.

JPC is used to capture policy from individual judges relative to various profiles of assessment outcomes and, when averaged, can provide weights for the components of the assessment. It is not used to look at individual OSCE cases or “test items” and is not used to determine a minimum passing score or set a standard for a particular assessment instrument as is the case with standard setting methods.

The specific steps in the procedure for a single iteration of JPC are:

1. Judges are asked to independently provide overall ratings of the performance of examinees on a complex skill assessment often using graphic representations or profiles of the scores from those assessments for a large number of examinees. The judges' ratings are often couched in terms of the compe-

tence of the examinees. Repeated ratings of some of the same profiles provide a measure of stability of the judgments of each individual judge although this is not a mandatory part of the process. Overall ratings can also be compared among the various judges to determine the degree of agreement among the judges but consensus among the judges is not a goal of this process.

2. Multiple regression analysis is used to determine appropriate regression or beta weights for the assessment components using the various assessment component scores as the predictor or independent variables and the judge's overall ratings for the assessment profiles as the dependent variable or variable to be predicted for each judge. This is a widely used statistical procedure that provides a means for capturing how each expert valued each skill component in arriving at his/her global ratings of performance.
3. The regression weights from each of the judges can be expressed as percentages and, thus, provide a straightforward statement about the relative importance of each assessment component in determining the overall ratings made by that judge. If the assessment has three facets, one judge's percentage weights for those components might be 30 percent, 35 percent and 35 percent, respectively, while a second judge's weights or values for the importance of the three facets in terms of overall performance might be 40 percent, 40 percent and 20 percent, respectively. A third expert's percentage weights might be 35 percent, 40 percent and 25 percent, respectively.
4. The sets of weights provided by the judges are then typically averaged to arrive at a composite set of weights for that particular skill assessment. For the example above, the average weight for the facets would be 35 percent, 38.33 percent and 26.67 percent, respectively. This set of weights could then be applied to new assessments of that skill to provide an overall rating or score for examinee performance. For example, if an examinee received percentage scores of 65 percent, 85 percent and 80 percent on the three assessment components, his/her overall percentage score would be:

$$(.35 \times 65) + (.3833 \times 85) + (.2667 \times 80) = 76.67 \text{ percent.}$$

5. The process can be continued with an additional iteration of the above steps after the panel of judges have had some opportunity to see their own set of regression weights and those of the other panel members; some discussion among the panel members might or might not be included.

JPC has been used to describe raters' relative weightings of performances in such areas as clinical judgment in psychology,³ personnel selection and promotion,⁴ and teacher competence and certification.⁵ Clauser, Subhiyah, Nungester, Ripkey, Clyman and McKinley⁶ used JPC to model clinician raters' decisions in scoring an assessment that used a computer-based simulation of the patient care environment in a medical education setting.

This article provides a description of an undergraduate medical education setting, the procedures used to obtain JPC judgments from a panel of four experts and a discussion of the results of applying JPC in a situation involving this panel and six physician-patient communication skills assessed in an objective structured clinical examination (OSCE).

Methods

Iterative JPC was used to derive appropriate weights for the assessment of six key physician-patient communication skills referenced in the Toronto⁷ and Kalamazoo⁸ consensus statements: 1) rapport building, 2) information management, 3) eliciting or setting the agenda, 4) active listening, 5) addressing feelings, 6) negotiating or reaching common ground with the patient. These consensus statements identified communication skills that, if implemented, would result in improved doctor-patient relationships and better health outcomes. The authors have developed standardized patient cases and related assessment instruments to reliably and validly assess these communication skills.⁹

Four medical school faculty members from across the country who are recognized experts in the area of patient-centered communication took part in this exercise.

The experts first viewed videotaped standardized patient interviews of real third year medical students as a means of grounding them in how to interpret the percentage scores and the score profiles for the six communication skills. The videotapes were recorded

during a communication skills OSCE to measure the six communication skills mentioned earlier. Two students were observed doing two interviews each and six students were observed during only one interview. The experts were then provided graphic profiles for the six core skills for each student (the graphic profiles, similar to Figure 1, showed the mean percentage scores based on four OSCE cases for each of the core skills as previously rated by trained raters). The experts provided their own global assessment of each of the six skills and the student overall performance using a five-point rating scale with five as the most skilled performance. In addition, they provided judgments about the competence of each of the eight students using the categories of notable, competent, and incompetent. For example, if the hypothetical student profile presented in Figure 1 represented one of these eight students, one of the experts might have given him/her ratings of five, four, two, one, four and two, respectively, for the six individual communication skills, further provided an overall rating of three on the five-point scale, and lastly indicated that the student's overall performance was competent.

Following this training concerning the communication skills and the scoring of those skills, the experts were introduced to the process of JPC and the way they were to use this method to recommend appropriate weights for the communication skills.

One hundred hypothetical graphical profiles of percentage scores on each of the six communication skills were created. The profiles were constructed to represent a wide variety of student performance on the six communication skills; there were profiles with low scores, moderate scores, high scores, and with low, moderate, and high scores mixed together. An example of a profile is shown in Figure 1.

Each of the profiles represented a single student's performance on a total of four standardized patient cases. For each of the six skills, the hypothetical student was scored on the percentage of the total possible points they obtained with possible scores ranging from zero to 100. The average percentage scores for the six skills for the four cases in the figure are: rapport building, 90 percent; information management, 75 percent; eliciting the agenda, 18 percent; active listening, 10 percent; addresses feelings, 70 percent; reaching common ground, 30 percent. The values presented within each graph were the mean percentage values for the four cases along with plus and minus one standard deviation for those four values. The vertical axis actually ranged from

-20 to +120 percent so that standard deviations could generally stay within the body of the graph.

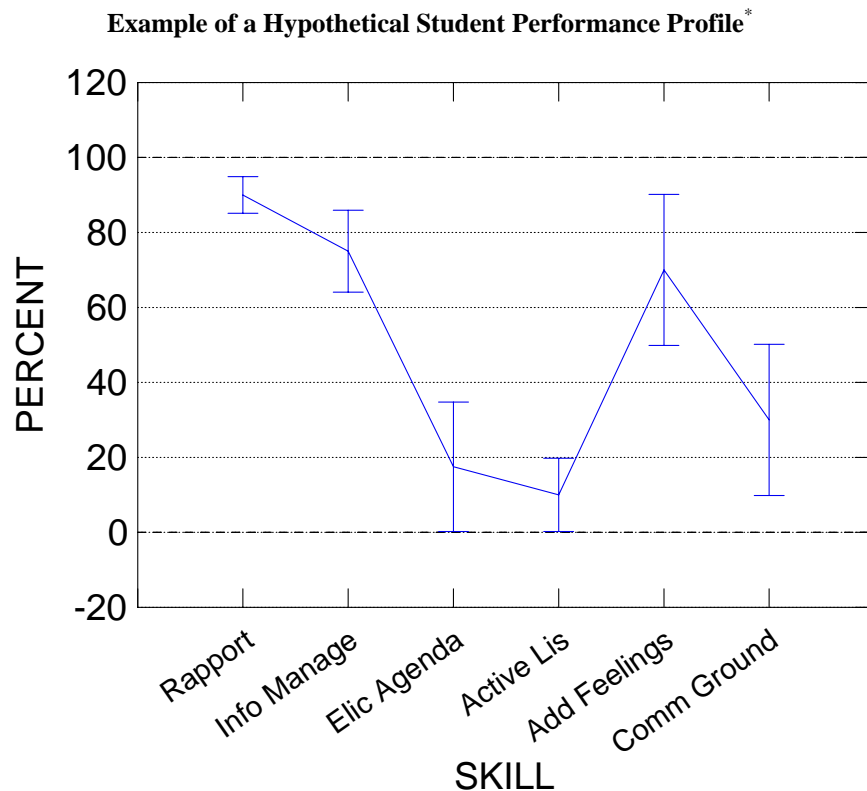
The judges applied JPC by independently rating 100 randomly arranged profiles and immediately re-rating 40 of those profiles selected randomly from the group of 100 profiles. They were not told that they were repeating any of the profile ratings. They were asked to make judgments concerning the overall competency of the student represented in each profile using the ratings of notable, competent, and incompetent. They were also asked to make comments concerning their thoughts about their judgments as they progressed through the process. After they finished, they were asked to record additional notes about any criteria or guiding principles they used or developed as they completed the task. They were asked to refer to these notes the following day during the group discussion.

The following day, after viewing the regression weights (expressed as percentages) each of the experts had assigned to the six communication skills and having a structured discussion about their reasoning for valuing some skills more highly than others, they performed the second iteration of rating the same 140 hypothetical student profiles (see above) in the same order as on the first day.

The relative weights each expert applied to the six communication skills in specifying the overall competence of hypothetical medical students were estimated using ordinary least squares regression analysis for each of the two iterations of the 100 initial profiles. The percentage scores for the six skills were the independent variables while the expert rating of competence, converted to numerical values of three for notable, two for competent, and one for incompetent, served as the dependent variable for the regression analysis for the data set from each expert.

Measures of consistency were calculated for each expert between their two ratings of the 100 profiles on each of the two days and between their two ratings of the 40 common profiles for each day by calculating kappa coefficients and determining percentages of exact agreement in their ratings. Interrater consistency was determined among pairs of experts by determining percentages of exact agreement in their ratings of notable, competent and incompetent and by calculating kappa coefficients for all four experts combined across 100 profiles.¹⁰ Kappa is a measure of agreement as is an intraclass correlation coefficient. If there is complete agreement between raters, kappa is equal to +1. If the observed agreement is greater than chance agreement, kappa will be

Figure 1



Rapport – Rapport Building
 Elic Agenda – Eliciting or Setting the Agenda
 Add Feelings – Addresses Feelings

Info Manage – Information Management
 Active Lis – Active Listening
 Comm Ground – Negotiating or Reaching Common Ground

*Graphic values are mean values plus and minus one standard deviation based on four standardized patient cases for the six core communication skills

a positive value and if the observed agreement is less than chance agreement, kappa will be negative. The lowest possible value for a kappa coefficient is -1, indicating a complete absence of agreement between raters or ratings.

Results

The kappa coefficient and percentage of agreement between the ratings of the 100 profiles each of the two times for each of the four experts is presented in Table 1. These same values are given for the re-

**Table 1
 Intra-Rater Consistency**

	Percentage of Rating Agreement*	Percentage of Rating Agreement [†]	Percentage of Rating Agreement [‡]
Rater #1	79.0% (0.64)	90.0% (0.83)	95.0% (0.91)
Rater #2	79.4% (0.63)	100.0% (1.00)	97.4% (0.95)
Rater #3	82.8% (0.66)	90.0% (0.77)	100.0% (1.00)
Rater #4	81.0% (0.64)	97.5% (0.95)	90.0% (0.78)

*Rating of 100 profiles, iterations 1 and 2

[†]Rating of 40 profiles twice, iteration 1

[‡]Rating of 40 profiles twice, iteration 2

Percentage of Exact Rating Agreement with Kappa Coefficients in parentheses

peated ratings of the 40 profiles on each of the two iterations. Some disagreement between each rater's competency judgments was expected for the two separate iterations since the structured discussion may have influenced some changes. However, all of the raters had approximately 80 percent of the ratings exactly the same between the two iterations; kappa coefficients ranged from 0.63 to 0.66. There was excellent intra-rater consistency when they re-rated the same 40 profiles each of the two days; all eight of those percentages of agreement were 90 percent or greater and the kappa coefficients ranged from 0.77 to 1.00.

The degree of agreement between pairs of experts for each of the two iterations (Table 2) was only moderate ranging from approximately 50 percent to

in Table 3. Although some of the component weights for three of four individual experts changed considerably between the two iterations (e.g., from .159 to .064, from .061 to .148, from .090 to .248), the mean changes from the first to the second iteration were less dramatic. The largest negative mean change between the two days was for information management (from .117 to .089) while the largest positive mean change was for reaching common ground (from .209 to .257). Based on these average values, it is clear they placed the greatest weight on the skill of reaching common ground and the least weight on the information management skill.

Using the means as regression coefficients, the following equation could be used to determine a student's percentage score on a communication skill

Table 2
Inter-Rater Consistency

Percentage of Exact Rating Agreement: First Iteration (N = 140 profiles)

	Rater #1	Rater #2	Rater #3
Rater #2	54.1%		
Rater #3	71.7%	50.5%	
Rater #4	65.0%	77.6%	53.5%

Percentage of Exact Rating Agreement: Second Iteration (N = 140 profiles)

	Rater #1	Rater #2	Rater #3
Rater #2	73.7%		
Rater #3	70.0%	61.6%	
Rater #4	65.0%	77.8%	49.0%

80 percent; this was expected since they were expressing their own expert competency judgments and not attempting to "match" their colleagues. It is interesting to note that some percentages were higher, some were lower, and some remained the same when comparing these values from the first iteration to the second. The overall kappa coefficients for all four experts combined were 0.37 for the first iteration and 0.44 for the second iteration. The kappa coefficient for the second day was slightly higher indicating greater agreement among the experts after their discussion.

Distributions of relative weights that resulted from applying least squares regression analysis to experts' judgments elicited during each of the two rounds of ratings summed to unity for each expert and can be interpreted as proportions. Those proportions or weights for each of the experts are presented

OSCE measuring these six skills (each expressed as a percentage score):

$$\text{Score (\%)} = (.158)(\text{RB}) + (.089)(\text{IM}) + (.153)(\text{SA}) + (.199)(\text{AL}) + (.144)(\text{AF}) + (.257)(\text{CG})$$

Using the percentage scores for the six communication skills reported for the hypothetical examinee in Figure 1, his/her overall percentage score for the assessment would be:

$$(.158)(90\%) + (.089)(75\%) + (.153)(18\%) + (.199)(10\%) + (.144)(70\%) + (.257)(30\%) = 43.43\%$$

Since this process does not provide a minimum passing score for the determination of competence, the examinee cannot be declared competent or incompetent based on this overall percentage score alone.

Table 3
Component Weights for Four Experts with Means and Standard Deviations

First Iteration							
Expert	RB*	IM†	SA‡	AL§	AF¶	CG**	Total
1	.175	.159	.159	.161	.163	.182	.999
2	.104	.024	.136	.215	.148	.372	.999
3	.242	.193	.140	.061	.173	.191	1.000
4	.159	.091	.214	.281	.165	.090	1.000
Mean	.170	.117	.162	.179	.162	.209	.999
SD	.057	.075	.036	.093	.010	.118	
Second Iteration							
Expert	RB*	IM†	SA‡	AL§	AF¶	CG**	Total
1	.168	.064	.256	.166	.173	.173	1.000
2	.111	.042	.089	.273	.091	.395	1.001
3	.203	.141	.113	.148	.183	.212	1.000
4	.149	.110	.155	.210	.130	.248	1.002
Mean	.158	.089	.153	.199	.144	.257	1.000
SD	.038	.045	.074	.056	.042	.097	

*RB – Rapport Building
 ‡SA – Setting the Agenda
 ¶AF – Addresses Feelings

†IM – Information Management
 §AL – Active Listening
 **CG – Reaching Common Ground

Discussion

The JPC procedure is a viable approach to “capture the policy” of experts in assessing a clinical skill that is dependent on several components that are probably not of equal importance. In such situations, JPC provides appropriate weights for the components.

Without using JPC, the easiest way for an evaluation director to handle the weighting of various components would be to assign equal weight to each. Using the percentage scores for the six communication skills presented in Figure 1, the assignment of equal weight to each score would be accomplished by averaging the six percentage scores arriving at an overall percentage score of 48.83 percent. In doing so, if the evidence and experience of experts suggest that distinctions should be made concerning the value of each component, then the resultant scores may misrepresent the true competency of the examinee and provide an invalid score as a final end product from an otherwise valid assessment process. For example, suppose an examinee had the following percentage scores: Rapport Building – 55, Information Management – 90, Setting the Agenda – 50, Active Listening – 65, Addresses Feelings – 70, Reaching Common Ground – 30. This examinee would have an overall score of 60 percent if the six skills are given equal weight and the scores are simply averaged but his/her overall score using the component weights from the equation above would be 55.1 per-

cent. Now suppose that the examinee had the percentage scores for Information Management and Reaching Common Ground reversed. The overall score would still be 60 percent if the skills are given equivalent weights but it would be 65.2 percent using the weights from the equation.

JPC is a formal means of providing an appropriate weighting of the evaluation components. By providing the appropriate set of weights for the evaluation components to the examinees in advance, the learners are provided with a priority of emphasis in learning and practicing various skills.

Two iterations were used in this particular JPC procedure. It does appear that the group moved closer to consensus on the second iteration after some discussion of their values and ideas. However, JPC could be done without the iterative or discussion component. It is our belief that two iterations provided an appropriate set of component weights but we have no clear justification for recommending one, two or even three iterations of the process. It is important to note again that this is not necessarily intended to be a consensus building process.

One of the strengths of this empirical procedure is that it does not have to be done for each OSCE that is constructed for measuring these same six communication skills. It would be appropriate to repeat the process occasionally to verify that experts still view the skills in the same way; it would also be necessary

to repeat it if the skills or their definitions are changed.

It should be pointed out that the weighting of these skills represents the captured policy of the specific experts gathered for this exercise. All participating experts in this study endorse a "Patient Centered Model of Medical Communications."¹¹ Their emphasis on active listening to the patient's perspective and the need to use skills to negotiate common ground in situations in which the patient and clinician have different perspectives represent central tenants of this model. It is possible that other experts emphasizing different conceptual models might place greater importance in other areas. Institutions whose communications model and instruction differ significantly from that of our experts would be advised to capture the values of that model as we did here with the Patient Centered Model.

Additional applications for JPC come readily to mind. For example, for most graded clinical rotations, students are evaluated on a variety of distinct skills like history taking, differential diagnosis, interpersonal relationships, and learning skills. How these skills assessments are valued or combined in determining an overall clerkship grade is often poorly defined. JPC can help with such value determinations and appears to have a place in medical education.

Such a process was used for determining appropriate weights for the end-of-clerkship evaluation of third year family medicine students at the authors' institution utilizing the input of several clinical faculty members. The components in the evaluation were: 1) history, data collection, and physical examination skills, 2) assessment and plan, 3) learning habits, 4) interpersonal relationships, 5) responsibilities to the rotation. The weights derived from the regression analysis of individual faculty members' responses to profiles were quite different from the subjective weights that they provided before rating the profiles.

JPC can be a valuable tool for determining appropriate weights for the various components of complex assessments so common in undergraduate and graduate medical education. It might be particularly useful at the graduate medical education level with the current emphasis on the assessment of core competencies. The process for gathering judgments from participants is straightforward and the data analysis technique is fairly simple and available in any statistical software package.

Acknowledgements

The following individuals made contributions that were invaluable to this work: Lawrence R. Fischetti, Ph.D., Ann C. Jobe, M.D., Christine C. Matson, M.D., James A. MacKenzie, Ph.D., Kathy Zoppi, Ed.D. Dr. Ronald McCord was an invaluable member of our research team; his contributions are greatly missed.

A National Board of Medical Examiners Medical Education Research Fund Grant provided financial support for this work.

References

1. Hobson CJ, Gibson FW. Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review* 1983;8(4):640-649.
2. Roeber E. Setting standards on alternate assessments (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. 2002. Retrieved January 6, 2003, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html>.
3. Hammond KR, Hursch C, Todd F. Analyzing the components of clinical inference. *Psych Rev* 1964;71(6):438-456.
4. Zedeck S, Kafry D. Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance* 1977;18(2):269-294.
5. Jaeger RM. Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educ Measurement: Issues & Practice*, 1995;14(4):16-20.
6. Clauser BE, Subhiyah RG, Nungester R.J, Ripkey DR, Clyman SG, McKinley D. Scoring a performance-based assessment by modeling the judgments of experts. *J Educ Measurement*, 1995;32(4):397-415.
7. Simpson M, Buckman R, Stewart M., Maguire P, Lipkin M, Novack D, Til, J. Doctor-patient communication: The Toronto consensus statement. *British Med J* 1991;303(6814):1385-1387.

8. Makoul G. Essential elements of communication in medical encounters: The Kalamazoo consensus statement. *Acad Med* 2001;76(4):390-393.
9. Lang FF, McCord RS, Harvill LM, Anderson D. Communication assessment using the common ground instrument: Psychometric properties. *Family Med* 2004;36(3):189-195.
10. Fleiss JL *Statistical methods for rates and proportions*, second edition. New York: John Wiley, 1981.
11. Stewart M, Brown JB, Weston W, McWhinney I, McWilliam C, Freeman T. *Patient-*

centered medicine: Transforming the clinical method. Thousand Oaks, CA: Sage Publications, Inc, 1995.

Correspondence

Dr. Leo M. Harvill
Department of Medical Education
P. O. Box 70571
James H. Quillen College of Medicine
East Tennessee State University
Johnson City, Tennessee 37614-0965

423-439-6231 (Voice)

423-439-8004 (Fax)

harvill@mail.etsu.edu