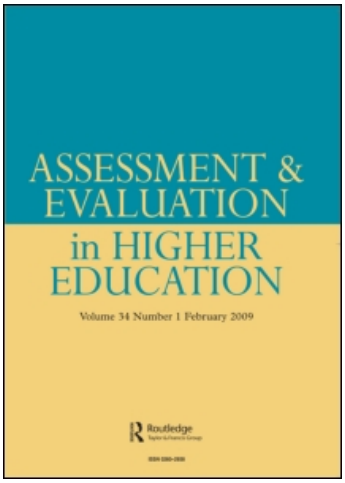

Assessment & Evaluation in Higher Education
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713402663

Assessment of competence in clinical reasoning and decision-making under uncertainty: the script concordance test method

Stephan Ramaekers[a]; Wim Kremer[b]; Albert Pilot[a]; Peter van Beukelen[b]; Hanno van Keulen[a]
[a] IVLOS Institute of Education, Utrecht University, 3508TC Utrecht, The Netherlands [b] Faculty of Veterinary Medicine, Utrecht University, 3508TC Utrecht, The Netherlands

Online publication date: 11 October 2010

To cite this Article Ramaekers, Stephan , Kremer, Wim , Pilot, Albert , Beukelen, Peter van and Keulen, Hanno van(2010) 'Assessment of competence in clinical reasoning and decision-making under uncertainty: the script concordance test method', Assessment & Evaluation in Higher Education, 35: 6, 661 — 673
To link to this Article: DOI: 10.1080/02602938.2010.500103
URL: http://dx.doi.org/10.1080/02602938.2010.500103

Assessment & Evaluation in Higher Education
Vol. 35, No. 6, October 2010, 661–673

Routledge
Taylor & Francis Group

Assessment of competence in clinical reasoning and decision-making under uncertainty: the script concordance test method

Stephan Ramaekers[a]*, Wim Kremer[b], Albert Pilot[a], Peter van Beukelen[b] and Hanno van Keulen[a]

[a]*IVLOS Institute of Education, Utrecht University, PO Box 80127, 3508TC Utrecht, The Netherlands;* [b]*Faculty of Veterinary Medicine, Utrecht University, PO Box 80127, 3508TC Utrecht, The Netherlands*

Real-life, complex problems often require that decisions are made despite limited information or insufficient time to explore all relevant aspects. Incorporating authentic uncertainties into an assessment, however, poses problems in establishing results and analysing their methodological qualities. This study aims at developing a test on clinical decision-making in veterinary medicine and establishing its reliability and validity. The test is based on the script concordance test method and covers a large sample of authentic cases and uncertainties. The answer key was compiled with reference to the professional judgements and decisions of a panel of experienced practitioners. From a substantive appraisal of the cases and items, the analysis of the test results and the responses from the experienced practitioners, it is concluded that this test validly represents the problems, decisions and uncertainties of clinical practice. In spite of the hindrances caused by the uncertainties included in the test, the reliability and validity of the test and its results could be evaluated and proved to meet measurement criteria.

Keywords: authentic assessment; competence development; decision-making under uncertainty; script concordance test

1. Introduction

Dealing with ill-defined problems and having to make decisions in uncertain situations, on the basis of limited information or under time pressure, is a part of everyday practice for many professionals (Eraut 2004; Jonassen 2004). To determine whether students are adequately prepared for this, the assessment of their problem-solving and decision-making capacities should include problems and circumstances which pose similar cognitive challenges.

Although authentic assignments and problems are considered valuable, particularly for the validity of an assessment (e.g. Linn, Baker, and Dunbar 1991), including real-life, open-ended problems and issues in an assessment, with uncertainties and possibly several solutions, creates various difficulties in establishing and analysing results. For example, how are good and poor student performances to be reliably distinguished when questions and answers contain ambiguities?

This study concerns the design of a test to measure progress in the development of competence in problem-solving and decision-making in situations of uncertainty and

*Corresponding author. Email: s.p.j.ramaekers@uu.nl

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602938.2010.500103
http://www.informaworld.com

evaluation of its measurement properties. The test was developed for a course in clinical problem-solving in veterinary medicine. Its design is based on the script concordance test (SCT) format developed by Charlin et al. (1998) to assess problem-solving and decision-making skills in realistic situations.

2. Theoretical foundations

2.1. The nature of clinical problem-solving and decision-making

The SCT format is grounded in theory and empirical research on clinical reasoning, problem-solving and the organisation of knowledge. How doctors analyse clinical problems, establish a diagnosis and decide about treatments has been studied since the late 1950s. Initially, systematically testing hypotheses, until explanations were found, was considered the essence of the problem-solving process. As some differences and similarities between experts and novices could not be explained by means of a superior reasoning process, research changed its focus towards the structure of expert knowledge (Neufeld et al. 1981).

The illness script theory assumes that experienced clinicians have their knowledge organised in coherent networks, 'scripts', covering numerous aspects of diseases, meaningful for practice. These scripts emerge through clinical experience and become, over the years, refined and rich in detail about particular patients, diseases, associated situations and enabling conditions (Custers, Boshuizen, and Schmidt 1996; Norman and Schmidt 1992). In this process, the knowledge of underlying biomedical principles and mechanisms and the causal reasoning at the base of judgements and decisions become embedded (encapsulated) into clinical concepts, but are still accessible if needed (Rikers, Schmidt, and Moulaert 2005). Comparing new cases with previous experiences and pattern recognition increasingly dominates the problem-solving process as expertise advances (Norman 2005).

Recognition of the complexity of real-world problems and human limitations in dealing simultaneously with too many different issues has fuelled research into the way decisions are made under uncertainty. Based on quantitative models and weighting of pre- and post-test probabilities, standards have been developed which describe an optimal (expert) approach to a particular clinical problem. Their values as methods for retrospective analysis of the decisions made, including reasoning fallacies and sources of bias, have been widely recognised (Hunink 2001). Criticism has been made of the limited applicability of these methods in a real-life clinical setting (Berg 1997; Elstein 2004).

Currently, most researchers agree that clinical problems are highly context-specific and that transfer from one problem solution to another is limited (Norman 2005). Finding appropriate solutions depends mainly on a knowledge base covering many different aspects of clinical problems and organised in structures, adjusted to practice (Elstein and Schwarz 2002). Experienced clinicians may solve their problems in a variety of ways; even in similar situations, they do not necessarily follow the same line of thought to achieve similar outcomes (Grant and Marsden 1988; Norman, Young, and Brooks 2007). Their strategies largely depend on pattern recognition and previous successful choices. They rarely use conscious reasoning, deduction or extensive testing (Forde 1998; Norman, Young, and Brooks 2007).

Circumstances which contribute to uncertainty are that decisions sometimes have to be made under time pressure or on the basis of very limited information. The reliability of information may be uncertain, results of patient tests may be inconclusive and a prognosis may not be predicted precisely (Eraut 2004; Forde 1998).

2.2. Rationale of the SCT format

The SCT format is designed to develop assessments of problem-solving competence in a way that fits current notions about clinical problem-solving and decision-making. SCTs supposedly measure correct interpretations of available data (Sibert et al. 2002), the extent and richness of mental 'scripts' (Charlin et al. 2000) and competence in testing hypothesis and decision-making under uncertainty (Charlin and van der Vleuten 2004). The problems that participants are presented with are chosen to match the issues, circumstances and cognitive challenges of real practice. Consequently, the design of SCTs fits into views on assessment (and learning) which emphasise the importance of a high level of authenticity for the validity of the assessment (e.g. Swanson, Norman, and Linn 1995; van der Vleuten 1996).

To incorporate real-life issues and problems, beyond the level of 'single right answer' questions, the appropriateness of solutions in an SCT is based on the professional judgements of a group of experts (reference panel). Several answers may be considered appropriate. The decisions of the participants are compared with those of the reference panel; the degree of agreement between the participants and the experts determines how answers are valued and indicates the participants' level of competence.

With the SCT format, tests have been constructed in various domains within medicine (e.g. Meterissian et al. 2007) and characteristics which have been studied are the timing of the assessment and the effects of different formats (Sibert et al. 2006), optimisations of the scoring methods (Charlin et al. 2002) and the composition of the reference panel (e.g. Gagnon et al. 2005; Nendaz et al. 2004). Results were compared between different levels of clinical experience, and also across different cultures and learning environments (Sibert et al. 2002). Furthermore, results on SCTs have been related to other indicators of clinical competence (e.g. Gagnon et al. 2006).

As regards the assessment of clinical competence in the transition phase from preclinical learning into internship, previous studies have shown that, despite an increase of clinical experiences, the performances on conventional tests do not show improvement (Boshuizen 2003; Patel, Arocha, and Zhang 2005). This phenomenon is referred to as the 'intermediate dip'. Two explanations have been suggested: a temporary lack of knowledge organisation owing to insufficient integration of practical experiences with theoretical knowledge; and shortcomings of conventional tests to measure problem-solving competence validly (Schmidt and Boshuizen 1993). The absence of this intermediate dip in the SCT, when we compare the results at different levels of experience, is considered an indication that supports the validity of the SCT with regard to clinical decision-making (Charlin et al. 1998).

2.3. Focus of this study

Previous studies of the SCT typically concerned a limited domain (a medical specialisation or a group of related conditions), participants with clinical experience and a comparison of scores between participants with different levels of experience. In this study, the SCT is applied on a broad domain (primary veterinary care), participants are undergraduates without substantial clinical experience and the scores of the same students on the same test, before and after a one-year course in clinical problem-solving, are compared. Against this background, the main issues this study addresses concern:

(1) The development of an SCT and its corresponding answer key to be used at undergraduate level to assess progress in problem-solving competence.

(2) The evaluation of the (internal-consistency) reliability of test results. Although ambiguity in the problems and answers of the test is conditional, this should not lead to doubt about the consistency of the measured results with regard to the students' performances. Furthermore, does repeated administration of the same test affect the participants' results?
(3) The evaluation of the (content) validity of the test. Do the cases and test items adequately represent the larger domain of the conditions, clinical decisions and uncertainties in primary veterinary care?
(4) The evaluation of the test sensitivity. Can the test detect changes in competence within the frame of a one-year course in solving clinical problems?

3. Methodology

3.1. Materials

In an SCT, problems and situations are described in short case vignettes. A vignette contains the main features of a case's first presentation and relevant aspects of its history which would be known in reality. Each case comes with four test items, formulated as a hypothesis or suggestion for action (Table 1).

Besides this hypothesis or proposed action, a test item holds additional information about the case. Participants are asked to assess the effect of the additional information on the plausibility of the hypothesis or the appropriateness of the proposed action. This entails carefully combining and weighing all available information.

3.2. Test development procedure

Development of the SCT included the following steps:

(1) The assessment matrix was based on epidemiological data concerning the clinical problems that frequently occur in primary veterinary care to achieve a representative sample of cases.

Table 1. Case vignette with two items.

'Carl', a six-year-old male Rottweiler dog, is presented to you. For three days, he has not eaten and vomits 5–8 times per day. According to his owner, he is usually a gobbler and never picky in what is served. He has not stopped drinking. Carl is kept as a family pet and allowed to walk about freely in and around the house, as long as he stays on the premises. First impression: an agitated dog with some signs of discomfort. There is no visible loss of weight. Pulse rate: 140/min (equal, regular); respiratory rate: 28/min (costo-abdominal); temp. 39°C; skin turgor: average–poor.

Suppose you consider this a case of:	and then you find that:	then this diagnostic hypothesis becomes:
b. stimulation of central receptors (due to poisoning)	despite fierce attempts, he hardly produces any vomit	−2 −1 0 +1 +2

−2 = very unlikely; −1 = less likely; 0 = not more nor less likely; +1 = more likely; +2 = very likely.

Suppose you consider for further assessment/treatment:	and the assessment of the patient revealed:	then this approach becomes:
d. abdominal X-ray	yellow mucosa + extended CRT (capillary refill time)	−2 −1 0 +1 +2

−2 = contraindicated; −1 = not advisable; 0 = not less nor more significant; +1 = advisable; +2 = indicated.

Clinical teachers, representing the main subdomains in veterinary medicine, provided the information needed to turn these clinical problems into realistic cases. Test items were chosen to reflect authentic biomedical and veterinary issues, including dilemmas related to owner preferences, ethical issues or time pressure.

(2) To disclose whether the students' unfamiliarity with the SCT format of items or the 'indifferent' answer category would affect their reasoning and choices, three trial sessions were conducted with fourth-year students from the previous cohort, following the 'think-aloud' procedure. These trials confirmed engagement of the students in the intended cognitive processes. Changes in the format or phrasing of cases were not indicated; the trials did, however, reveal the necessity for high-quality test instructions.

(3) The final version of the SCT in veterinary medicine (SCT-VM) was composed, covering 30 cases and 120 test items to create a sample large enough for the content to be tested. Previous studies (Charlin, Tardif, and Boshuizen 2000) indicated that an SCT covering a medical subdomain needs about 50–60 test items to achieve a reliability (Cronbach's α) of 0.80 or more.

(4) To establish the answer key, the test was completed by the reference panel. Based on previous studies (e.g. Gagnon et al. 2005), a minimum of 10 experts per subdomain (animal species) was regarded as sufficient. Inclusion criteria for the reference panel were: veterinary practitioner, non-teaching, with at least 10 years of clinical experience in primary veterinary care, acknowledged and recommended by colleagues (from university clinical staff). Thirty-five practitioners were invited to participate; 28 agreed and completed the test. For each expert, only the answers which concerned cases in their particular areas of expertise are included in the answer key.

(5) In addition to the test itself, a short questionnaire was developed for participant feedback, in particular about the SCT format of test items and the representativeness of the cases.

3.3. Context and participants

The SCT-VM was developed as an instrument to establish the progress students make in a course on clinical problem-solving, including practice with real patients, and covering most of the last (fourth) pre-clinical year before the clerkships (Utrecht University). The test is conducted twice, near the beginning and at the end of the course.

Students participate on a voluntary basis. Test results are neither part of the course assessment programme nor revealed to the teaching staff. The students receive individual feedback about their scores and guidance in the interpretation of results.

Of all the students in the course, 168 (97.7%) participated in the test; 148 of them in both the pre- and the post-tests. To avoid student performances being affected by unfamiliarity with this type of case description, the pre-test took place after the students had some opportunity to become accustomed to case vignettes in clinical tutorials (maximum seven).

3.4. Data analysis

(1) Development of the answer key:

(a) The *degree of concurrence between members of the reference panel* was analysed to identify the items that should be reviewed and, if necessary, excluded from the answer key. Large variability in answers may result from measurement error e.g. in the construction or phrasing of an item. The total concurrence indicates that the item does not involve an aspect of uncertainty.
(b) The *optimal scoring model.* The usual SCT scoring model is based on a score of one for the experts' modal answer, whereas the alternative answers receive a score corresponding to the proportion of panel members who choose the same alternative. Given some apparent patterns in the answers of the reference panel, alternative scoring models with a potentially better fit were studied to disclose their effects on the students' scores.

(2) Evaluation of reliability and validity:
(a) With the provisional answer key and scoring model, the estimated *internal-consistency reliability* and item-total correlations were calculated. Commonly used measures such as the discrimination index or distractor analysis were not used for item analysis, as they assume a single right answer. The individual scores of the panel members were checked to uncover deviant response patterns.
(b) If indicated (large variability in expert answers, low item–total correlation), items were reviewed independently by two senior veterinarians to reassess their *validity* (Borsboom, Mellenbergh, and van Heerden 2004).
(c) With the final answer key, the scores of participants were established and internal consistencies re-estimated.
(d) Generalisability theory provides methods to disentangle the contributions of multiple factors (e.g. the number of items) and their interactions with the reliability of results (Brennan 2001). To determine the reproducibility of test results and the effects of repeated use of the test, a G-study (variance component analysis) was conducted, based on a two-facet fully-crossed design with the items, participants and the two occasions as facets. A D-study projected the effects of changes in one of the facets with regard to optimisation of reliability.

(3) Evaluation of test sensitivity:
(a) Finally, the results of the pre- and post-tests were compared to disclose whether the test measured a significant change in competence.

4. Results

4.1. Test development: answer key and scoring model

With the panel members' responses, a provisional answer key was composed based on 12 experts in companion animals, 12 in farm animals and 11 in horses. This answer key showed a degree of concurrence between two-thirds of all experts on one alternative in 22 test items, and on two adjacent alternatives (e.g. 'very unlikely' and 'less likely') in 71 of the test items. In 17 items, the distribution of answers of the reference panel called for a review. Figure 1 illustrates the different degrees of item concurrence.

Close examination of the distribution in the experts' responses led to two hypotheses (a three-point answer scale provides sufficient differentiation; a modus score of one point is an overestimation) tested with four alternative scoring models. The effects of the alternative models on the averages and ranges of the reference panel and the

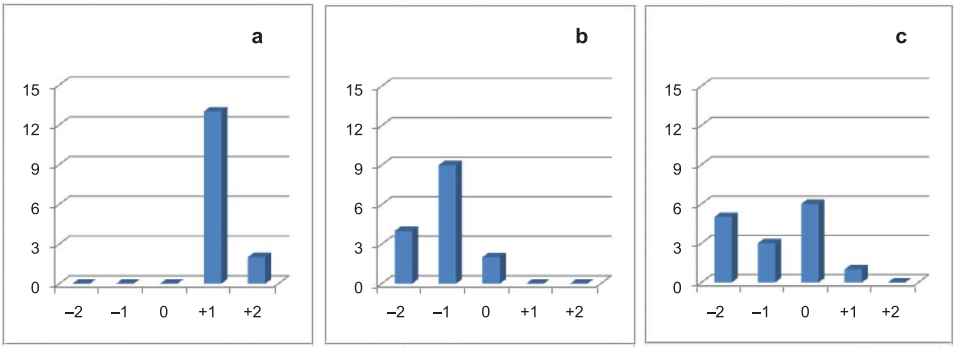

Figure 1. Variable degrees of concordance (expert responses): (a) large majority in one alternative, (b) majority in two adjacent alternatives and (c) other patterns.

students in the pre-test are presented in Figure 2. Correlations between results in Models 2, 3 and 5 with those of the SCT aggregate scoring model (Model 1) are strong ($r = 0.98$, $r = 0.89$ and $r = 0.96$, respectively; $p < 0.01$; $n = 164$). For Model 4, this correlation is moderate to strong ($r = 0.66$).

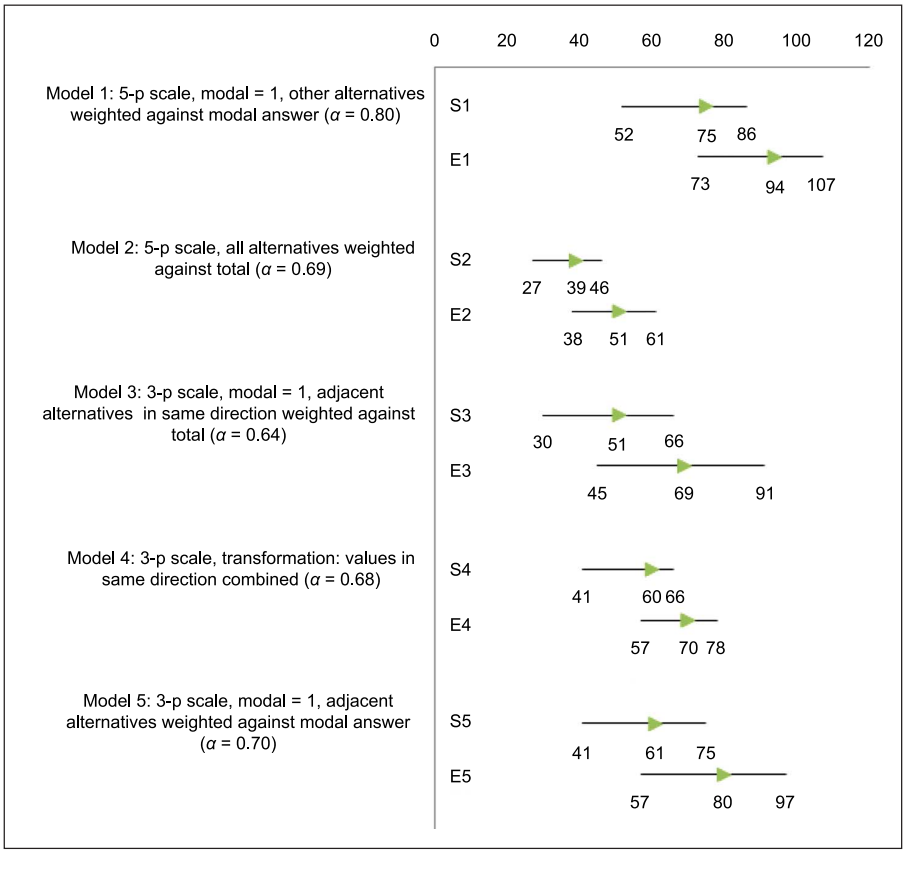

Figure 2. Effects of alternative scoring models on the pre-test results of students (S) and the expert panel (E): lowest score–mean–highest score.

Both types of adjustments in the scoring model resemble higher levels of concurrence between the experts and lead to a reduction of the scale range. The lower reliabilities (Cronbach's α) in these models might result from the information loss owing to non-valued responses. As the SCT-VM is intended to monitor competence development, reducing the scale range (differences between students) was avoided. Further analysis is based on Model 1.

4.2. Reliability and content validity

The review of the 17 items with a limited concurrence between panel members did not uncover apparent errors in the case or item construction, affecting validity. None of these items were removed in the final answer key.

The internal consistency (α) of the pre-test and the post-test is 0.80 and 0.79, respectively. Item-total statistics show that removal would not increase α by more than 0.004 for any of the items.

One practitioner, with over 40% outlier answers and a low personal score (<*M*–2SD), was excluded from the reference panel.

The results from the G-study about the generalisability of participant results and the relative contribution of different sources of variance are shown in Table 2. The *G*-coefficient indicates that 85.4% of the result-to-result variation is owed to real differences between the participants. The additional D-study established that a minimal 80 items would have been sufficient to obtain a reliability (*G*-coefficient) greater than 0.8; and if this test had been used only once, then 130 items would have been needed to achieve the same reliability.

4.3. Sensitivity to changes in competence

The students' scores improved from the pre-test ($M = 74.9$; SD = 5.5) to the post-test ($M = 79.6$; SD = 4.9). The improvement is significant ($t = 12.753$; df = 147; $p < 0.00025$). Furthermore, their individual scores on the pre- and post-tests correlate positively ($r = 0.653$; $n = 148$; $p < 0.001$) and the effect size is large (Cohen's $d = 0.89$).

Table 2. G-study: variance component analysis and generalisability.

Generalisability
Design type: two-facet fully-crossed design (P × F1 × F2)

Number of participants: 160
Number of items: 120
Number of occasions: 2

Source	df	ss	ms	Variance	Proportion
Participants (P)	159	156.641	0.985	0.0004	2.5%
Items (F1)	119	666.981	5.605	0.016	11.6%
Occasions (F2)	1	12.484	12.484	0.001	0.5%
P × F1	18,921	2749.176	0.145	0.027	19.2%
P × F2	159	12.522	0.079	0.000	0.0%
F1 × F2	119	38.523	0.324	0.001	1.0%
P × F1 × F2	18,921	1728.250	0.091	0.091	65.1%

Note: Error variances: relative(0.001) → absolute(0.001); *G*-coefficients: $G = 0.854 \rightarrow \varphi = 0.769$.

4.4. Participant feedback on using the test

The results from the questionnaire (Figure 3) show that the students and experts are more or less agreed on the authenticity of the cases (4.2 ± 2.0 on a five-point Likert scale) and on the perceived difficulty of the SCT format (3.8 ± 1.0). Students considered the cases more complex; they also perceived the test as knowledge-intensive, rather than reasoning-intensive. Monitoring progress in clinical reasoning is considered very useful (4.6) by the students.

5. Discussion

5.1. Reliability

The reliability criterion concerns the consistency of the measurements and results across the items within the test. A potential threat to the reliability of the SCT-VM results, including uncertainties in the test. They should reflect realistic uncertainties and the variability in responses which they cause should be distinguished from error

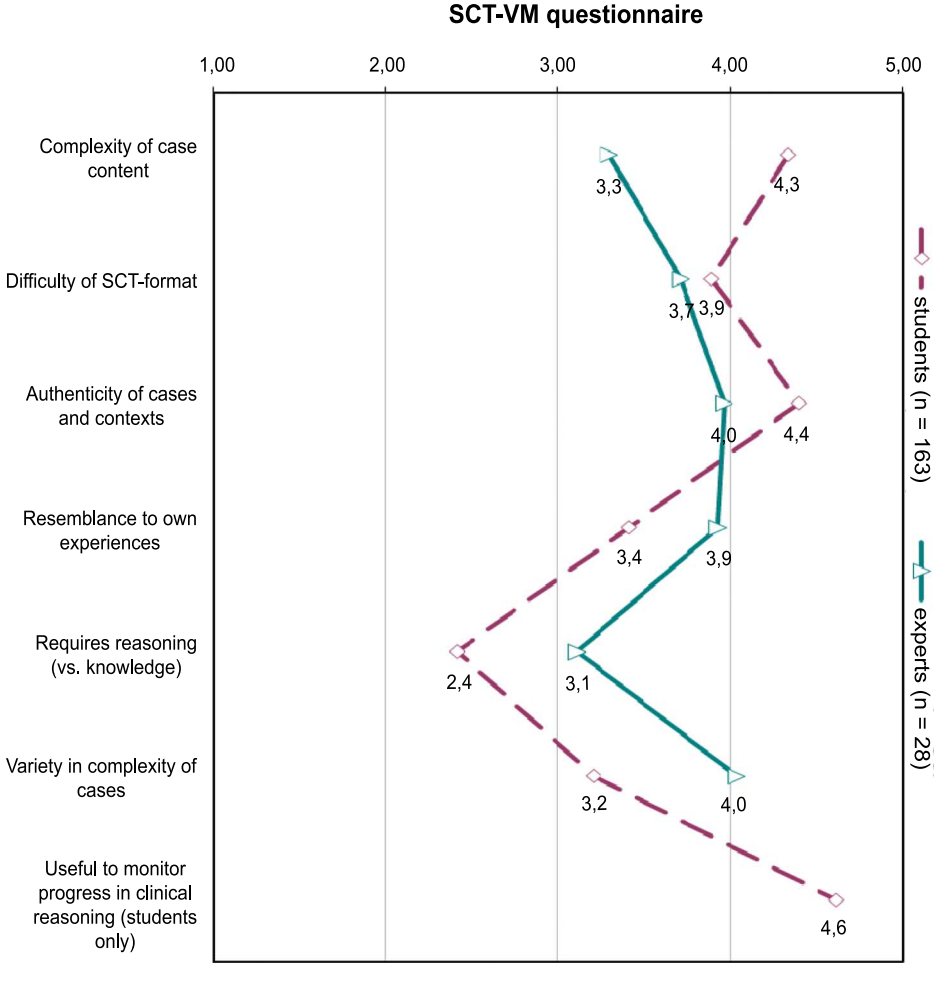

Figure 3. Results from the questionnaire.

in item construction or inconsistencies in the answer key. Ultimately, there should be no doubt as to whether test scores reflect the students' actual performances.

To achieve a high level of consistency of measurement, reliability issues have been reviewed repeatedly during test development up to the final evaluation:

- During the development of the SCT-VM and in the analysis of results, items with low concurrence between experts were reviewed to identify answer variability *owing to construction error*. No items were removed. One of the panel members, however, was excluded as this member's answers were beyond a reasonable level of distribution. Testing the effects of alternative scoring models confirmed the *fit of the classic SCT-model* with the data in the SCT-VM.
- The test results of the SCT-VM are based on a *substantial number* of cases and items and its *internal consistency* in both administrations is satisfactory (>0.79). The G-study, which combines different types of reliability analysis in one model, shows a *high generalisability* of results (0.85) and that repeated use did not affect test results. To assess progress with a pre- and a post-test, a total of 80 items would have been satisfactory, aiming at $G > 0.80$.

5.2. Validity

Appraisal of validity requires a substantive analysis of the instrument, relating test results to the content, processes and conditions of the competences to be measured (Borsboom, Mellenbergh, and van Heerden 2004).

A clear difference between written test formats and clinical problem-solving in practice, which may affect the reasoning processes, is the actual presence of a patient. Such presence requires attending to issues of comfort and safety and to communication with the owner, concurrently with the problem-solving process. Moreover, in an SCT, the hypotheses are already suggested, whereas in real practice, clinicians generate their own hypotheses.

Within these limitations, however, the findings in this study support in several ways the validity of the SCT-VM for assessing clinical problem-solving and decision-making:

- The SCT-VM contains a *large sample* of cases and items based on epidemiological data *representative* of the problems and conditions that veterinarians in primary care frequently encounter. Within this number of cases and items, *the different areas* of clinical judgements and decisions are covered. The authenticity of the problems and circumstances in the cases was confirmed by the experts from the reference panel and by the reviewers of items.
- The SCT-VM requires *cognitive activities similar* to those in practice: interpreting the information and weighing its reliability, reasoning about and recognising possible patterns, appraising the probability of hypotheses and alternatives, estimating the outcomes or effectiveness of interventions. The results of these activities are stated in terms of *judgements* or *decisions*. The think-aloud in the trial sessions and the students' feedback confirm engagement in the same activities and processes.
- The judgements and decisions of the *experienced practitioners* make up the reference against which student answers are compared. This allows real-life

problems and dilemmas, beyond the level of 'single right answer' issues, to be included in the test.

- *Coverage of the domain* of primary veterinary care was achieved by a reference panel with sufficient expertise from each subdomain. If the distribution of expert answers was beyond the expected range of differences, the case content was re-examined to disclose artificial uncertainties (e.g. lacking information which would be available in practice) or construction errors affecting test results.

6. Conclusion

In the light of the findings in this study, we conclude that the SCT-VM meets the described objectives and conditions. Hindrances related to the breadth of the domain to be covered as well as the limited clinical experiences of the students could be avoided. The results from using the same test twice made it clear that an SCT can be used as an instrument to monitor progress in problem-solving and decision-making competence.

The SCT-VM in this study was used formatively. In the case of an assessment with a summative function, students might have been more hesitant to participate in a test with ambiguities in the cases, questions and answers. How that would have influenced their choices in these cases is open to speculation.

The main limitations of an SCT concern the aspects of concurrent patient handling, communicating and problem-solving and a lack of necessity to generate one's own hypotheses. An assessment with real or simulation patients has better opportunities to include these aspects as well. Nevertheless, the SCT format has some important usability advantages; it is based on a large number of cases, can be administered comparatively easily and processed uniformly to a large numbers of students without creating a burden on real patients. These strengths, in our opinion, offset the limitations of the SCT. We recommend that the SCT format be used more widely in actual educational practices so that its features and applicability in other domains and its use for summative purposes may be further investigated.

Notes on contributors

Stephan Ramaekers is a lecturer and consultant on curriculum development in higher education at the IVLOS Institute of Education at Utrecht University. His PhD research focuses on the use of authentic, complex tasks to enhance the development of problem-solving competence.

Wim Kremer is a professor of farm animal health at the Faculty of Veterinary Medicine, Utrecht University. His research interests relate to the professional development of veterinarians. He is responsible for the master programme in farm animal health.

Albert Pilot is professor of curriculum development at the IVLOS Institute of Education, Utrecht University and professor of chemistry education in the Department of Chemistry, Utrecht University. His research focuses on curriculum development, professional development of teachers and context-based education.

Peter van Beukelen is a professor of quality improvement in veterinary education at Utrecht University. His research interests are: clinical reasoning, active learning, workplace and lifelong learning. Current research concerns staff development and assessment of teaching competence.

Hanno van Keulen is a lecturer in higher education development at IVLOS Institute of Education, Utrecht University. He is also a professor of science and technology education at

Fontys University of Applied Science, the Netherlands. His research interests are with staff and educational development in higher education and the innovation of science and technology education in primary education.

References

Berg, M. 1997. Problems and promises of the protocol. *Social Science and Medicine* 44, no. 8: 1081–8.

Borsboom, D., G.J. Mellenbergh, and J. van Heerden. 2004. The concept of validity. *Psychological Review* 111, no. 4: 1061–71.

Boshuizen, H.P.A. 2003. Expert development: The transition between school and work. In *Expert development: How to bridge the gap between school and work*, ed. H.P.A. Boshuizen, 7–38. Heerlen: Open University.

Brennan, R.L. 2001. *Generalizability theory.* New York: Springer.

Charlin, B., C.A. Brailovsky, L. Brazeau-Lamontagne, L. Samson, C. Leduc, and C. van der Vleuten. 1998. Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. *Medical Teacher* 20, no. 6: 567–71.

Charlin, B., M. Desaulniers, R. Gagnon, D. Blouin, and C.P.M. van der Vleuten. 2002. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine* 14: 150–6.

Charlin, B., L. Roy, C. Brailovsky, F. Goulet, and C. van der Vleuten. 2000. The script concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine* 12: 189–95.

Charlin, B., J. Tardif, and H.P.A. Boshuizen. 2000. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine* 75: 182–90.

Charlin, B., and C.P.M. van der Vleuten. 2004. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. *Evaluation and the Health Professions* 27, no. 3: 304–19.

Custers, E.J.F.M., H.P.A. Boshuizen, and H.G. Schmidt. 1996. The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates. *Memory and Cognition* 24, no. 3: 384–99.

Elstein, A.S. 2004. On the origins and development of evidence-based medicine and medical decision making. *Inflammation Research* 53, Suppl. no. 2: S184–9.

Elstein, A.S., and A. Schwarz. 2002. Evidence base of clinical diagnosis – clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal* 324, no. 7339: 729–32.

Eraut, M. 2004. *Developing professional knowledge and competence.* London: RoutledgeFalmer. (Orig. pub. 1992.)

Forde, R. 1998. Competing conceptions of diagnostic reasoning: Is there a way out? *Theoretical Medicine and Bioethics* 19, no. 1: 59–72.

Gagnon, R., B. Charlin, M. Coletti, E. Sauve, and C. van der Vleuten. 2005. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education* 39, no. 3: 284–91.

Gagnon, R., B. Charlin, L. Roy, M. St-Martin, E. Sauvé, H. Boshuizen, and C. van der Vleuten. 2006. The cognitive validity of the script concordance test: A processing time study. *Teaching and Learning in Medicine* 18, no. 1: 22–7.

Grant, J., and P. Marsden. 1988. Primary knowledge, medical education and consultant expertise. *Medical Education* 22: 173–9.

Hunink, M.G.M. 2001. In search of tools to aid logical thinking and communicating about medical decision making. *Medical Decision Making* 21, no. 4: 267–77.

Jonassen, D.H. 2004. *Learning to solve problems: An instructional design guide.* San Francisco: Pfeiffer.

Linn, R.L., E. Baker, and S.B. Dunbar. 1991. Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher* 16: 1–21.

Meterissian, S., B. Zabolotny, R. Gagnon, and B. Charlin. 2007. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *American Journal of Surgery* 193, no. 2: 248–51.

Nendaz, M.R., A.M. Gut, A. Perrier, O. Reuille, M. Louis-Simonet, A.F. Junod, and N.V. Vu. 2004. Degree of concurrency among experts in data collection and diagnostic hypothesis generation during clinical encounters. *Medical Education* 38, no. 1: 25–31.

Neufeld, V.R., G.R. Norman, J.W. Feightner, and H.S. Barrows. 1981. Clinical problem-solving by medical students: A longitudinal and cross-sectional analysis. *Medical Education* 15, no. 5: 315–22.

Norman, G. 2005. Research in clinical reasoning: Past history and current trends. *Medical Education* 39, no. 4: 418–27.

Norman, G., M. Young, and L. Brooks. 2007. Non-analytical models of clinical reasoning: The role of experience. *Medical Education* 41, no. 12: 1140–5.

Norman, G.R., and H.G. Schmidt. 1992. The psychological basis of problem-based learning: A review of the evidence. *Academic Medicine* 67, no. 9: 557–65.

Patel, V.L., J.F. Arocha, and J. Zhang. 2005. Thinking and reasoning in medicine. In *The Cambridge handbook of thinking and reasoning*, ed. K.J. Holyoak and R.G. Morrison, 727–51. New York: Cambridge University Press.

Rikers, R.M.J.P., H.G. Schmidt, and V. Moulaert. 2005. Biomedical knowledge: Encapsulated or two worlds apart? *Applied Cognitive Psychology* 19, no. 2: 223–31.

Schmidt, H.G., and H.P.A. Boshuizen. 1993. On the origin of intermediate effects in clinical case recall. *Memory and Cognition* 21, no. 3: 338–51.

Sibert, L., B. Charlin, J. Corcos, R. Gagnon, P. Grise, and C. van der Vleuten. 2002. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Medical Teacher* 24: 522–7.

Sibert, L., S.J. Darmoni, B. Dahamna, M.F. Hellot, J. Weber, and B. Charlin. 2006. On line clinical reasoning assessment with script concordance test in urology: Results of a French pilot study. *BMC Medical Education* 6, no. 45: 1–9.

Swanson, D.B., G.R. Norman, and R.L. Linn. 1995. Performance-based assessment: Lessons from the health professions. *Educational Researcher* 24, no. 5: 5–11.

van der Vleuten, C.P.M. 1996. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1, no. 1: 41–67.