The reliability of a hypothesis generation and testing task

M. B. DONNELLY, J. C. SISSON† & J. O. WOOLLISCROFT†

Department of Postgraduate Medicine and Health Professions Education and †Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan

Summary. The purpose of this paper is to present results of initial experience with a clinical reasoning task which assesses two clearly defined aspects of clinical problem solving. Fourteen senior and 40 junior medical students at the University of Michigan Medical School participated in this study. They were given three clinical reasoning problems — the hypothesis generation and testing tasks (HG&T). As suggested by the name, two specifically defined components of clinical problem-solving, developing the initial hypotheses or differential and then testing hypotheses, were evaluated by these tasks. The findings of this study indicate that hypothesis generation and testing can be reliably evaluated with between seven and ten tasks. The results of this study suggest that reliable assessments of specific components of clinical problem-solving can be developed.

Key words: *clinical competence; *problem solving; education, medical undergraduate/ *methods; Michigan

Introduction

The degree to which problem-solving skills can be generalized is one of the central issues in the conceptualization and assessment of clinical problem-solving. Measures of performance across problems typically have yielded relatively low reliability coefficients (Elstein *et al.* 1978; Berner 1984; McGuire 1985). Generally, these findings have been interpreted as indicating that problem content is the major determinant of performance and that a student's knowledge of the content varies from problem to problem. Thus, generalized skills in clinical problem-solving do not appear. However, Donnelly *et al.* (1982), upon finding this same pattern of results in the limited domain of chronic obstructive pulmonary disease, demonstrated that the commonly applied complex weighting schemes had an adverse effect on estimated reliability. Their results suggested that there may be other factors which explain the apparent lack of generalized skills.

In this paper, the effect of task complexity will be explored. The general literature on simulations (Fitzpatrick & Morrison 1971) has indicated that the fidelity of simulations is inversely related to the reliability of simulations; this is due most likely to the increasing complexity of both methods and goals. Thus, it may be that simulations used in medical education are paradoxically of too high a fidelity to obtain satisfactory reliability coefficients. It may be necessary to design tasks which are more limited in the dimensions or aspects of clinical problem-solving that they assess.

Clinical problem-solving performance is usually measured on high fidelity tasks, each of which encompasses a variety of underlying skills (McGuire & Babbott 1967; Harless *et al.* 1971; Stillman *et al.* 1986). Greater consistency in performance across problems may be observed by identifying and measuring the specific components of clinical problem-solving being assessed in a high fidelity clinical simulation and then comparing performance across problems on equivalent components. That is, the measurement of explicitly defined and carefully circumscribed judgemental task(s) in a set of

Correspondence: Michael B. Donnelly PhD, Department of Postgraduate Medicine/Health Professions Education, 1212 Towsley Center, University of Michigan Medical School, Ann Arbor, Michigan 48109–0201, USA.

clinical problems may lead to higher between-problem reliability.

The purpose of this paper is to present results of our initial experience with a clinical reasoning task which assesses two clearly defined aspects of clinical problem-solving. Specifically, we examined whether or not clearly defining students' performance on circumscribed components of a problem-solving exercise leads to more reliable results than typically have been found with more global measures of clinical problem-solving.

Methods

Fourteen senior and 40 junior medical students at the University of Michigan Medical School, Ann Arbor, Michigan participated in this study. They were given three clinical reasoning problems — the hypothesis generation and testing tasks (HG&T). As suggested by the name, two specifically defined components of clinical problem-solving were evaluated by these tasks. First, given the presenting complaint of a patient, the students were asked to generate diagnostic hypotheses that they would initially test in the medical interview and the physical examination. These hypotheses were to be listed in rank order of probability. The students were also told: 'In answering these questions, think as you would in an encounter with a patient who seeks your help. Generate only those hypotheses that you would regularly test in your interview and physical examination of such a patient . . . Assume that the patient is at least a high school graduate and has lived in Michigan for many years. You should assume that the patient does not have a rare disease and has not had a medical evaluation for his/her chief complaint within the past 3 months.' The content of the three problems was recurring chest pain, loose stools and productive cough.

The second type of task required students to indicate how they would test a specific diagnostic hypothesis from those likely to have been generated in the first reasoning task (e.g. angina pectoris or ulcerative colitis) through the medical interview and the physical examination. They were instructed to do this in terms of what questions they would ask or what physical examination manoeuvres they would perform in order to identify specific manifestations, complications and aetiologic factors of the specified problem. Although the two tasks were interrelated, the hypothesis generation and hypothesis testing tasks were administered independently so that retracing was not possible. Each student was given a sample problem with scoring guidelines prior to the administration of these tasks. Participation was voluntary and performance on the HG&T did not influence the students' clerkship evaluation.

Very important in the design of these problems is that they measured recall and application and not recognition of information. While this approach increased the difficulty of scoring the problems, we believe that such tasks more validly assess clinical problem-solving or reasoning as compared to recognition tasks (Feightner & Norman 1978; Norman & Tugwell 1982).

Task selection was based on a model of diagnostic reasoning which the authors are currently developing. In this model, hypothesis generation and testing are viewed as the basic elements in clinical reasoning. It should be noted that in positing these basic components, the hypothetico-deductive model is not assumed. Rather, hypothesis generation and testing are understood in terms of cognitive continuum theory (Hamm 1988).

To determine interrater reliability in scoring these tasks, a sample of 14 students' performances on the HG&T tasks was evaluated by three different raters (a doctor, a nurse and a senior medical student). A check-list was developed in which the rater simply indicated whether a given element of the answer was present or not — a '1' or a '0'. Reliability of scoring was judged by means of coefficient α and was calculated separately for each of the 14 students; five of the αs were above 0·90 and 13 above 0·80. Only one was below 0·80 and it was 0·79. Thus, it was possible reliably to score performance on these problems.

A per cent score was developed for each problem/task by summing the number of check-list elements which were correctly identified and dividing by the maximum possible . Other items identified by the students were given a weight of '0' and did not affect the student's score. Thus, three scores were developed for the hypothesis generation tasks, one for each problem.

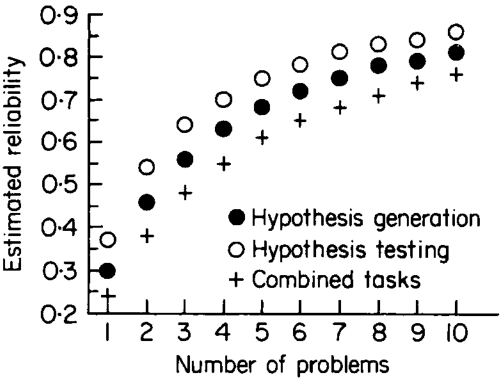

Fig. 1. Spearman-Brown estimates of the reliability of the HG&T tasks.

Hypothesis testing was evaluated on only two of the problems; this gave two scores for the medical interview and two for the physical examination. These scores form the basis for subsequent analyses.

The major statistical analyses involved estimating the reliability (coefficient α, intraclass correlation, and the Spearman-Brown formula) of the hypothesis generation and hypothesis testing tasks. The dimensionality of these tasks was determined by an exploratory principal axes factor analysis with varimax rotation. Finally, evidence for the construct validity of the HG&T was determined by means of a two-way analysis of variance with one repeated measure.

Results

Table 1 presents the mean score for each of the problem's sections. As might be expected with uncued recall tasks, average student performance was not at a high level; and this was especially true of the hypothesis testing sections (the history and physical examination).

The reliabilities of the hypothesis generation, hypothesis testing, and the two tasks in combination are presented in Table 2. This table contains the intraclass correlation (the reliability of a single score), coefficient α and the number of problems (*k*) used in the estimation of each reliability coefficient. In an absolute sense, the reliabilities are not very high (0·50s and 0·60s). However, when one considers that these are based on relatively few problems in each task (column *k* of the table) they are higher than might be expected for clinical problem-solving tasks. Also, note that the intraclass correlations are higher for each of the individual tasks than when they are combined (0·30 and 0·37 versus 0·24). This suggests that it is best to score the hypothesis generation and hypothesis testing tasks separately.

Figure 1 presents the Spearman-Brown estimates for the reliability of the hypothesis generation and testing tasks singly and in combination for from 1 to 10 problems. From this figure it can be seen that about seven problems are required to obtain a reliability 0·80 for the hypothesis testing tasks and 10 problems for the hypothesis generation tasks. The estimated number of problems to obtain a 0·80 reliability for the combined tasks is somewhat larger, 13 problems. In any case, these tasks, whether singly or in combination, do appear to be more reliable than have been reported previously for more complex clinical simulations.

Table 1. Student performance scores by task

Section	Mean (%)	Standard deviation (%)	*n*
Hypothesis generation			
Case 1 (chest pain)	59	17	54
Case 2 (loose stools)	31	20	54
Case 3 (productive cough)	56	26	54
Hypothesis testing			
History 1 (angina pectoris)	31	10	54
Physical 1 (angina pectoris)	13	14	54
History 2 (ulcerative colitis)	15	10	54
Physical 2 (ulcerative colitis)	8	9	54

Table 2. Reliability of hypothesis generation and hypothesis testing tasks

	Intraclass correlation	α	*k*
Hypothesis generation	0·30	0·56	3
Hypothesis testing	0·37	0·67	4
Hypothesis generation and testing combined	0·24	0·68	7

In order to explore further the importance of scoring the two dimensions, or tasks, separately, a principal axes factor analysis of the hypothesis generation and testing scores was performed. If two scores, a hypothesis generation score and a hypothesis testing score, are necessary to represent performance reliably then two factors ought to emerge. Table 3 presents the results of the factor analysis. There was only one eigenvalue greater than 1·00 and when more than one factor was extracted it was not interpretable. Thus, it was concluded that the one factor solution presented in Table 3 is a reasonable representation of the relationship among the measures; however, as noted below, the two tasks were differentially difficult.

Finally, preliminary estimation of the construct validity of these tasks was obtained by using the construct of experience. It was hypothesized that if these tasks are valid then senior medical students ought to perform better on them than do junior medical students. Table 4 presents the analysis of variance testing this hypothesis. As can be seen in this table, the senior students' performances were significantly better on both the hypothesis generation and hypothesis testing tasks than were those of the junior medical students. Thus, there is evidence for the construct validity of the HG&T. The tasks were not equivalent (i.e. not equal in difficulty) as is shown by the significant differences between the two task types. The students did much better on hypothesis generation as compared to hypothesis testing.

Discussion

The primary purpose of this study was to determine if a clinical problem-solving/reasoning task with clearly defined judgemental tasks would produce more reliable/generalizable results than do the currently used higher fidelity simulations. Swanson *et al.* (1987) determined that up to 40 high fidelity simulations may be necessary to estimate performance reliably (i.e. α ≥ 0·80). The preliminary data presented in this study indicate that hypothesis generation and testing performance can be reliably evaluated with between seven and ten tasks. It would appear, then, that the tasks described here are more reliable than have been reported for the more complex simulations. Further, the intraclass correlations indicate that estimating the reliability of the two tasks separately is somewhat better than combining them. This latter result favours our hypothesis that reliability is negatively related to task complexity.

We found the results of the factor analysis of the seven HG&T scores indicating one factor rather than two to be of interest. As is suggested by both the reliability and the analysis of variance analyses, there appears at least to be differential performance on the two tasks. However, the factor analysis suggests that hypothesis generation and testing are relatively highly intercorrelated resulting in only one factor. One could

Table 3. Factor analysis of HG&T scores

Section	Varimax Factor Loading
History 1	0·51
Physical 1	0·48
History 2	0·61
Physical 2	0·77
Diagnosis 1	0·12
Diagnosis 2	0·69
Diagnosis 3	0·58

λ 2·27
Percentage of variance 32%

Table 4. Two-way analysis of variance comparing educational level and task type (repeated measure)

Source	df	MS	*F*
Educational level (A)	1	0·149	8·38*
Error$_1$	52	0·018	
Task type (B)	1	2·018	287·09*
AXB	1	0·002	0·26
Error$_2$	52	0·010	

*$P < 0·01$

Mean scores

Student group	Hypothesis generation	Hypothesis testing
Juniors	0·47	0·14
Seniors	0·54	0·24

argue that the factor analysis should not have been performed given the relatively small sample size; but a similar conclusion would be reached by just inspecting the bivariate, zero-order correlations. We feel that these analyses suggest that there is a basic underlying feature, likely knowledge, which is critical to the performance of both the hypothesis generation and hypothesis testing tasks. However, the application of this knowledge to different clinically relevant tasks does differ. Students were much better at developing a differential diagnosis than the processes utilized to test the applicability of that hypothesis to a particular patient situation.

The findings in this study appear to parallel the findings in prior studies that knowledge is highly correlated with student performance in clinical problem-solving tasks. Unlike the findings from prior studies, however, the division of the clinical problem-solving exercise into specific tasks allows relatively few problems to be utilized to develop reliable measures of student performance. An important point is that this study tried to emulate the realities of the clinical situation. Unlike some studies of student clinical problem-solving which emphasize the manipulation of the data, this study emphasized a very basic level of processing. Furthermore, an open-ended format allowed the students, in an uncued fashion, to record their thoughts on how they would pursue further clinical evaluation.

To summarize, the results of this study suggest that reliable assessments of specific components of clinical problem-solving can be developed. Initial analyses suggest that there are clearly defined basic processes that can be identified. However, there also appear to be underlying unifying concepts or skills.

References

Berner E. (1984) Paradigms and problem solving: a literature review. *Journal of Medical Education* **59,** 625–33.

Donnelly M., Fleisher D., Schwenker J. & Chen C (1982) Problem solving within a limited content area. In: *Proceedings of the Twenty-First Annual Conference on Research in Medical Education,* pp. 51–6. Association of American Medical Colleges, Washington, DC.

Elstein A., Shulman L. & Sprafka S. (1978) *Medical Problem Solving: An Analysis of Clinical Reasoning.* Harvard University Press, Cambridge.

Feightner J. & Norman G (1978) Computer-based problems as a measure of the problem solving process. In: *Proceedings of the Seventeenth Annual Conference of Research in Medical Education,* pp. 51–6. Association of American Medical Colleges, Washington, DC.

Fitzpatrick R. & Morrison E (1971) Performance and product evaluation In: *Educational Measurement* (ed. by R. Thorndike), pp. 237–70. American Council on Education, Washington, DC.

Hamm R. (1988) Clinical intuition and clinical analysis: expertise and the cognitive continuum. In: *Professional Judgment* (ed. by J. Dowie & A. Elstein), pp.78–105. Cambridge University press, Cambridge.

Harless W., Drennon G., Marxer J., Root J. & Miller G. (1971) CASE: a computer-assisted simulation of the clinical encounter. *Journal of Medical Education* **46,** 443–8.

McGuire C. & Babbott D. (1967) Simulation technique in the measurement of problem solving skills. *Journal of Educational Measurement* **4,** 1–10.

McGuire C. (1985) Medical problem-solving: a critique of the literature. *Journal of Medical Education* **60,** 587–95.

Norman G. & Tugwell P. (1982) A comparison of resident performance on real and simulated patients. *Journal of Medical Education* **57,** 708–15.

Stillman P. *et al* (1986) Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine* **105,** 762–71.

Swanson D., Norcini J. & Grosso L. (1987) Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* **12,** 220–46.

Received 7 *October* 1988; *editorial comments to authors* 25 *November* 1988; *accepted for publication* 1 *May* 1990