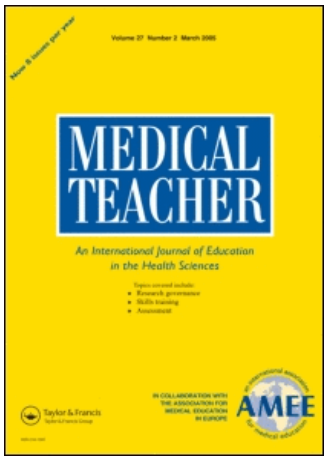

Medical Teacher
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713438241

Composition of the panel of reference for concordance tests: Do teaching functions have an impact on examinees' ranks and absolute scores?

To cite this Article: Charlin, Bernard, Gagnon, Robert, Sauvé, Evelyne and Coletti, Michel , 'Composition of the panel of reference for concordance tests: Do teaching functions have an impact on examinees' ranks and absolute scores?', Medical Teacher, 29:1, 49 - 53
To link to this article: DOI: 10.1080/01421590601032427
URL: http://dx.doi.org/10.1080/01421590601032427

2007; 29: 49–53 MEDICAL TEACHER

Composition of the panel of reference for concordance tests: Do teaching functions have an impact on examinees' ranks and absolute scores?

BERNARD CHARLIN[1], ROBERT GAGNON[1], EVELYNE SAUVÉ[1] & MICHEL COLETTI[2]
[1]University of Montreal, Canada, [2]University of Bobigny, France

Abstract

Background: Concordance tests are designed to assess the component of uncertainty of clinical reasoning. Scoring is based on a comparison of examinees' answers with those of a panel of reference, including their variability. This allows construction of tests that are close to real clinical life, with all its complexity and ambiguity.

Aim: This study was carried out to determine the effect of teaching functions of members composing the reference panels on students' scores and ranking.

Methods: A group of 80 residents in family medicine from a French University (Bobigny) completed a 72-item concordance test. The answers of two panels, each made up of 29 family physicians (teaching function *versus* non-teaching function), were used to generate the correction keys.

Results: Correlation between the sets of data obtained with the two panels is high (ICC = 0.98). Concordance scores obtained from the teaching-function panel are higher than scores obtained from the non-teaching-function panel (72.0 *versus* 76.3; $p < 0.001$). Ranking provided by the two panels was very similar.

Conclusions: This legitimizes the use of non-teaching physicians on panels. Panel composition influenced absolute score values: Residents showed more concordance with their academic trainers than with community-based physicians.

Problem statement and review of literature

Concordance tests (CJ) are designed to assess the component of uncertainty in clinical reasoning (Charlin & van der Vleuten, 2004). For each item, a clinical case is presented, containing either not enough data to solve the clinical problem (diagnostic, treatment), or data ambiguity, or conflict among values (assessment of ethical reasoning for instance). A series of questions is related to the case. Each contains an option relevant to the clinical problem, followed by the presentation of new data. Examinees' task is to assess the effect the new data have on the status of the option (Charlin et al. 2000). Subsequent questions within the case explore the effect of other data on other options. Figures 1 and 2 illustrate the test format.

A series of studies has revealed the usefulness of concordance tests to discriminate along levels of experience (Charlin & van der Vleuten, 2004; Gagnon et al. 2005), their applicability in domains as diverse as surgeons' per-operative reasoning (Meterissian et al. 2006), choice of treatment protocols in radio-oncology (Lambert, 2006) or reasoning in the emergency room (Carrière, 2005). Another study (Charlin et al. 2006) has shown that, for better discrimination along clinical experience, the best items are those that induce variability among panel of reference members. This deliberate introduction of variability permits the construction of tests containing tasks as complex and ambiguous as clinical reality often is.

However, this concept of variability raises an issue: the composition of concordance tests' panels of reference. Up to now this issue has not been formally investigated, given

Practice points

- Concordance tests assess complexity and ambiguity in clinical reasoning.
- Scoring is based on a comparison of examinees' answers with those of a panel of reference. This raises the issue of the composition of panels of reference.
- Should a panel be made up of residents' academic trainers or of in-practice physicians?
- Study results show that ranking is not modified, but residents show more concordance with their academic trainers than with community-based physicians

Correspondence: Bernard Charlin, URDESS Faculté de médecine-direction Université de Montréal, CP 6128, succursale centre-ville, Montréal, Québec, H3C 3J7, Canada. Tel: (514) 343-6111, extension 14140; Fax: (514) 343-7650; email: bernard.charlin@umontreal.ca

ISSN 0142–159X print/ISSN 1466–187X online/07/010049–5 © 2007 Informa UK Ltd.
DOI: 10.1080/01421590601032427

If you were thinking of	And then the patient reports or you find upon clinical examination	This hypothesis becomes
(a diagnostic hypothesis)	(new clinical information)	−2 −1 0 +1 +2

− 2 Ruled out or almost ruled out
− 1 Less probable
0 Neither less or more probable
+ 1 More probable
+ 2 Certain, almost certain

Figure 1. Format of items used for diagnostic knowledge assessment.

A 17 year-old-girl presents with dyspnea. She is out of breath and she is brought very rapidly to your office. At the onset of dyspnea, she was in a car returning home after a desensitization injection at her doctor's office.		
If you were thinking of	And then the patient reports or you find upon clinical examination	This hypothesis becomes −2 −1 0 +1 +2
Anaphylactic reaction	A respiratory rhythm at 32	−2 −1 0 +1 +2
Asthma	A difficulty in swallowing	−2 −1 0 +1 +2
Hyperventilation	A normal pharynx	−2 −1 0 +1 +2
Anaphylactic reaction	Arterial tension = 120/180	−2 −1 0 +1 +2
Asthma	A diffuse arterial murmur	−2 −1 0 +1 +2
Hyperventilation	Arterial tension = 150/90	−2 −1 0 +1 +2

−2 Ruled out or almost ruled out
−1 Less probable
0 Neither less or more probable
+ 1 More probable
+ 2 Certain, almost certain

Figure 2. Example of a clinical vignette and related items.

that most research conducted on the theme was realized (Charlin & van der Vleuten, 2004; Carrière, 2005; Gagnon et al. 2005; Lambert, 2005; Charlin et al. 2006, Meterissian et al. 2006) with reference panels composed of teachers belonging to the examinees' educational programs. A study conducted by Sibert (Sibert et al. 2002) in urology residency programs represents an exception. Scores of residents at a French university (Rouen) and a Canadian university (McGill) were compared on the same test, revealing a panel effect, with ranking being statistically similar but scores being higher for residents when they were judged by the panel from their own country.

The present study explored further the panel effect. It used two panels to score twice a concordance test given to in-training family-medicine residents One panel was made up of physicians with teaching functions in the residency program, the other was comprised of physicians with a private practice and no teaching function. The study goal was to examine the effect of the variation of panel composition on residents' scores and ranking. Similar ranking would legitimize the use of non-teaching physicians on CT panels. Similar scores would allow direct comparison of scores across examination conditions.

Methodology

Material

Test items covered the domain of family medicine and were aimed at the residency level. They were taken from the material used in two previous studies (Gagnon et al. 2005; Charlin et al. 2006). The test had 72 questions relating to 24 clinical cases (24 items, each with several questions). Questions covered diagnostic, investigation and treatment aspects. The test was given in French.

Scoring process

Reference panel members (referees) were asked to complete the test individually. Their answers were used to build the correction key, with the following methodology (Gagnon et al. 2005) for each question; the number of referees who had

If you were thinking of	And then the patient reports or you find upon clinical examination	Number of experts for each response				
		–2	**–1**	**0**	**+1**	**+2**
Anaphylactic reaction	A respiratory rhythm at 32	0	0	2	22	5
		0	**0**	**0.09**	**1**	**0.23**
Asthma	A difficulty in swallowing	12	12	3	0	2
		1	**1**	**0.25**	**0**	**0.17**
Hyperventilation	A normal pharynx	0	0	17	12	0
		0	**0**	**1**	**0.71**	**0**
Anaphylactic reaction	Arterial tension = 120/180	0	17	12	0	0
		0	**1**	**0.71**	**0**	**0**
Asthma	A diffuse arterial II/VI murmur	0	5	24	0	0
		0	**0.21**	**1**	**0**	**0**
Hyperventilation	Arterial tension = 150/90	0	0	22	7	0
		0	**0**	**1**	**0.32**	**0**

Figure 3. Answer key construction.

provided each answer was recorded. Examinees received a credit reflecting the number of referees who gave the same answer as theirs. The modal referees' choice(s) on each item equaled 1 point; other referees' choices provided a proportional partial credit. Answers not chosen by referees received zero. The process consisted of dividing all answers for an item by the modal value for that item. For example (see Figure 3), if on an item 22 referees (out of 29) had chosen the anchor '+1', this answer received 1 point (22/22). If two referees had chosen the anchor '0', this answer received 0.09 (2/22), and if five referee experts had chosen '+2', this answer received 0.23 points (5/22). The total score for the test is the sum of credit obtained for each item. Numbers were then transformed to get a maximum of 100.

Selection of members of the two panels

Thirty-eight physicians with a teaching function in the Family of Medicine program at the University of Bobigny were asked to complete the test during an academic meeting. Twenty-nine physicians in private practice, without teaching functions in respect of residents, were asked to complete the test during a continuing medical education (CME) meeting. All physicians agreed to participate.

Examinees

A group of 80 residents in their third year (last year) of training in Family Medicine at the University of Bobigny (France) were asked to complete the test. They all agreed to participate.

Statistical analysis

A previous study has shown that the absolute score on concordance tests varies with the number of referees on the panel (Gagnon et al. 2005). In order to compare equivalent panels, a random sample of 29 out of 38 academic physicians was selected with a random selection procedure in SPSS. Reliability of scales was estimated with Cronbach's alpha coefficient. It was calculated for both series of scores, one with scores obtained from non-academic and one with scores obtained from the academic panels.

The difference in the means of concordance scores obtained with the two panels was compared with a paired sample *t*-test, since the level of concordance of residents is measured under two situations (panel 1 vs. panel 2). The intra-class correlation (ICC) coefficient was used to assess the correlation between scores obtained with the two different panels.

Based on preliminary data, an anticipated number of 80 respondents was deemed sufficient to provide a power of 80% to detect a two-point difference between means on both answer keys (alpha = 0.05). This number of subjects was sufficient to detect a significant intra-class correlation higher than 0.27 (beta = 0.80; alpha = 0.05).

Table 1. Panel effect.

	Non-teaching	Teaching	*p*-value
Mean	72.0	76.3	<0.001
SD	9.1	8.5	
Median	73.6	78.2	
Minimum	41.2	51.1	
Maximum	87.7	96.9	
Range	46.5	45.8	

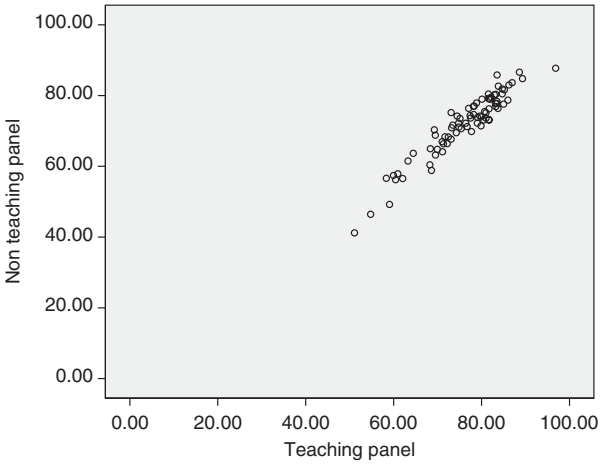

Figure 4. Distribution of scores with both panels.

Results

All 80 respondents' data were used in the analyses. There was no missing value. Values of the Cronbach's alpha coefficient for the test were the same with both panels (alpha = 0.70).

Data are summarized in Table 1 and depicted in Figure 4. The value of the intra-class correlation coefficient between scores obtained with the two different panels reveals a high relationship (ICC = 0.98). Concordance scores obtained from the non-teaching function panel of physicians are significantly lower than the score obtained from the panel of teaching physicians (72.0 *versus* 76.3; $p < 0.001$).

Discussion

In any assessment situation, jury composition has a major importance and competent persons who will provide reliable scores on student performance are recruited. With the modified aggregate scoring process (Norman, 1985; Norcini et al. 1990) used with the concordance test, this aspect becomes even more crucial. Test developers have to make a deliberate decision on who will be part of the panel of reference. For instance, if a test is taken for certification purposes at the end of training in family medicine, should the panel be made up of physicians who train residents or of physicians representing the profession that residents wish to enter?

This study, which compared scores obtained with such contrasted panels, showed a fairly high correlation between scores (0.98), meaning that classification of respondents does not differ significantly between the two panels. Even with the presence of a probable restriction of range effect, this represents a near perfect concordance in classification. This finding is important. It means that panels made up of physicians from the community provide similar rankings to academic physicians. This legitimizes their use on CT panels. This implication only held for community physicians who, like those in this study, take active means to maintain their competence and participate in continuing medical education events.

However, the study revealed a difference in the absolute value of scores. Higher scores were obtained with the teaching-function panel in comparison with the non-teaching function panel. An explanation for this effect may be that there is a higher concordance when examinees are assessed by the physicians who trained them. Members of teaching staff have many opportunities for exchange with residents, and therefore influence residents' reasoning processes. This induced more points when residents were scored by their panel. These results are similar to those of Sibert's study held in urology across two countries (Sibert et al. 2002), which found that ranking was statistically the same and that scores were higher when scores were computed with panels made up of trainers of the examinees' own country.

The present study has several limitations. The test was done in one specialty and one country, with a limited number of residents. The results have to be confirmed by other studies. If the stability of ranking across panels is confirmed, this will lend an argument of construct validity to the concordance test concept. The quality of CTs may be improved by asking a group of experts to determine a priori that certain responses were so clearly wrong that examinees should not receive credit for them. This would act as safeguard against haste or a lack of knowledge or judgement on the part of panel members. Roussel (Roussel et al. 2006) gave minus 1 point for such wrong choices by residents. This seems to enhance the reliability of the test. Such processes were not used in this study and remain to be tried.

In summary, when the assessment goal is to probe the component of uncertainty of clinical reasoning, the inclusion of non-teaching physicians on the reference panel is legitimate, given that they take means to maintain their competence. These results also suggest that concordance scores cannot be directly compared in respect of their absolute value. Residents show more concordance with their trainers than with community-based physicians. Any comparison has to be based on the distribution on the panel of referees' answers rather than on the performances of a group of examinees.

Acknowledgements

This research project was funded by a grant from the Medical Council of Canada. The study received ethical approval.

Notes on contributors

BERNARD CHARLIN is Professor of Surgery and Head of the Unit of Research and Development in Health Sciences Education at the University of Montreal.

ROBERT GAGNON is a psychometrician. He works as a research associate with the Continuing Medical Education Division at the University of Montreal.

ÉVELYNE SAUVÉ is a research assistant at the Unit of Research and Development in Health Sciences Education of the University of Montreal.

MICHEL COLETTI is a professor in the unit of Family Medicine of the University of Bobigny, France.

References

Carrière B. 2005. Development and initial validation of a script concordance test for residents in a pediatric emergency medicine rotation. Thesis, University of Illinois in Chicago.

Charlin B, Roy L, Brailovsky C, Van der Vleuten C. 2000. The Script Concordance Test, a tool to assess the reflective clinician. Teach Learn Med 12:189–195.

Charlin B, Van der Vleuten C. 2004. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. Eval Heal Prof 27:304–319.

Charlin B, Gagnon R, Sauvé E, Coletti M. & van der Vleuten, C. 2006. Assessment of clinical reasoning: is it necessary to accept variability of answers within the panel of reference to detect clinical experience? Med Educ, forthcoming.

Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. 2005. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ 39:284–291.

Meterissian S, Zabolotny B, Gagnon R, Charlin B. 2007. Is the Script-concordance test a valid instrument for assessment of intra-operative decision-making skills? American J Surg 4193:248–251.

Lambert, C. 2005. The script concordance test: validation study of a new tool to assess clinical reasoning of radiation oncology residents. Thesis, University of Montreal.

Norcini J, Shea J, Day S. 1990. The use of the aggregate scoring for a recertification examination. Evaluation in the Health Prof 13:241–251.

Norman G.R. 1985. Objective measurement of clinical performance. Med Educ 19:43–47.

Roussel F, Sibert L, N'Guyen P. et al. 2006. Évaluation du raisonnement clinique au cours du troisième cycle de Médecine Générale par le Test de Concordance de Script: étude comparative de deux corrections. Communication at the Congrès de la Société Internationale Francophone d'Éducation Médicale, Beyrouth, 1–2 June.

Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. 2002. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. Med Teach 24:522–527.