# CONSUMER NOTES CLINICAL INDICATORS: DETERMINING INTER-RATER RELIABILITY WITH MULTIPLE RATERS, NOMINAL CATEGORIES AND SEVERAL CASES

*Cadeyrn J. Gaskin, BBS (Hons), MBS, is a Research Officer, School of Health Sciences, Massey University, New Zealand, and PhD candidate, School of Human Movement, Recreation and Performance, Victoria University, Melbourne, Australia*

*Anthony P. O'Brien, RN, RPN, RGN, BA, MEdStud, PhD, is Regional Academic Coordinator of Nursing and Health Care Practices, Southern Cross University, Lismore, Australia, and Research Associate, School of Health Sciences, Massey University, New Zealand*

*Derrylea J. Hardy, BBS (Hons), is a Research Officer, School of Health Sciences, Massey University, New Zealand*

***Accepted for publication June 2003***

ACKNOWLEDGEMENTS

This research was conducted by Massey University and supported by the ANZCMHN Inc., the New Zealand Branch of ANZCMHN Inc., the Maori Caucus of the New Zealand Branch of ANZCMHN Inc. and (in part) by a grant from the Health Research Council of New Zealand (00/429). The following people are acknowledged for their involvement in the original research project: Anthony J. O'Brien, School of Nursing, University of Auckland, New Zealand; Neil McNulty, Canterbury District Health Board, New Zealand; Erina Morrison-Ngatai, School of Health Sciences, Massey University, New Zealand*; Tom Ryan, Townsville Mental Health Service, Australia; Georgina Skews, (formerly) Central Sydney Area Health Services, Australia; Julie Boddy, School of Health Sciences, Massey University, New Zealand.

*at Waiariki Institute of Technology, Rotorua, New Zealand (during the data collection stages)

**Key words: inter-rater reliability, rater agreement, mental health**

## ABSTRACT

**A method of determining inter-rater reliability when there are multiple raters, nominal rating categories and several cases is described and applied in the development of an instrument for auditing the ANZCMHN (1995) standards of practice for mental health nursing in New Zealand. Clinical statements (n=41) from the O'Brien et al (2002a, 2003) study, which reflected nursing behaviours contributing to the achievement of the standards of practice, were used to audit consumer files. During two Phases, the clinical indicator statements were refined and rules for judging the achievement of each statement from case note documentation were established. The resultant statements have adequate inter-rater reliability for the assessment of nursing practice with respect to the ANZCMHN (1995) standards of practice.**

## INTRODUCTION

The Australian and New Zealand College of Mental Health Nurses (ANZCMHN 1995) has developed six broad standards of practice for mental health nursing in New Zealand. O'Brien et al (2002a, 2003) developed an instrument that facilitates the measurement of these standards by determining the presence or absence of clinical indicators of pivotal mental health nursing behaviours in consumer case notes: the Consumer Notes Clinical Indicator (CNCI) audit booklet (O'Brien et al 2002b). An important stage in the development of this instrument was the establishment of the internal reliability of each clinical indicator statement. The internal reliability, determined by the measurement of the inter-rater reliability of each clinical indicator statement, was strengthened during a two-stage pilot study. This paper focuses on how the inter-rater reliability of each clinical indicator statement was assessed, the method of improving reliability, and the outcome of the rating process.

### Literature review

In a multi-stage research project, O'Brien et al (2002a, 2003) developed clinical indicators for the ANZCMHN (1995) New Zealand standards of practice. In the first stage, two focus groups of experienced mental health nurses who identified as Maori and non-Maori, respectively, generated nursing behavioural statements that contributed to the fulfilment of the standards of practice. The use of two focus groups, which included participants from different ethnic backgrounds (Maori and non-Maori), reflected the bicultural focus of the research. Maori are the indigenous people of New Zealand and, although they represent 16.4% of the New Zealand population (Statistics New Zealand 2001), are over-represented as consumers of mental health services

(Ministry of Health 2002). The behaviours identified in the two focus groups were separately content analysed and worded as clinical indicator statements. The statements were then assessed to determine whether they were likely to be found in nurses' documentation in consumer case notes. Those statements that were unobservable in consumer case notes were included in a second instrument, the Professional Practice Audit Questionnaire (PPAQ) (O'Brien et al 2002c), the development of which is reported elsewhere (O'Brien et al 2002a; Gaskin et al 2003).

In the second study, the clinical indicator statements were included in a 3-round Delphi process to rate the importance of each statement to the fulfilment of their respective standard (Hardy et al in press). Maori and non-Maori nurses and consumers were the participants. Criteria for consensus and importance were used to judge the appropriateness of each clinical indicator statement for measuring the standards of practice. At the completion of the Delphi process, 41 clinical indicator statements met the consensus and importance criteria, and were accordingly incorporated into the draft CNCI audit booklet (O'Brien et al 2002a, 2003).

### Inter-rater reliability

An instrument's internal reliability refers to how consistent it is in measuring a specific attribute (Polit and Hungler 1999). The inter-rater reliability of an instrument is a measure of its internal reliability. Accordingly, many methods to compute inter-rater reliability have been developed (Banerjee et al 1999; Agresti 1992). These methods have been successfully used when there are two raters, for example, Kappa (Cohen 1960) and intraclass Kappa (Bloch and Kraemer 1989); dichotomous rating categories, for example, tetrachoric correlation coefficient (Pearson 1900); ordinal data (Nelson and Pepe 2000; Szalai 1993); interval/ratio data (eg Shrout and Fleiss 1979); or a large number of raters or ratings, for example, $\text{Kappa}_{SC}$ (Szalai 1998), log-linear models (Tanner and Young 1985), and latent-class models (Agresti 1992; Uebersax and Grove 1990).

There is not a measure of agreement, however, when there are multiple raters, several nominal categories for the raters to choose from, and more than one situation being rated, as was the case in the present study, in which there were multiple consumer files. One way to measure inter-rater reliability in this situation is to calculate the proportion of agreement between raters. Although this method has been criticised for not taking into account rater agreement by chance (Cohen 1960), this problem dissipates with increases in the number of nominal categories, raters, or cases to be rated. The binomial distribution can be used to test the statistical significance of the agreement between multiple raters, when there are several nominal categories and more than one case.

Although the determination of statistical significance is useful as an indication of whether the agreement between raters could be attributable to chance, the effect of increases in the number of nominal categories, raters, or cases to be rated leads to lower levels of agreement being found to be significant. Accordingly, the magnitude of agreement between raters should also be used in determining the adequacy of agreement between raters. Based on their observations from the literature in which inter-rater reliability measures have been reported, Shaughnessy and Zechmeister (1997) suggested that agreement of 0.85 or better is acceptable.

In the O'Brien et al (2002a) study, occurrence of the observable clinical indicator statements in consumer files was assessed during two Phases of a pilot study. The objective of this pilot study was to increase the reliability of the statements to consistently measure important aspects of mental health nursing practice. This paper reports on the method used to assess inter-rater reliability, the way in which inter-rater reliability was improved, and the inter-rater reliability of the statements.

## METHOD

### Consumer files

Consumer files (Phase 1, n=8; Phase 2, n=7) that met the inclusion criteria of 'consumers who had had an episode of care within the last 12 months for at least two days as an inpatient, or at least two months in community care' were audited in the pilot study.

### Measure

The draft CNCI audit booklet consisted of observable clinical indicator statements (Phase 1, n=41; Phase 2, n=25) that emanated from the Delphi stage of the O'Brien et al (2002a, 2003) study. Although O'Brien et al found that 86 clinical indicator statements were important to mental health nursing practice, some of these statements could be merged as they covered the same behaviour and other statements were transferred to the PPAQ because they could not be observed in consumer case notes. This refinement produced a smaller set of 41 statements for inclusion in the present study. Clinical indicator statements applied during the pilot study are listed in table 1. The status of each clinical indicator statement was recorded on a four-point nominal scale as present, absent, not applicable, or not rated. The rating, not applicable, was given when the particular clinical indicator statement was not relevant to the consumer whose notes were being audited. For example, some clinical indicator statements were only relevant to consumers who identified themselves as being Maori. The rating, not rated, was used when a rater decided a clinical indicator statement could not be clearly applied to a file. This rating indicated that the clinical indicator statement needed to be reviewed.

### Procedures

The study was conducted in two Phases at a North Island District Health Board Mental Health Service (MHS), with each Phase lasting two days. Four raters were involved in the first Phase. In the second Phase, three raters were involved, two of whom also participated in the first phase. In both Phases, the raters were members of the O'Brien et al (2002a, 2003) research team.

Staff at the MHS randomly selected the files that were used in the research, in line with the inclusion criteria. The researchers independently assessed each file for documented evidence of each clinical indicator statement having occurred, or not having occurred. At the end of each day of the pilot, the ratings of clinical indicator statements in each file were assessed to determine the extent to which the ratings of the statements were the same. Differences between raters on the assessment of clinical indicator statements were discussed. When consensus was reached about how a clinical indicator statement should be interpreted, this information was recorded so that rules for each clinical indicator statement could be established for the final instrument. Clinical indicator statements were removed from the instrument if they were found to be too ambiguous to be consistently interpreted in the same way or if it was found that the nursing behaviour could not be observed in consumer case notes. If the clinical indicator statements were not observable in case notes they were transferred to the PPAQ (O'Brien et al 2002a, 2003).

### Analysis

The magnitude of agreement between raters, and the statistical significance of the agreement, were calculated for each day of the two pilot study Phases. The magnitude of agreement between raters for each clinical indicator statement was calculated by averaging the agreement on each file. The statistical significance was determined, using the binomial distribution, by calculating the probability that the magnitude of agreement occurred by chance. That is, the probability that the raters agree by chance ($P_a$) over a series of files can be expressed as the mean of the probability of agreeing by chance on each file:

$$P_a = \frac{\sum_{i=1}^{F}(p_r)_i}{F}$$

where pr is the probability of rater agreement on a single file and F is the number of files. The probability of rater agreement on a single file ($P_r$) follows a binomial distribution with n - 1 raters and y -1 raters in agreement. Therefore, the probability of $P_a$ or greater agreement between raters ($P_g$) occurring is:

$$P_g = \sum_{i=y-1}^{n-1}(p_a)_i$$

If the value obtained for $P_g$ is less than 0.05 then, by convention, it has met the generally accepted level for statistical significance (Polit and Hungler 1999). The stricter level of 0.01 is often used when erroneously rejecting that the null hypothesis has important consequences.

## RESULTS

Of the 41 clinical indicator statements that were included on the first day of Phase 1, 16 statements were removed because of ambiguity, repetition of other statements, or lack of observability in consumer case notes. Of the remaining 25 clinical indicator statements, 21 had inter-rater reliability values of 0.85 or better. The magnitude of agreement between raters on each clinical indicator statement, for each day, are shown in table 1.

## DISCUSSION

During the process of pilot testing clinical indicator statements for use in an audit tool, 25 statements emerged as being of potential value for measuring the six ANZCMHN (1995) standards of practice. Using Shaughnessy and Zechmeister's (1997) suggestion that a benchmark level of rater agreement of 0.85 appears to be supported by the literature, 21 of the 25 clinical indicator statements, on both days of Phase 2, could be classified as having adequate agreement between raters. Some of the clinical indicator statements, however, remained problematic at the end of Phase 2 of the pilot study.

Of particular concern were the first two clinical indicator statements, 'Tangata whaiora is given a choice of whether they want their cultural issues addressed,' and 'If tangata whaiora has identified specific cultural issues, then access to relevant cultural support is provided for all issues.' On the second day of Phase 1, full agreement between raters in rating these clinical indicator statements only occurred on one of the three files. On another file, raters totally disagreed on the rating of the latter clinical indicator statement. The confusion related to the identification of the consumers' ethnicity. To address consumers' cultural issues, the mental health nurse must establish whether consumers want their ethnicity acknowledged and their cultural needs met. Given the salience of these clinical indicator statements to the New Zealand mental health context, rules were established to increase the consistency with which these statements were rated as having occurred or not having occurred, from the documented evidence in consumer files.

These rules increased the consistency with which the raters interpreted the clinical indicator statements. The rules for finding occurrences of clinical indicator statements enabled raters to discuss each occasion of discovery and provided parameters for their discovery. When it was not clear whether or not a clinical indicator statement had occurred, as a result of the ambiguity of the statement itself or poor quality of documentation in case notes, raters were able to use the rules of discovery as points of departure to argue the case for inclusion, or to exclude on the basis of insufficient evidence.

**Table 1: Mean percentage of inter-rater reliability of ratings of consumer notes clinical indicator statements in pilot study**

| | Phase 1 | | Phase 2 | |
|---|---|---|---|---|
| **Clinical Indicator Statement** | **Day 1** | **Day 2** | **Day 1** | **Day 2** |
| [1]Tangata whaiora (consumers) are given a choice of whether they want their cultural issues addressed. | 56.25 | 87.50** | 75.25* | 78.00* |
| [1]If tangata whaiora has identified specific cultural issues, then access to relevant cultural support is provided for all issues. | 50.00 | 93.75** | 91.75** | 66.67 |
| [1]If a deficit in the provision of culturally safe practice has been identified, then there is evidence of change. | 81.25** | 75.00** | 100.00** | 100.00** |
| [1]The nurse supports tangata whaiora decision to utilise rongoa (Maori medicine/therapies). | 100.00** | 87.50** | 100.00** | 100.00** |
| [1]Maori cultural assessment for Maori tangata whaiora has been conducted. | 100.00** | 87.50** | 100.00** | 100.00** |
| [1]Maori mental health nurses and/or cultural advisors have been consulted regarding care of Maori tangata whaiora and/or whanau (family). | 100.00** | 100.00** | 100.00** | 100.00** |
| [2]The nurse has sought informed consent of tangata whaiora. | 56.25 | 75.00** | 91.75** | 89.00** |
| [2]Tangata whaiora has been informed of their legal rights. | 56.25 | 81.25** | 83.50** | 89.00** |
| [2]Consultation about treatment has taken place with whanau and/or significant others. | 62.50* | 87.50** | 100.00** | 78.00** |
| [2]Tangata whaiora has been informed of support services. | 93.75** | 75.00** | 91.75** | 89.00** |
| [2]Goals are set and reviewed in partnership with tangata whaiora. | 56.25 | 62.50* | 83.50** | 100.00** |
| [2]Tangata whaiora has been given the opportunity to provide feedback on nursing care. | 87.50** | 81.25** | 91.75** | 89.00** |
| [2]Maori tangata whaiora has been asked if they would like a Maori mental health nurse as their advocate. | 100.00** | 87.50** | 100.00** | 100.00** |
| [2]The mental health nurse has observed and supported Maori tikanga/kawa (traditional beliefs/practices). | 100.00** | 87.50** | 100.00** | 100.00** |
| [3]There is a documented nursing assessment. | 75.00** | 93.75** | 91.75** | 78.00* |
| [3]Where restrictions are placed on the tangata whaiora's freedom, there is evidence in the case notes of regular nursing review. | 56.25 | 62.50* | 100.00** | 89.00** |
| [3]There is a completed nursing care plan. | 68.75* | 81.25** | 100.00** | 100.00** |
| [3]There is a rationale for nursing care. | 50.00 | 93.75** | 91.75** | 100.00** |
| [4]The nurse has provided information to tangata whaiora about his/her care. | 75.00** | 68.75* | 100.00** | 89.00** |
| [4]There is a relapse prevention program based on the principles of recovery. | 93.75** | 81.25** | 100.00** | 100.00** |
| [4]Available health and social resources have been used to support tangata whaiora in the community. | 93.75** | 75.00** | 100.00** | 100.00** |
| [4]Nurses collaborate with significant others in providing wellness education. | 75.00** | 81.25** | 100.00** | 100.00** |
| [4]The nurse has provided mental health promotion that focuses on tangata whaiora strengths and wellness. | 93.75** | 93.75** | 91.75** | 89.00** |
| [4]The nurse has provided a health promotion intervention that reflects relevant personal issues. | 75.00** | 62.50* | 100.00** | 100.00** |
| [6]There is a partnership between the nurse and the multidisciplinary team. | 75.00** | 75.00** | 83.50** | 89.00** |

Note: The number of raters during Phase 1 and Phase 2 were 4 and 3, respectively. Files (n=4) were rated on each day of the two Phases, except on day 2 of Phase 2 when a smaller number of files were rated (n=3). *$p<0.05$ **$p<0.01$.

'Tangata whaiora' is the Maori term that refers to all consumers, users, and patients of the mental health service. The term 'Maori tangata whaiora' refers to mental health consumers of Maori ethnicity.

[1,2,3,4,6] relate to ANZCMHN (1995) Standards of Practice 1, 2, 3, 4, and 6, respectively, indicating the ANZCMHN Standard to which each statement most applies.

The dissonance between raters, caused by difficulty in identifying cultural issues, illustrates the importance of having clearly defined rules for determining whether or not a clinical indicator statement has occurred in a file.

Disagreement between raters can occur because of inter-rater differences or variability in methods of rating (Shaughnessy and Zechmeister 1997). Mainly because of greater sophistication in the method of rating the status of clinical indicator statements, agreement between raters increased over the two Phases. Increased levels of agreement may also have been caused by a training effect, as the raters became more familiar with looking for the clinical indicator statements in consumers' files (Judd et al 1991). Steps taken to increase rater agreement were:

- the elimination of value-laden words;
- the development of rules for interpreting the status of each clinical indicator statement;
- the recording of specific nursing behaviours from consumers' case notes that would indicate a clinical indicator statement had been achieved;
- the recording of where evidence of each clinical indicator statement might be found in consumer case notes;
- the development of precise definitions of key terms or phrases within the clinical indicator statements; and,
- the provision of clear rationales for each statement to illuminate the basic principles inherent in the clinical indicator statements and their importance to quality mental health nursing practice.

An Audit Guidebook (O'Brien et al 2002d) was developed for use in conjunction with the CNCI audit booklet. A page from the Audit Guidebook is reproduced in figure 1 to illustrate the level of detail regarding rules and rationale for determining occurrence of each clinical indicator statement. Such detailed information facilitated the improvement of the inter-rater reliability of the clinical indicator statements.

A factor that affected the level of rater agreement, over which the raters had no control, was the varying quality of nurses' documentation in consumer case notes. There was only partial evidence of achievement of clinical indicator statements in some files because of poor documentation by nurses, and, in other cases, it was not possible to discern whether specific entries in the case notes were nurses' notes because there was no designation identification for the entry. Incomplete or ambiguously recorded entries in the case notes increased the degree of rater interpretation and judgement that was required regarding clinical indicator statement occurrence.

The method used in this study to determine inter-rater reliability may be appropriate in other situations where conventional methods are inappropriate. Like all measures of inter-rater reliability, however, obtaining an

**Figure 1: Audit guidelines for CNCI 1.**

| **CNCI 1: Tangata whaiora/consumer is given a choice of whether they want their cultural issues addressed.** | |
|---|---|
| **Rationale** | **All tangata whaiora/consumers have the right to choose to have cultural issues addressed; nurses should be mindful of sexual orientation, gender, age/generation, ethnicity or migrant experience, religion/spirituality, occupation and socio-economic status and disability.** |
| **Definition of Terms** | Choice - The opportunity to make a decision in any direction without coercion, inducement or imposed bias. Right of self determination. |
| | Cultural issues - All of the issues identified by tangata whaiora/consumers that arise from cultural identity/background. |
| | Addressed - Demonstrates that the nurse has taken action to enable the person the opportunity to make a choice about cultural issues. Cultural issues are acknowledged and responded to within the episode of care. |
| **Type of Indicator** | A critical rate-based event assessing whether choice of having cultural issues addressed is provided or not. |
| **Suggested data sources** | Clinical nursing notes - clinical progress notes, nursing assessment forms and care plans; and nursing cultural assessment documentation. The range of cultural issues for example, could include a person with HIV, a young person with a cultural identity crisis, or a person identifying as a particular ethnicity. |
| **Numerator** | Number of files where tangata whaiora/ consumer has been given a choice of cultural issues being addressed. |
| **Denominator** | Total files audited |
| **RULE:** | **If 1 is YES, also answer 18. Clinical notes must provide clear evidence of a choice being given to identify cultural issues and this includes the nurse's recording of an issue such as ethnicity. A person has the right to identify or not identify their particular ethnicity.** |

From *Clinical indicators for mental health nursing standards of practice in Aotearoa/New Zealand: Consumer notes clinical indicators audit guide.* (p.6), by O'Brien et al 2002b, Palmerston North: Massey University. Copyright 2002 by Massey University. Adapted with permission.

adequate number of ratings is important. With an adequate number of ratings the statistic is useful when there are multiple raters and nominal rating categories. This method only gives an indication of the likelihood that the agreement was due to chance, however. Attention should also be paid to the magnitude of agreement because this statistic is vital for determining whether a measure can be used consistently across cases.

## CONCLUSION

This paper has presented a method for determining the inter-rater reliability of a measure when there are multiple raters, nominal rating categories, and several cases being rated. Application of this inter-rater reliability method, in the O'Brien et al (2002a, 2003) pilot study, confirmed the reliability of the 25 clinical indicator statements in the CNCI audit booklet (O'Brien et al 2002b) as measures for the achievement of mental health nursing practice standards. When auditing consumer case notes for documented evidence of specific nursing practices having occurred, inconsistencies between raters were greatly reduced by the determination of strict rules regarding what constitutes 'achievement' of the clinical indicator statements, and clear definitions of all terms. The reliability of measures that audit consumer case note documentation will be strengthened when nurses document their practice more clearly, and include their designations with their signatures in the files. Further research is recommended to establish national benchmarks of the rate of occurrence of the clinical indicators in clinical practice, and to ascertain what low and high rates of occurrence mean in terms of consumer outcomes.

## REFERENCES

Australian and New Zealand College of Mental Health Nurses Inc. (ANZCMHN). 1995. *Standards of practice for mental health nursing in New Zealand.* Greenacres. South Australia: ANZCMHN.

Agresti, A. 1992. Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research.* 1:201-218.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. 1999. Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics.* 27:3-23.

Bloch, D.A. and Kraemer, H.C. 1989. 2x2 kappa coefficients: Measures of agreement or association. *Biometrics.* 45:269-287.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 20:37-46.

Gaskin, C.J., O'Brien, A.P. and Hardy, D.J. 2003. The development of a professional practice audit questionnaire for mental health nursing in Aotearoa/New Zealand. *International Journal of Mental Health Nursing.* 12:259-270.

Hardy, D.J., O'Brien, A.P., Gaskin, C.J., O'Brien, A.J., Morrison-Ngatai, E., Skews, G., Ryan, T. and McNulty, N. In press. The Delphi prioritisation of New Zealand mental health nursing clinical indicators: A bi-cultural study. *Journal of Advanced Nursing.*

Judd, C.M., Smith, E.R. and Kidder, L.H. 1991. *Research methods in social relations.* 6th edn. Fort Worth, TX: Holt, Rinehart and Winston.

Ministry of Health. 2002. AIM: Advice to the Incoming Minister of Health. Retrieved October 11, 2002 from http://www.moh.govt.nz.

Nelson, J.C. and Pepe, M.S. 2000. Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research.* 9:475-496.

O'Brien, A.P., O'Brien, A.J., McNulty, N.G., Morrison-Ngatai, E., Skews, G., Ryan, T., Hardy, D.J., Gaskin, C.J. and Boddy, J.M. 2002a. *Clinical indicators for mental health standards of practice in Aotearoa/New Zealand: A report to the Health Research Council of New Zealand.* Palmerston North. New Zealand: Massey University.

O'Brien, A.P., O'Brien, A.J., McNulty, N.G., Morrison-Ngatai, E., Skews, G., Ryan T., Hardy, D.J., Gaskin, C.J. and Boddy, J.M. 2002b. *Clinical indicators for mental health nursing standards of practice in Aotearoa/New Zealand: Consumer Notes Clinical Indicators audit booklet.* Palmerston North. New Zealand: Massey University.

O'Brien, A.P., O'Brien, A.J., McNulty, N.G., Morrison-Ngatai, E., Skews, G., Ryan T., Hardy, D.J., Gaskin, C.J. and Boddy, J.M. 2002c. *Clinical indicator statements for mental health nursing standards of practice in Aotearoa/New Zealand: A guide to the Professional Practice Audit Questionnaire.* Palmerston North. New Zealand: Massey University.

O'Brien, A.P., O'Brien, A.J., McNulty, N.G., Morrison-Ngatai, E., Skews, G., Ryan T., Hardy, D.J., Gaskin, C.J. and Boddy, J.M. 2002d. *Clinical indicators for mental health nursing standards of practice in Aotearoa/New Zealand: Consumer Notes Clinical Indicators audit guide.* Palmerston North. New Zealand: Massey University.

O'Brien, A.P., O'Brien, A.J., Hardy, D.J., Morrison-Ngatai, E., Gaskin, C.J., Boddy, J.M., McNulty, N., Ryan, T. and Skews, G. 2003. The New Zealand development and trial of mental health nursing clinical indicators - A bicultural study. *International Journal of Nursing Studies.* 40:835-861.

Pearson, K. 1900. Mathematical contribution to the theory of evolution VII: On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A: Containing papers of a mathematical or physical character.* 195:1-47.

Polit, D.F. and Hungler, B.P. 1999. *Nursing research: Principles and methods.* 6th edn. Philadelphia: Lippincott-Williams & Wilkins.

Shaughnessy, J.J. and Zechmeister, E.B. 1997. *Research methods in psychology.* 4th edn. New York: McGraw-Hill.

Shrout, P.E. and Fleiss, J.L. 1979. Interclass correlations: Uses in assessing rater reliability. *Psychological Bulletin.* 86:420-428.

Statistics New Zealand. 2001. Resident Maori population estimates as at 30 June 2001. Retrieved December 21, 2001 from http://www.stats.govt.nz/domino/external/web/prod_serv.nsf/8048d3976fd325824c2567e2000de0fc276 502b12a6e113bcc25694200808060$FILE/alltabls5.xls.

Szalai, J.P. 1993. The statistics of agreement on a single item or object by multiple raters. *Perceptual and Motor Skills.* 77:377-378.

Szalai, J.P. 1998. Kappa$_{SC}$. A measure of agreement on a single rating category for a single item or object rated by multiple raters. *Psychological Reports.* 82:1321-1322.

Tanner, M.A. and Young, M.A. 1985. Modelling agreement among raters. *Journal of the American Statistical Association.* 80:175-180.

Uebersax, J.S. and Grove, W.M. 1990. Latent class analysis of diagnostic agreement. *Statistics in Medicine.* 9:559-572.