Design, format, validity and reliability of multiple choice questions for use in nursing research and education

Julie Considine, Mari Botti, Deakin University
Shane Thomas, La Trobe University

Multiple choice questions are used extensively in nursing research and education and play a fundamental role in the design of research studies or educational programs. Despite their widespread use, there is a lack of evidence-based guidelines relating to design and use of multiple choice questions. Little is written about their format, structure, validity and reliability of in the context of nursing research and/or education and most of the current literature in this area is based on opinion or consensus. Systematic multiple choice question design and use of valid and reliable multiple choice questions are vital if the results of research or educational testing are to be considered valid. Content and face validity should be established by expert panel review and construct validity should be established using 'key check', item discrimination and item difficulty analyses. Reliability measures include internal consistency and equivalence. Internal consistency should be established by determination of internal consistency using reliability coefficients while equivalence should be established using alternate form correlation. This paper reviews literature related to the use of multiple choice questions, current design recommendations and processes to establish reliability and validity, and discusses implications for their use in nursing research and education.

Key words: multiple choice questions, parametric tests, nursing education, nursing research, measurementg

Julie Considine RN RM BN CertAcuteCareNsg(Emerg) GradDipNsg(Acute Care) MN, PhD Candidate, School of Nursing, Faculty of Health and Behavioural Sciences, Deakin University
Email: Julie.Considine@nh.org.au

Mari Botti RN BA GDCAP PhD, Professor, School of Nursing, Faculty of Health and Behavioural Sciences, Deakin University

Shane Thomas BA(Hons) DipPubPol PhD, Professor, School of Public Health, Faculty of Health Sciences, La Trobe University

Introduction

Multiple choice questions (MCQs) are often used to measure knowledge as an end-point in nursing research and education, usually in the context of testing an educational intervention. It is of paramount importance that MCQs used in nursing are valid and reliable if nursing, as a profession, is to produce credible results that may be used to change nursing practice or methods of nursing education. The majority of the literature regarding the format, structure, validity and reliability of MCQs is found in medical education, psychometric testing and psychology literature and little is written regarding the use of MCQs for nursing research and, or, education. There is also a lack of empirically supported guidelines for development and validation of MCQs (Violato 1991, Masters, Hulsmeyer et al 2001). While there are several papers pertaining to the design and use of MCQs, many of the publications to date are based on opinion or consensus.

The purpose of the discussion in this paper is to provide a review of findings from the literature about the use of MCQs and current recommendations for their design and format and examine the processes that should be undertaken to establish reliability and validity of MCQs for use in nursing research and education.

Current guidelines for the format and development of MCQs

Over the last decade there has been an increase in research into the format, design and construction of MCQs (Haladyna 1999). To date the conventional format of MCQs has three components: the stem, the correct answer or the key, and several incorrect but plausible answers or distractors (Isaacs 1994, Nunnally & Bernstein 1994, Haladyna 1999). While it is acknowledged that format characteristics of MCQs such as number of options for each test item, number of correct responses, use of inclusive alternatives, completeness of the stem and orientation of the stem, influence the difficulty and discrimination of MCQs, there is little empirical evidence on which to make definite recommendations about question design and format (Violato 1991).

The stem

The stem provides the stimulus for the response and should provide the problem to be solved (Gronlund 1968, Isaacs 1994,

Haladyna 1999, Linn & Gronlund 2000). The stem may be written as a question or partial sentence that requires completion. Research comparing these two formats has not demonstrated any significant difference in test performance (Violato 1991, Haladyna 1999, Masters et al 2001). To facilitate understanding of the question to be answered, it is recommended that if a partial sentence is to be used, a stem with parts missing either at the beginning or in the middle of the stem should be avoided (Haladyna 1999).

The stem should present the problem or question to be answered in a clear and concise manner (Gronlund 1968, Nunnally & Bernstein 1994, Haladyna 1999, Linn & Gronlund 2000, Masters et al 2001). If the stem requires the use of words such as 'not' or 'except', these words should be made very obvious by the use of bold, capital or underlined text (Gronlund 1968, Haladyna 1994, Nunnally & Bernstein 1994, Masters et al 2001).

The key

The correct answer is referred to as the 'key' (Isaacs 1994). There should be only one correct answer for each MCQ (Gronlund 1968, Haladyna 1994, Isaacs 1994, Linn & Gronlund 2000). The location of the key should be evenly distributed throughout the test to avoid 'placement bias' (Gronlund 1968, Haladyna 1999, Masters et al 2001). This means that if, for example, there were twenty questions, each with four options, each option would be the correct answer on five occasions. When allocating the position of the correct answer for each MCQ, care should be taken not to use patterns that may be recognised by participants (Gronlund 1968).

Distractors

Distractors are incorrect answers that may be plausible to those who have not mastered the knowledge that the MCQ is designed to measure yet are clearly incorrect to those who possess the knowledge required for that MCQ (Haladyna 1999). A good distractor is one that is selected by those who perform poorly and ignored by those who perform well (Gronlund 1968, Haladyna 1999, Linn & Gronlund 2000). Both the correct answer and the distractors should be similar in terms of grammatical form, style and length (Gronlund 1968, Nunnally & Bernstein 1994, Haladyna 1999, Linn & Gronlund 2000). The use of obviously improbable or implausible distractors should be avoided (Haladyna 1994, Isaacs 1994, Nunnally & Bernstein 1994, Linn & Gronlund 2000) as obviously incorrect distractors increase the likelihood of guessing the correct answer (Haladyna 1994).

There is ongoing debate in the literature regarding the optimum number of distractors for a multiple choice test item. Many authors state that the performance of distractors is more important than the number of distractors (Haladyna 1994, Masters et al 2001). While there is some evidence that there are advantages to having greater numbers of options per test item, this is only true if each distractor is performing appropriately (Haladyna 1999). Several studies have examined the use of fewer options per test item. Haladyna and Downing (1993) (cited in Haladyna 1999 p48) found that the majority of MCQs had only one or two 'working' distractors and concluded that three option MCQs consisting of one correct answer and two distractors were suitable. The reliability of three option MCQs has been shown to be comparable to that of four option MCQs (Catts 1978, Nunnally & Bernstein 1994). The use of three option MCQs are advantageous as they take less time to complete and less time to construct (Catts 1978, Masters et al 2001) and reduce the probability of inclusion of weak distractors (Masters et al 2001).

Decreasing the probability of guessing the correct answer is often cited as a reason to increase the number of alternatives in a MCQ. The probability that a participant will guess the correct answer is equal to one divided by the number of alternatives (Nunnally & Bernstein 1994 p340). For example, the probability of guessing is 0.50 for two alternatives, 0.33 for three alternatives, 0.25 for four alternatives, 0.20 for five alternatives, 0.16 for six alternatives and so on. It therefore may be argued that the decrease in likelihood of guessing the correct answer becomes minor when the number of alternatives is increased beyond four or five (Nunnally & Bernstein 1994).

Other considerations

More general issues to be considered in the development of MCQs are simplicity, formatting and order of options, number of principles tested and independence of MCQs when appearing as a series (Gronlund 1968, Haladyna 1994, Isaacs 1994, Haladyna 1999). Given that the purpose of MCQs is to test knowledge, rather than ability to read or translate what is written, the vocabulary used in MCQs should be simple enough to be understood by the weakest readers in the group and the amount of reading required should be minimised where possible (Haladyna 1994, Haladyna 1999). Current literature recommends that MCQ options be formatted vertically rather than horizontally to facilitate ease of reading and that options should be presented in a logical order. For example, a numerical answer should be presented in ascending or descending numerical order (Haladyna 1994, Isaacs 1994, Haladyna 1999). Each MCQ should be designed to test one specific element of content or one type of mental behaviour (Gronlund 1968, Haladyna 1994, Haladyna 1999). When designing a series of MCQs, each should be independent from one another to avoid one question providing a cue for another question (Gronlund 1968, Haladyna 1994) as this is likely to introduce unfavourable psychometric properties into the question series.

Reliability

Reliability is the degree to which an instrument produces the same results with repeated administration (Beanland et al 1999, Polit & Hungler 1999, Gravetter & Wallnau 2000). A high level of reliability is particularly important when the effect of an intervention on knowledge is measured using a pre-test / post-test design. In this type of research design, reliability of the pre-test and post-test are fundamental to the credibility of

results and the ability of the researcher to attribute differences in pre-test and post-test performance to the intervention being tested. The ability to attribute such changes is also affected by research design (Polger & Thomas 2000).

Concepts related to reliability are consistency, precision, stability, equivalence and internal consistency (Beanland et al 1999 p328). MCQs can be considered to have a high degree of reliability because they have an objective scoring process (Haladyna 1994, Haladyna 1999). Other forms of measurement of learning outcomes, for example, essay test items may be influenced by subjectivity or variation between scorers and are subject to a number of biases such as hand writing quality and length of sentences, both of which have been shown to affect essay test scores (Haladyna 1994).

Reliability is measured using correlation coefficients or reliability coefficients (Beanland et al 1999, Polit & Hungler 1999, Gravetter & Wallnau 2000). For a set of MCQs to be considered reliable, the values of these coefficients should be positive and strong (usually greater than + 0.70) (Gravetter & Wallnau 2000). The square of the correlation (r*2*) measures how accurately the correlation can be used for prediction by determining how much variability in the data is explained by the relationship between the two variables (Gravetter & Wallnau 2000 p539).

The data used for reliability and validity analyses are typically obtained during a pilot study. The sample for the pilot study should consist of participants who ultimately will not form part of the true research sample. The administration and use of the MCQs in the pilot study should be conducted under conditions that are similar to the intended use of the MCQs (Nunnally & Bernstein 1994). For example, the pilot sample should be representative of the eventual target population in terms of range and level of ability. Conditions such as time limits should also be similar (Nunnally & Bernstein 1994).

Stability

Stability of a single set of MCQs is established using test re-test correlation. The MCQs are administered to the same group of participants on two or more occasions and the test scores compared using a correlation coefficient, usually Pearson's *r* product moment correlation (Beanland et al 1999). Currently, the optimal time interval between the test and re-test when using MCQs remains under debate in the literature. A major issue in the interpretation of test re-test correlation is the influence of practice effects and memory on the re-test result (Nunnally & Bernstein 1994) and if the time between test and re-test is short, there is the possibility that these effects will result in artificially improved re-test results (Linn & Gronlund 2000). It should also be remembered that correlation coefficients are insensitive to changes in overall scores from pre-test to post-test. The use of longer intervals between test and re-test may minimise these effects but re-test results may be affected by changes in participants over time (Linn & Gronlund 2000). As currently there is no evidence regarding the ideal interval between testing and re-testing, the researcher needs to consider factors such as effects of time on participants and what the results will be used for in order to make a judgement regarding an appropriate interval between tests.

Equivalence

Issues surrounding the use of test re-test correlation may be minimised by the development of two alternative forms of MCQs. This method is appropriate if the researcher believes that practice effects or memory of the first administration will influence post test performance. Equivalence will determine whether the two sets of MCQs are measuring the same attributes (Polit & Hungler 1999) and is established using alternative form correlation (also known as parallel form or equivalent form correlation) (Nunnally & Bernstein 1994, Beanland et al 1999, Linn & Gronlund 2000). To determine equivalence, the two sets of MCQs are administered to the same participants, usually in immediate succession and in random order (Polit & Hungler 1999). The two test scores are then compared using a correlation coefficient, usually Pearson's *r*.

Internal consistency

Internal consistency is an estimate of 'reliability based on the average correlation among items within a test' (Nunnally & Bernstein 1994 p251) and examines the degree to which the MCQs in a test measure the same characteristics or domains of knowledge (Beanland et al 1999, Polit & Hungler 1999). Typically, internal consistency is measured by the calculation of a reliability coefficient (Cronbach 1990, Beanland et al 1999, Polit & Hungler 1999). For each MCQ, the reliability coefficient examines the proportion of participants selecting the correct answer in relation to the standard deviation of the total test scores (Linn & Gronlund 2000). While there are several statistical formulae that can be used to make these calculations, it should be remembered that while they use different calculations, they fundamentally produce the same result (Cronbach 1990). The most common formula used to measure internal consistency is 'coefficient alpha' (Cronbach 1990) however when describing the internal consistency of MCQs, many research reports refer to the Kuder-Richardson coefficient (KR-20). Kuder-Richardson (KR-20) is a specific form of the coefficient alpha formula and is used for dichotomous data (Beanland et al 1999, Polit & Hungler 1999), for example in the context of MCQs when answers are scored as 'correct' or 'incorrect' (Linn & Gronlund 2000).

Reliability coefficients are an expression of the relationship between observed variability in test scores, true variability and variability due to error (Cronbach 1990, Beanland et al 1999, Polit & Hungler 1999). Reliability coefficients range from zero to one (Catts 1978, Beanland et al 1999). The closer the reliability coefficient is to one, the more reliable the research instrument. A reliability coefficient of 0.70 or greater is generally considered acceptable (Beanland et al 1999, Polit & Hungler 1999). Although a high reliability coefficient is

considered more desirable, a coefficient approaching 1.0 may suggest a high level of redundancy in the test items and indicates that MCQs may be eliminated to shorten the test without adversely affecting reliability.

Reliability coefficients are useful in alerting researchers to errors in sampling of content that will adversely affect reliability (Nunnally & Bernstein 1994). Low reliability coefficients may indicate that the test is too short or that the MCQs have very little in common (Nunnally & Bernstein 1994). Nunnally and Bernstein (1994) advocate that both reliability coefficients and alternative form correlation be reported. In the case of MCQs, if the alternative form correlation is significantly lower than the reliability coefficient (0.20 or greater) measurement error from differences in content or variation over time is indicated (Nunnally & Bernstein 1994).

Validity

Validity of a research instrument is the degree to which the instrument measures what it is supposed to measure (Beanland et al 1999). Validity is closely related to reliability because for an instrument to be valid, it must be reliable (Beanland et al 1999, Polit & Hungler 1999). It is also important to remember that instruments may in fact be reliable even when they are not valid (Beanland et al 1999, Polit & Hungler 1999). MCQs should be subject to a number of reviews to establish validity and identify any sources of bias (Haladyna 1994). Factors that contribute to increased or decreased difficulty of MCQs include: some form of bias (Haladyna 1999), for example, poor instructions, use of complicated vocabulary, ambiguous statements, inadequate time limits; MCQs that are inappropriate for the learning outcomes being measured; poorly constructed MCQs; too few MCQs and identifiable patterns of correct answers (Linn & Gronlund 2000). Validity consists of numerous elements including content validity, face validity and construct validity. The validity of each of these elements needs to be determined to establish overall validity of MCQs.

Content validity

Content validity ascertains whether the MCQs are relevant, appropriate and representative of the construct being examined and / or the cognitive processes that they are intended to test (Beanland et al 1999, Polit & Hungler 1999). There is currently no completely objective method of establishing content validity (Polit & Hungler 1999). Content validity is reliant on judgement (Polit & Hungler 1999). Content validity of MCQs is usually established by a content review, which should be undertaken by experts in the domain being examined and who also have some expertise in tool development (Beanland et al 1999, Haladyna 1999, Polit and Hungler 1999). It is recommended that expert panels comprise at least three persons (Polit & Hungler 1999).

Face validity

Face validity may be considered a sub-type of content validity (Beanland et al 1999). Face validity pertains to the appearance of an instrument and includes such issues as clarity, readability and ease of administration (Beanland et al 1999). Editorial review and pilot study are used to establish face validity of MCQs and should confirm acceptable readability, clarity of content and writing, consistency of style, and identify errors in spelling, grammar, punctuation or abbreviations (Haladyna 1999).

Construct validity

Construct validity is the extent to which an instrument measures a theoretical attribute (Beanland et al 1999, Polit & Hungler 1999). In the case of MCQs, construct validity is related to whether or not the questions measure the domain of knowledge being examined. The construct validity of MCQs should be established using 'key check' and item response analyses such as item difficulty analysis, item discrimination analysis and distractor evaluation (Gronlund 1968, Violato 1991, Haladyna 1999, Masters et al 2001).

Key check

The key check determines if the correct answer to the MCQ is actually correct and ensures that there is not more than one answer that may be considered to be correct. The key check should be conducted by a number of persons who are experts in the content area. Where there is variation in the answer perceived to be correct, the MCQ should be reviewed until there is consensus (Haladyna 1994).

Item discrimination analysis

Item discrimination analysis examines how each MCQ is related to overall test performance (Nunnally & Bernstein 1994, Haladyna 1999). Item discrimination has the underlying premise that if a question is highly discriminating, the overall test scores of those choosing the correct answer to that question should be higher than the overall test scores of those who choose the incorrect answer (Haladyna 1999, Linn & Gronlund 2000, Masters et al 2001).

Nunnally and Bernstein (1994) recommend the use of item to total correlations to examine item discrimination. Item to total correlations are used to statistically establish item discrimination of MCQs by analysing the relationship between each MCQ and the total test score (Nunnally & Bernstein 1994, Beanland et al 1999). The point biserial correlation coefficient (*rpb*) uses the Pearson correlation coefficient (*r*) formula to measure the relationship between two variables in instances when one variable is measured on an interval or ratio scale, for example overall test score, and the other variable is dichotomous, for example, incorrect versus correct answers (Cohen & Cohen 1983, Nunnally & Bernstein 1994, Polit & Hungler 1999). The point biserial correlation coefficient (*rpb*) statistically compares correct and incorrect answers for each question (scored as one and zero respectively) with overall test score performance (Polit & Hungler 1999). Most item to total correlations range from zero to 0.40 (Nunnally & Bernstein 1994). An uncorrected item to total correlation of 0.25 or greater (Beanland et al 1999) is considered to be acceptable.

The size of the discrimination index provides information about the relationship between specific questions and the test in its entirety (Haladyna 1999). MCQs that have high levels of positive discrimination are considered to be the best MCQs in that they are the least ambiguous and do not have extreme degrees of difficulty or simplicity (Gronlund 1968, Nunnally & Bernstein 1994). Zero discrimination occurs when even numbers of participants select either the correct or incorrect answer (Gronlund 1968) while negative discrimination occurs when a question elicits the incorrect answer from those who perform well and the correct answer from those whose overall test performance is poor (Gronlund 1968, Haladyna 1999). MCQs that elicit negative or zero discrimination should be discarded or revised and retested (Gronlund 1968).

In most instances questions with low or negative item to total correlation are eliminated but unlike other measures of reliability, exceptionally high item to total correlations may be considered unfavourable. MCQs with an item to total correlation of greater than 0.70 may be considered redundant because this demonstrates a high degree of similarity or 'overlap' of the concepts that they are measuring (Beanland et al 1999).

Item difficulty analysis

Item response theory states that '... the lesser the difficulty of the item, or the higher the ability of the subject, the greater is the probability of the answer being correct.' (Hutchinson 1991 p8). This concept is fundamental to item difficulty analysis. Item difficulty determines the percentage of participants who selected the correct answer for that question (Gronlund 1968, Nunnally & Bernstein 1994, Haladyna 1999, Linn & Gronlund 2000, Masters et al 2001). High percentages of participants who select the correct answer may indicate a high level of knowledge or well understood instructions, making the test item appear easy. Conversely, low percentages of participants who select the correct answer may indicate inadequate instructions or a poor level of knowledge making the test item appear difficult.

Item difficulty established by the proportion of participants selecting the correct answer is a secondary criterion for MCQ inclusion (Nunnally & Bernstein 1994). Item to total correlations are biased towards items with intermediate degrees of difficulty, so if item discrimination was the only criterion used for item selection, item difficulties of 0.5 to 0.6 would be over represented and test discrimination would be biased towards middle achievers (Nunnally & Bernstein 1994). While variation in item difficulty will cause a decrease in the reliability coefficient it will increase the test's overall ability to discriminate at all levels as long as each MCQ correlates with the overall test score (Nunnally & Bernstein 1994). Nunnally and Bernstein (1994) recommend that if the MCQ with the highest indices of discrimination have variation in item difficulty, item difficulty should be ignored in favour of high levels of discrimination.

Distractor evaluation

The distribution of answers selected for each question should be examined. A good distractor should be selected by those who perform poorly and ignored by those who perform well (Gronlund 1968, Haladyna 1999, Linn & Gronlund 2000). Distractors that are not chosen or are consistently chosen by only a few participants are obviously ineffective and should be omitted or replaced (Nunnally & Bernstein 1994, Linn & Gronlund 2000). If a distractor is chosen more often than the correct answer, this may indicate poor instructions or a misleading question (Nunnally & Bernstein 1994).

Final selection of multiple choice questions

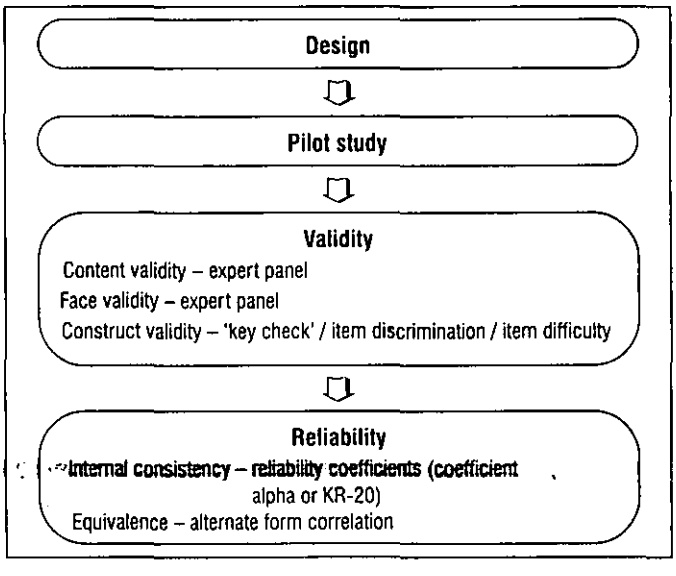
When designing a set of MCQs, more questions than required should be written to allow for the elimination of questions found to be redundant by reliability and validity studies. In the first instance, MCQs should be ranked by their discrimination indices (item to total correlations) and a reliability coefficient should be calculated for the set of MCQs with the highest item to total correlations (Nunnally & Bernstein 1994). The least discriminating MCQs should be replaced with MCQs that have more desirable item difficulty values (Nunnally & Bernstein 1994).

An initial MCQ set should range from five to 30 MCQs. Low average item to total correlation and a high intended reliability may indicate a need for more questions (Nunnally & Bernstein 1994). If this set of MCQs produces the intended level of reliability then addition of further MCQs is not needed. If the reliability is lower than desired, five or 10 MCQs should be added, based on item difficulty analysis. It should be noted that the addition of poorly discriminating MCQs (MCQs with an item to total correlation of less than 0.05) will not significantly increase the reliability coefficient and may result in a decreased reliability coefficient (Nunnally & Bernstein 1994). A summary of the process of design, analysis and selection of MCQs to be used in nursing research is presented in Figure 1.

Implications for nursing research and education

MCQs are often the key criteria by which learning outcomes

Figure 1: Summary of design, analysis and selection of MCQs.



are evaluated in nursing research or nursing education programs. Many of the research studies involving the use of MCQs are testing an intervention designed to affect knowledge either through an experimental or quasi-experimental design using pre-test / post-test measurements. MCQ design and establishment of validity and reliability are therefore fundamental to the rigour of the research and should be undertaken using a systematic process if the results are to be considered valid. The use of MCQs to assess knowledge is also common in nursing education. MCQs are used by universities in both undergraduate and postgraduate courses and clinical settings such as hospitals use MCQs to assess knowledge or determine the effectiveness of education programs such as learning packages or inservice education sessions.

High levels of reliably are mandatory when use of the decisions made on the basis of results of MCQs are important, final, irreversible, or have lasting consequences (Linn & Gronlund 2000 p132). This raises many questions for nurses designing MCQs and nurses who are on the receiving end of MCQs, whether in the nursing education or research context. In terms of nursing research, the results of studies have the potential to change nursing education or practice, making it vital that they are based on sound research methods that have used valid and reliable tools. This raises questions as to the role of results of reliability and validity studies. Should they be more important in the process of gaining approval by Research and Ethics Committees, should they be made available to research participants when they are deciding to be part of a study or should they have more emphasis in the methodology sections of nursing research publications?

The importance of MCQs in the nursing education context has been experienced by most nurses at some stage in their career. Poor performance in an exam or test using MCQs has the potential to thwart achievement of an academic qualification and as a consequence impact adversely on career pathways or inhibit accreditation or credentialling processes in a clinical domain. When scrutinising the use of MCQs in these contexts, further questions are raised. What guarantees do students have that the MCQs that may have such profound and long term effects are valid and reliable? Do universities and hospitals subject the MCQs that they use to reliability and validity analyses and should students have the right to see the results of these analyses to be certain that judgements regarding their knowledge are accurate? Not withstanding that the answers to these questions require debate and discourse among nursing academics, educators and researchers, they highlight the importance of ensuring that MCQs used in nursing are designed using a structured approach and that validity and reliability are established using rigorous and scientific processes.

Acknowledgements

This study was supported by a Deakin University Postgraduate Scholarship and the Annie M Sage Scholarship from the Royal College of Nursing, Australia National Research Scholarship Fund. The authors would like to acknowledge support received from the Emergency Department at The Northern Hospital, the Executive of The Northern Hospital, Professor Barry McGrath, Monash University and Professor Neil Paget, Monash University.

References

Beanland C, Schneider Z et al 1999 *Nursing research: methods, critical appraisal and utilisation*. Mosby, Sydney

Catts R 1978 *Q. How many options should a multiple choice question have? (a) 2, (b) 3, (c) 4*. Assessment Research & Development Unit, NSW Department of Technical and Further Education, Sydney

Cohen J, Cohen P 1983 *Applied multiple regression / correlation analysis for the behavioural sciences*. Lawrence Erlbaum, Hillsdale, NJ

Cronbach L 1990 *Essentials of psychological testing*. Harper Collins, New York

Gravetter F, Wallnau L 2000 *Statistics for the behavioral sciences*. Wadsworth / Thomson Learning, Stamford

Gronlund N 1968 *Constructing achievement tests*. Prentice Hall, New Jersey

Haladyna TM 1994 *Developing and validating multiple-choice test items*. Lawrence Erlbaum, New Jersey

Haladyna TM 1999 *Developing and validating multiple-choice test items*. Lawrence Erlbaum, New Jersey

Hutchinson T 1991 *Ability, partial information, guessing: statistical modelling applied to multiple-choice tests*. Rumsby Scientific, Adelaide

Isaacs G 1994 *Multiple choice testing: a guide to the writing of multiple choice tests and to their analysis*. Higher Education Research and Development Society of Australasia, Campbelltown, NSW

Linn R, Gronlund N 2000 *Measurement and assessment in teaching*. Prentice Hall, New Jersey

Masters JC, Hulsmeyer BS et al 2001 Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *The Journal of Nursing Education* 40(1):25-32

Nunnally J, Bernstein I 1994 *Psychometric theory*. McGraw Hill, New York

Polger S, Thomas SA 2000 *Introduction to research in the health sciences*. Churchill Livingstone, Edinburgh

Polit DF, Hungler BP 1999 *Nursing research: principles and methods*. Lippincott Williams & Wilkins, Philadelphia

Violato C 1991 Item difficulty and discrimination as a function of stem completeness. *Psychological Reports* 69(3 Pt 1):739-743