Using subjective judgement to determine the validity of a tutorial performance evaluation instrument

Authors:
Judith C. Bruce[1]
Melanie L. Lack[1]

Affiliations:
[1]Department of Nursing Education, University of the Witwatersrand, South Africa

Correspondence to:
Judith Bruce

e-mail:
judith.bruce@wits.ac.za

Postal address:
Department of Nursing Education, Faculty of Health Sciences, 7 York Road, Parktown, Johannesburg, 2193, South Africa

Keywords:
problem-based learning; quantitative analysis; subjective judgement; tutorial performance; validity; nursing

Dates:
Received: 15 Oct. 2007
Accepted: 21 Nov. 2008
Published: 25 May 2009

How to cite this article:
Bruce, J.C. & Lack, M.L., 2009, 'Using subjective judgement to determine the validity of a tutorial performance evaluation instrument', *Health SA Gesondheid 14*(1), Art. #409, 6 pages. DOI: 10.4102/hsag.v14i1.409

This article is available at:
http://www.hsag.co.za

ABSTRACT

Evaluating students' learning performance is dependent on assessment criteria from which valid inferences can be made about student learning. An existing 36-item instrument used to evaluate baccalaureate nursing students' performance in problem-based learning tutorials was presented to experts in nursing for their subjective judgement of item validity. Quantitative analysis of data sets from experts' judgements was used to construct a valid measurement scale for evaluating students' tutorial performance. The objectives of the study were to determine the content validity of items in a tutorial performance evaluation (TPE) instrument and to determine the construct validity of items through paired comparison of main and sub-items in the instrument. Academic experts (n = 8) from two South African universities were selected by means of purposive, maximum variation sampling. Data were collected in three rounds of the Delphi technique, which incorporated the Subjective Judgement Model for paired comparison of instrument items. Experts' ratings were captured on a visual analogue scale for each item. Relative item weights were determined using paired comparisons. Statistical analysis resulted in ratio scale data, each item being assigned a ratio relative to its weight. It was concluded that quantitative analysis of subjective judgements is useful to determine the construct validity of items through paired comparison of items in a TPE instrument. This article presents the methodological perspectives of subjective judgement to establish instrument validity.

OPSOMMING

Die evaluering van studente se leervermoë is afhanklik van die waardebepalingskriteria waarvan geldige afleidings betreffende die student se leerervaring gemaak kan word. 'n Bestaande instrument met 36 items waarmee baccalaureus-verpleegkundestudente se prestasie in die probleemgebaseerde leertutoriale geëvalueer is, is aan kundiges in verpleegkunde gegee vir subjektiewe beoordeling van die geldigheid van die items. 'n Geldige meetinstrument vir die evaluering van studente se tutoriale prestasie is ontwerp deur van die kwantitatiewe ontleding van die datastelle op grond van die kundiges se oordeel gebruik te maak. Die doelwitte van die studie was om die inhoudsgeldigheid van items in 'n evalueringsinstrument van tutoriale prestasie te bepaal en om die konstrukgeldigheid van items te bepaal deur die gepaarde vergelyking van hoof- en sub-items in die instrument. Akademiese kundiges (n = 8) van twee Suid-Afrikaanse universiteite is deur middel van doelgerigte, maksimale variasie-steekproeftrekking geselekteer. Data is deur middel van drie rondtes van die Delphi-tegniek ingesamel, wat die subjektiewe oordeelmodel vir gepaarde vergelyking van die instrumentitems ingesluit het. Die kundiges se beoordeling is op 'n visueleanaloë-skaal vir elke item weergegee. Relatiewe itemgewigte is deur middel van gepaarde vergelyking bepaal. Statistiese ontleding het verhoudingskaaldata tot gevolg gehad, en elke item is van 'n verhouding relatief tot die gewig voorsien. Daar is bevind dat kwantitatiewe ontleding van subjektiewe beoordeling bruikbaar is om die geldigheid van 'n konstruk deur gepaarde vergelyking van items in 'n evalueringsinstrument van tutoriale prestasie te bepaal. Hierdie artikel bied die metodologiese perspektiewe van die subjektiewe beoordeling aan om die geldigheid van die instrument te bepaal.

INTRODUCTION

Subjective judgement by experts has become increasingly important in the sphere of nursing education, particularly in the development of assessment tools and in educational evaluation processes. Alongside this importance is mounting criticism that subjective judgements are biased, less transparent and less accurate in their predictions. Assessing and evaluating student performance require predictions or results that are accurate, valid and free from bias. It follows that tools used to assess performance in any learning environment must produce results from which valid and unbiased inferences can be drawn. A particular challenge that confronts nursing educators is the assessment of learning outcomes that are abstract and less tangible, such as those outcomes requiring the development of interpersonal skills, leadership, reasoning, cross-cultural competence and group skills.

In a problem-based learning (PBL) context, group skills are particularly important. The literature generally agrees that successful learning outcomes in PBL are dependent on effective tutorial group functioning or tutorial performance (Niemenin, Saure & Lonka 2006:65; Rideout 1999:232; Savin-Baden 2000:2). This is premised on the notion that students possess certain attributes and levels of skills necessary for effective functioning within PBL groups or that, in the least, these attributes and skills will develop over time. There is also consensus that such attributes and skills can only be evaluated within the context of their learning groups – in this case PBL tutorial groups. Students, their peers and the PBL tutor or facilitator all participate in tutorial performance evaluation (TPE) and in this way contribute to formative and summative evaluations (Dornan *et al*. 2004:673; Rideout 1999:232). However, how tutorial performance should be evaluated or what criteria it should be measured against has been the subject of much discussion. Since the inception of PBL in health professional education, few directions have been provided about the validity and reliability of criteria. Criteria are usually specified with little evidence in support of reliability and validity. Criteria are considered to be reliable and valid when they are relevant, and if they have been useful in drawing inferences from students' scores to identify gaps in their

Journal of Interdisciplinary Health Sciences
Article #409

learning. Content validity is usually established through review by a panel of experts, sometimes followed by determination of the content validity index (Lynn 1986:383; Smith, Thurkettle & De la Cruz 2004:616). Construct validity is usually evaluated using factor analysis; the criteria for extracting factors include the use of Eigen values of 1:00 or above (Burns & Grove 2005:533).

The nursing school in this study had designed and used an instrument to assess students' PBL tutorial performance without any evidence at its disposal to support or refute the validity of items in the instrument. Hence, the validity of its assessment of student performance during PBL tutorials was questionable. Central to determining instrument validity and hence valid inferences, to a greater or lesser extent, is the reliance on the subjective judgement by experts. This article aims to discuss the method of subjective judgement and to describe how meaningful subjective judgements can be used to construct a valid measurement scale for evaluating students' tutorial performance in a PBL context. The article describes the research methods in relation to sampling, data-collection techniques and procedure, and the plan for data analysis. Discussion of methodological aspects of subjective judgement will follow in the place of conventional results of research.

Aim and objectives: The aim was to determine the validity of an existing 36-item TPE instrument based on the subjective judgements by experts. This instrument had been developed and used by the academic department participating in the study. The research objectives were to:

- determine the content validity of items in the TPE instrument; and
- determine the construct validity of items through paired comparison of main items (Mi) and sub-items (si) in the instrument.

The discussion focuses primarily on subjective judgement as method to collect data in relation to these objectives.

Definition of key concepts

Subjective judgement is generally understood as a process whereby informed persons, called experts, give an opinion or estimate of something based on intuition and guessing (Miranda 2001:88) in the absence of objective data. Problem-based learning is a teaching-learning approach in which students address health problems or issues in small groups, guided by a facilitator. Tutorial performance refers to student behaviours in a small group-learning context that facilitate individual learning, group learning and team work (Rideout 1999:233). Evaluation refers to the process of collecting and interpreting information to assist with judgements about students' learning and performance (Oermann & Gaberson 2006:2). Validity refers to the appropriateness and usefulness of inferences made from assessments and evaluations (Oermann & Gaberson 2006:24).

LITERATURE REVIEW

Emerging ideas on the concepts validity and reliability suggest that validity is not entirely a property of the measuring instrument, but of the instrument's scores and their interpretation by educators and other users. Measuring instruments derived from subjective judgement by experts have been used in research in a range of fields: medical education (Downing, Teikan & Yudkowsky 2006:51), the military (Crawford & Williams 1985b:387), ecology (McCarthy *et al.* 2004:76), psychology and social science research (Miranda 2001:87). In all of these fields problems have surfaced in acquiring and treating judgemental data (Crawford & Williams 1985b:388). Subjective judgement as an alternative to mathematical models has been criticised for being less accurate than other models, for its bias towards over-estimating the value of an item and for its variability according to relative benefit the experts might derive from the outcome (McCarthy *et al.* 2004:77). As a result of these criticisms, researchers have made considerable strides towards the quantitative analysis of subjective data to minimise inherent biases and to improve its accuracy.

Quantitative analysis of subjective data is important when other methods for decision making and problem solving would be inappropriate or difficult, for example in cases of allocating public monies (Crawford & Williams 1985b:400) or, as in this case, measuring the academic performance of students. In these examples the problems or issues to be dealt with are often amorphous or of multifaceted interest (Crawford & Williams 1985b:400) to individuals who have diverse backgrounds, experiences and motives. In education the outcome of judgements hold different meanings for educators, students, academic planners and managers. Furthermore, contemporary educational contexts are indeed diverse yet polarised according to language, culture and ethnicity, creating greater opportunity for preference and bias when subjective judgements are used to measure academic performance. In small group-learning situations the dynamics of diversity are far more tangible and exert greater influence on learning – positive and negative – than in large groups using traditional methods. Although this proposition is sufficiently important to justify quantitative analysis of subjective data, Crawford and Williams (1985a:2) posit additional reasons for the desirability of quantitative analysis: 1) It provides a formal analytic framework that lends structure and definition to data sets that are usually amorphous and unstructured. The researcher then has the opportunity to consider the data systematically and to examine options or comparisons one at a time. 2) This formal framework for quantitative analysis allows for repeatability and hence an audit trail. In this regard Crawford and Williams (1985a:3) cite the extensive allocation of public resources as an example where an audit trail would be mandatory. 3) A quantitative framework also enhances sensitive analysis of data where the researcher is faced with different viewpoints. In so doing the researcher has an opportunity to study the effects of variations in subjective judgements on the research outcomes.

In order to subjectively judge the ranked importance of items in the TPE instrument, a paired comparison analysis was employed in this study. The method of paired comparisons within a subjective judgementmodel is not new and was described over four decades ago by David (1963:9) as being useful primarily when objects or items to be compared can be judged only subjectively. Paired comparison was first introduced in its embryonic form by Fechner in 1860 and, after considerable extensions, made popular by Thurstone in 1927 (David 1963:10). In this method items are presented in pairs to experts who are requested to judge the value or importance of an item relative to the other. Hence, the idea behind paired comparisons in determining construct validity is to estimate the value of a construct in relation to others in a set of constructs or sub-constructs, as in the case of a performance-assessment instrument. As stated in the Mindtools E-books (2006:3), paired comparison analysis helps one to work out the importance of a number of options relative to each other. It is particularly useful in the absence of objective data on which to base decisions. In studies where paired comparisons were used as opposed to a single comparison to some vague notion held by an expert, the accuracy of the judgement was improved (Miranda 2001:89). These findings are consistent with the results of Lederer and Prasad's (1992:55) study, which showed that paired comparisons produce better estimates than those produced on the basis of intuition or without any comparative measure, which may be viewed as guessing. Paired comparison analysis was carried out in this study using the linear model of the visual analogue scale, which is based on the notion that the expert's or the judge's predictions are a linear combination of available cues, either presented to or chosen by the judge. Cues are employed to make more explicit the corresponding weighting structure used in the judge's weighting policy. Paired comparisons are therefore important to improve the accuracy of subjective judgements.

Article #409
Journal of Interdisciplinary Health Sciences

TABLE 1
The original Tutorial Performance Evaluation Instrument

ITEMS	0	1	2	3	4	5	6	7
A. GROUP GROWTH								
1. Offers facts, suggestions, opinions								
2. Willing to work with group members								
3. Offers encouragement and support to group members								
4. Takes risks in expressing ideas								
5. Acknowledges contributions from group members								
6. Willing to share resources								
B. LEADERSHIP								
1. Gives direction								
2. Suggests opinions/decisions								
3. Volunteers to undertake tasks								
4. Identifies learning issues								
5. Identifies resources								
C. LEARNING AND TEACHING SKILLS								
1. Demonstrates use of multiple resources								
2. Demonstrates ability to integrate resources								
3. Contextualises learning								
4. Demonstrates ability to assist others to learn								
D. CONTENT								
1. Is accurate								
2. Is up-to-date								
3. Is sequential								
4. Is comprehensive/interdisciplinary								
5. Integrates legislation, ethics, social and physical sciences								
6. Evaluates and selects								
E. PROBLEM SOLVING SKILLS								
1. Defines/delineates problem								
2. Selects appropriate framework/strategies to solve problem								
3. Selects/designs appropriate strategy to solve problem								
4. Implements solution/option								
5. Evaluates problem solving process								
F. INTERACTION/COMMUNICATION								
1. Identifies own strengths and weaknesses								
2. Assumes different roles in group								
3. Demonstrates verbal skills appropriate to the situation								
4. Demonstrates non-verbal skills appropriate to the situation								
5. Demonstrates attitude appropriate to the situation								
6. Demonstrates integrity of (own) values/morals								
G. CRITICAL THINKING								
1. Identifies and challenges assumptions								
2. Demonstrates contextual awareness and thinking								
3. Explores and imagines alternatives								
4. Demonstrates analysis (active inquiry) and action								

TABLE 2
Likert scale descriptors (Adapted from Lynn 1986:384)

SCORE	DESCRIPTOR
1	Not relevant
2	Unable to assess relevance without item revision; in need of extensive revision that it would be no longer relevant
3	Relevant, but needs minor alteration
4	Very relevant

Journal of Interdisciplinary Health Sciences Article #409

Current TPE instruments have been criticised for producing scores that are not meaningful for valid inferences about students' performance within learning groups. This is evidenced by mismatched scores between and negative reports from facilitators and students. Additionally, items in TPE instruments are assumed to be of equal value or weight and are therefore unable to predict trends in the development of certain skills over a period of time. Current evidence suggests that students in the beginning years of study remain stagnant (high or low) in certain skills such as leadership and communication, which generally should show positive growth as learning opportunities in PBL groups increase.

THE TUTORIAL PERFORMANCE-EVALUATION INSTRUMENT

Students' performance in PBL tutorial groups was originally assessed using an evaluation instrument comprising of seven main items (constructs) and 36 sub-items. These items were all equally weighted and rated against an eight-point Likert scale (0 = never; 1–2 = seldom; 3–4 = sometimes; 5–6 = often; 7 = always); the points on the Likert scale were equidistant. Instrument items were generated through an extensive literature review and consultations with faculty and final-year baccalaureate students. The instrument was piloted using a sample of 22 baccalaureate students in the PBL programme. Results from the pilot study indicated minor changes to be made in the usage of a few words. This instrument is referred to as the 'original' TPE instrument (Table 1).

METHODS

This study used quantitative, descriptive methods to determine the content and construct validity of items in the TPE instrument. Content validity was determined using experts' judgement of instrument items according to a four-point rating scale; construct validity was determined through paired comparison analysis methods within a subjective judgement model.

The sample

Using purposive, maximum variation sampling, an expert group (n = 8) was selected from a target population of academics involved in the development and/or implementation of PBL in Health Science faculties at two South African universities. Disciplines represented in the sample included Nursing Education, Occupational Therapy, Dentistry and Medical Education (the Graduate Entry Medical Programme). Seminal writings on instrument validation by Lynn (1986:384) suggest that five or more experts are necessary to minimise or neutralise the probability of agreement without question. Although this may not be a critical factor in a decision Delphi as in this case, the researchers used maximum variation sampling as a way to obtain a sample that can provide rich data through their decisions and not through consensus. Ten (n = 10) experts were initially identified; however, only eight (n = 8) gave their consent and followed through with their participation.

Data collection

Data collection was approached using the Delphi technique, which incorporated the Subjective Judgement Model. The Delphi technique was used to elicit the judgement of experts for the purpose of making decisions about the items contained within the original TPE instrument. Consensus between experts was therefore not the intended outcome of the Delphi technique.

The Subjective Judgement Model was employed for the subjective, paired comparison of instrument items using a visual analogue scale. The Subjective Judgement Model, also referred to as the Human Judgement Model, enables subjective comparison of two (i.e. paired) objects or items using either a judgement matrix model (Miranda 2001:88) or a linear model, incorporating a visual analogue scale to record the experts' judgements. Paired comparison analysis helps to determine the importance of a number of items relative to each other. In this study, the linear model was used by the experts to subjectively judge the relative importance of main items (marked Mi) and sub-items (marked si) on the TPE instrument. The magnitude of the visual analogue scale is 100 mm in length with right-angle anchors at each end of a vertical or horizontal line. In this study, horizontal visual analogue scales, measuring 0 to 100 mm, were used to subjectively rate paired comparison between items in the instrument.

Three rounds of the Delphi technique elicited data to determine content and construct validity of the TPE instrument. In each instance experts were provided with clear instructions.

Round one Delphi

Determination of content validity commenced in this round. The main items (Mi) and sub-items (si) in the original TPE instrument were converted into a question format to be sent to the experts for their rating. Now referred to as the questionnaire, experts were asked to rate each main item (n = 7) and its sub-items (n = 36) on the questionnaire according to the descriptors of a four-point rating scale (Table 2). A 'comments box' following each set of main and sub-items enabled experts to give an opinion on how items should be revised and on items to be added. An open-ended question was included to elicit experts' opinion on the rating scale used for items in the questionnaire. Upon return of the questionnaires, main items and sub-items scoring 1 or 2 were excluded from the instrument. Items that received a rating of 3 were revised and modified based on the experts' opinion. Items that obtained a rating of 4 were retained in the instrument. The questionnaire was refined and used in round two Delphi.

Round two Delphi

Determination of content validity continued in this round. The refined questionnaire (from round one Delphi) was returned to the same expert group. A time interval of 14 days had elapsed between rounds one and two Delphi. Items that were modified or added as new items were colour-coded for easy identification and rescoring by the experts. Few modifications were required in this round; those sub-items scoring 1 or 2 were removed. Items that received a rating of 4 remained and the few items (n = 4) that received a rating of 3 in this round required merely a grammatical change. At the end of this round the revised questionnaire comprised all items, which scored 4 (very relevant); it became the (new) instrument for TPE, comprising seven main items and 34 sub-items. As per experts' proposal, these items would be rated against a four-point (0–3) Likert scale with descriptors. Based on the experts' ratings, the instrument was deemed to possess content validity and inferences from the results yielded could be considered valid.

Round three Delphi

A 14-day period elapsed following round two. In round three the Subjective Judgement Model, described above, was utilised for the determination of construct validity through paired comparisons of all main and sub-items. This was done using visual analogue scales. A total of 100 visual analogue scales were developed: 21 for main items and 79 for sub-items. The same experts were asked to exercise their subjective judgementof the relative importance of each item versus another item in a pair-wise manner, by placing their mark (/) between 0 to 100 mm on the visual analogue scale. Experts were further required to conduct a similar weighting assessment of the four-point Likert scale proposed in round two Delphi. The completed scales were mathematically referred to as 'units' for the purpose of analysing ratio-level data produced (Crawford & Williams 1985a:5). There were 808 units for analysis.

Data-analysis plan

Each visual analogue scale was measured accurately in millimetres from 0 to the expert's mark. The measurements of each unit were entered onto an Excel spreadsheet under codes ranging from 01 to 08, representing each of the eight experts.

Article #409 Journal of Interdisciplinary Health Sciences

Ratio-level data (ratio scales) were derived from the experts' judgements, using a relatively complex mathematical procedure performed by a resident statistician. The absolute size or value of items was then calculated using the ratio scale and a reference value. The results after the analysis of data are presented in a follow-up article.

ETHICAL CONSIDERATIONS

The study was approved by and ethical clearance was obtained from the respective university committees. Permission was obtained from the Head of the Department of Nursing to conduct the study using the existing evaluation instrument. Experts were recruited on the basis of informed consent and after a meeting with the researcher to clarify aspects of the study. They were assured that participation was voluntary, that their identity would be protected and that they could withdraw without any reprisal. All raw data and documents were kept in a locked cupboard to prevent unauthorised access to information.

DISCUSSION

Methodological perspectives of subjective judgements are discussed with reference to: 1) eliciting meaningful judgements, 2) synthesising subjective data, and 3) constructing the measurement scale.

Subjective judgement by an individual expert or a panel of experts is somewhat paradoxical within the context of quantitative approaches in research. It brings into question the objectivity and rigor of research and the meaningfulness of data from experts' judgements. The researchers considered two aspects important to elicit meaningful judgements from experts. Firstly, it was important to be clear about the sampling method to obtain a participant sample, and secondly, it was important to be clear about the type and quality of data to be elicited from the experts' judgements. To this end the researchers asked two questions: 'What purpose should the data serve?' and 'What criteria should experts meet to provide purposeful data?' The latter question required specificity about who should be considered as 'experts'. Although purposive sampling is broadly defined as '[s]electing participants based on personal judgement about which ones will be more representative or informative' (Polit & Beck 2004:729), it continues to be criticised for the lack of methods in support of the representativeness or typicalness of the sample selected. Maximum variation (heterogeneity) sampling (Patton 2002:234) was used to identify individuals who could provide rich information about the validity of items to evaluate PBL tutorial performance. This sampling method aims to identify themes or patterns that run through a range of variations (Patton 2002:234); it is informed by the logic that any common themes that emerge from great variation are valuable in capturing the core experiences of a setting or phenomenon (Patton 2002:234). In this study, three PBL curriculum categories were identified, namely PBL curriculum planning, implementation and evaluation, and were used as an overarching sampling frame. Within this sampling frame, 10 participants were purposively selected from universities offering a PBL Health Sciences curriculum.

In relation to eliciting meaningful judgements, the visual analogue scales were crucial to the collection of purposeful data. Usually data generated from the use of visual analogue scales are based on experts' educated guesses. To clarify the use and purpose of the visual analogue scales, and hence the purposefulness of data it elicited, a meeting was held with each expert explaining how the visual analogue scales should be used. The experts were asked to be seated when they worked on the visual analogue scales since alteration in perception of the length of the line is less likely to occur (Burns & Grove 2005:436). To further enhance the purposefulness of data from the visual analogue scales the researchers elected to include a rating scale to guide the judgement of experts in deciding the relative weighting of main items and sub-items. In most visual analogue scales, however, only the endpoints are specified or defined to guide those using the scale to exercise their judgement. Synthesising subjective data is dependent on the quality of the data-analysis plan, which includes data cleaning and organisation. Since the use of the Subjective Judgement Model was novel in this study, it was important to be clear about those factors that may influence the quality of data analysis. In the absence of modernity, photocopying was relied upon to produce multiple copies of the instruments. Wewers and Lowe (1990:230) caution against artefacts that may distort, in particular, the length of the visual analogue scale. Each line was co-checked for exactness of its length (100 m) prior to being dispatched to experts. Data cleaning was done through manual co-checking in the absence of appropriate software.

The endpoint of validating the items of the TPE instrument is a measurement scale, which will enable valid and meaningful inferences about students' tutorial performance. Constructing measurement scales from subjective data is informed by several statistical methods – mostly under the name of 'paired comparisons' (Crawford & Williams 1985b:389). The paired comparison approach used in this study merely elicited the judges' relative preference between items without any indication of the strength or weighting of their preference. Although this part of the research is dependent on the application of a statistical model for analysis, and as such is incomplete, it is the intention of the researchers to develop a software program that will enable computerised calculation of the weight of scores assigned to students' tutorial performance. Similar work done by software engineering research groups, who have designed software programs used by project managers to improve the subjective estimates of costing, will be a useful precept to follow.

Recommendations

In terms of research it is recommended that:

- subjective judgement by experts be quantified for content and construct validity of items in an evaluation instrument;
- paired comparison be used to assign different weights to items, doing away with predominantly equally weighted items in evaluation instruments; and
- professional statistical support be enlisted at all times during quantification and weighting procedures.

The fact that the Subjective Judgement Model was applied to one instrument only and in a specific context may be a limitation to extending its use for determining the validity of other instruments.

CONCLUSION

Although subjective judgement by experts has been criticised for being less accurate and less valid than mathematical models, it continues to be useful for decisions and consensus in developing and testing instruments to evaluate student learning. To counter claims about its ability to provide valid inferences about student learning and to enhance its utility, quantitative analysis becomes a necessary endeavour. Subjective rating of the value and importance of items within an instrument and the relative importance between main and sub-items gives the relative weight of each item; in so doing it contributes to valid instrument scores for meaningful interpretations about student learning. It may therefore be concluded that through the quantitative analysis of experts' judgements during three rounds of Delphi and using the Subjective Judgement Model, the TPE instrument possesses both content and construct validity.

REFERENCES

Burns, N. & Grove, S.K., 2005, *The practice of nursing research conduct: Critique and utilization*, WB Saunders, New York.

Crawford, G. & Williams, C., 1985a, *The analysis of subjective judgement matrices. A project AIR FORCE report*, Rand, Santa Monica.

Crawford, G. & Williams, C., 1985b, 'A note on the analysis of subjective judgement matrices', *Journal of Mathematical Psychology* 29, 387–405.
David, H.A., 1963, *The method of paired comparison*, Charles Griffin, London.
Dornan, T., Boshuizen, H., Cordingley, L., Hider, S., Hadfield, J. & Scherpbier, A., 2004, 'Evaluation of self-directed clinical education: Validation of an instrument', *Medical Education* 38, 670–678.
Downing, S.M., Tekian, A. & Yudkowsky, R., 2006, 'Procedures for establishing defensible absolute passing scores on performance examinations in health professions education', *Teaching and Learning in Medicine* 18(1), 50–57.
Lederer, A. & Prasad, J., 1992, 'Nine management guidelines for better cost estimating', *Communication ACM* 35(2), 51–59.
Lynn, M.R., 1986, 'Determination and quantification of content validity', *Nursing Research* 35(6), 382-385.
McCarthy, M.A., Keith, D., Tietjen, J.,Burgman, M.A., Maunder, M., Master, L., Brook, B.W., Mace, G., Possingham, H.P., Medellin, R., Andelman, S., Regan, H., Regan, T. & Ruckelshaus, M., 2004, 'Comparing predictions of extinction risks using models and subjective judgement', *International Journal of Ecology* 26, 76–74.
Mindtools E-books, 2006, *Paired comparison analysis. Working out the relative importance of different options*, Mindtools, viewed 17 June 2006, from http://www.mindtools.com.
Miranda, E., 2001, *Improving subjective estimates using paired comparisons*, IEEE Software, Ericsson Research Canada, Mississauga, Ontario.
Niemenin, J., Saure, P. & Lonka, K., 2006, 'On the relationship between group functioning and study success in problem-based learning', *Medical Education* 40, 64–71.
Oermann, M.H. & Gaberson, K.B., 2006, *Evaluation and testing in nursing education*, Springer Publishing Company, New York.
Patton, M.Q., 2002, *Qualitative research and evaluation methods*, Sage, Thousand Oaks.
Polit, D.F. & Beck, C.T., 2004, *Nursing research principles and methods*, Lippincott Williams and Wilkins, Philadelphia.
Rideout, E., 1999, *Transforming nursing education through problem based learning*, Jones and Bartlett Publishers International, London.
Savin-Baden, M., 2000, *Problem-based learning in higher education: Untold stories*, viewed 7 July 2006, from: www.mcgraw-hill.co.uk/openup/chapter.
Smith, A.J., Thurkettle, M. & De la Cruz, F.A., 2004, 'Use of intuition by nursing students: Instrument development and testing', *Methodological Issues in Nursing Research* 47(6), 614–622.
Wewers, M.E. & Lowe, N.K., 1990, 'A critical review of visual analogue scales in the measurement of clinical phenomena', *Research in Nursing and Health* 13(4), 227–236.

Article #409
Journal of Interdisciplinary Health Sciences