Volume 11 • Number 3 • 2008
VALUE IN HEALTH

Reliability of Two Instruments for Critical Assessment of Economic Evaluations

Flora Au, MA,[1] Shirlina Prahardhi, MEc,[1] Alan Shiell, PhD[1,2]

[1]Centre for Health and Policy Studies, Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada;
[2]Population Health Intervention Research Centre, University of Calgary, Calgary, AB, Canada

[Correction added after online publication 4-Sep-2007: The spelling of Shirlina Prahardhi's name has been corrected, and her academic degree has been updated.]

ABSTRACT

Objective: To assess the reliability of two instruments designed for critical appraisal of economic evaluations: the Quality of Health Economic Studies (QHES) scale and the Pediatric Quality Appraisal Questionnaire (PQAQ).

Methods: Thirty published articles were chosen at random from a recent bibliography of economic evaluations in health promotion. The quality of each of these studies was assessed independently by two raters using each of the two instruments. Inter-rater reliability and the agreement between the instruments were measured using an intraclass correlation coefficient (ICC). Cronbach's generalizability theory was also used to assess the sources of variation in quality scores of the studies and to indicate where improvements in reliability could best be made.

Results: Inter-rater reliability was excellent for both instruments (ICC = 0.81 for the QHES and 0.80 for the PQAQ). Agreement between the instruments varied (ICC = 0.77 for rater 1 and 0.56 for rater 2). The biggest source of variation in the scores assigned to the articles was the quality of the study (56% of total variance). Conventional measurement error explained 31% of the total variance. Variation due to rater (<0.1%) and measurement instrument (1.8%) was very low.

Conclusions: The results suggest that the two instruments perform equally well. Choice of instrument can therefore be based on other criteria—simplicity and speed of application in the case of one, and detail in the information provided in the case of the other. There is little improvement in reliability to be gained from using more than one rater or more than one assessment of quality.

Keywords: critical appraisal, economic evaluation, generalizability theory, reliability.

Introduction

As the number of published studies reporting the cost-effectiveness or cost-utility of health interventions increases, it becomes imperative that the eventual users of this evidence have an easy way of assessing its quality. Several checklists that facilitate the critical appraisal of economic studies are available [1–7]. What is less well known is how well these instruments perform in practice. Our aim in this article was to assess the reliability of two methods of assessing the quality of economic evaluations. Reliability is an important attribute of any measurement instrument, being necessary but not sufficient to ensure an instrument's validity [8,9]. It is commonly understood in terms of agreement or stability, but this is not strictly true and the concept is better understood as a measure of an instrument's ability to discriminate consistently among the subjects of the measurement [8]. In the context examined here, inter-rater reliability refers to the ability of different raters to assess the quality of studies consistently with each other. Agreement between the instruments similarly refers to the ability of the two instruments to discriminate consistently among the studies being evaluated.

We examined two quality appraisal measures: the Quality of Health Economic Studies (QHES) scale [7] and the Pediatrics Quality Appraisal Questionnaire (PQAQ) [5]. The QHES contains 16 questions, the choice of which was based on an extensive literature review and the opinions of a panel of international experts in health economic analysis. Each question or criterion carries a weighted point value that was derived from a discrete choice experiment carried out with a second group of international experts. The scale was then validated prospectively using a third group of health economists who compared their subjective global assessment of a sample of studies (using a visual analog scale) to scores obtained by the QHES. A study either meets or fails to meet each criterion, thus scoring either the full weighted value or zero for each question. The perfect score for a study is 100 and the lowest score is 0. The QHES was chosen for this study because of its formal validation and its extensive application [10–14].

Address correspondence to: Alan Shiell, Population Health Intervention Research Centre, University of Calgary, G012E, 3330 Hospital Drive NW, Calgary, AB, Canada T2N 4N1. E-mail: ashiell@ucalgary.ca

10.1111/j.1524-4733.2007.00255.x

We chose the PQAQ for comparison because it stands out among the remaining quality appraisal checklists. It is longer and more detailed than other checklists, and it too has been formally validated [15]. The PQAQ instrument contains 57 items that map into 14 domains. Of the 57 questions, 46 items have response options that are scored: 0 if the article fails the criterion or is impossible to judge; 0.5 if the criterion is met partially; or 1 if the criterion is met fully. Ten items refer to descriptive information about the study. The final item is an overall assessment of the quality of the study. This is scored on a 6-point Likert scale, where 1 means excellent and 6 means worthless [15]. A panel of seven experts in health economic evaluation independently assessed potential items for their importance, the clarity of the questions, and the appropriateness of the response categories. Although each of the 46 quantitative items is given a numerical score, the experts involved in the development of the PQAQ cautioned against computing a summary score. Their argument was that each domain was important and a high score on one domain should not be allowed to mask a low score on other domains.

The PQAQ was developed specifically to evaluate the quality of economic appraisals in pediatrics [15–17]. The authors claim that 9 of the 47 questions are unique to the pediatric population [18]. Scrutiny of these questions suggests this is not the case, however, because it is easy to generalize many of these supposedly unique questions to other populations. For example, one of the questions identified by the developers of the PQAQ as referring only to pediatric studies asks about costs incurred by agencies other than the health-care sector, and refers specifically to identifying and valuing "school and community resources when necessary" (emphasis added). The reference to school resources does make the question specific to children, but the inclusion of community resources gives the question relevance to a wider population. Other questions are equally generalizable, albeit with some liberal interpretation. For example, question 19 asks whether future changes in the productivity of the child and his or her salary are taken into account. The broader relevance of the question is ensured by interpreting "for the child" to refer to whatever population group was the subject of the economic evaluation under scrutiny. One question was specific to children (are school/day-care absences taken into account), but by regarding this question as not applicable for all adult populations, the question can be ignored without affecting the domain score. Finally, for one question (whose quality of life was being measured) it was the response categories, but not the question itself, that were specific to children (was it the child, his or her parent, teacher, carer, or "other" who completed the questionnaire). In this instance, the question is rendered relevant by extensive use of "other."

Methods

We applied the two critical appraisal instruments to a sample of 30 articles drawn at random from a census of all economic evaluations of health promotion that were published in English between 1990 and 2003 [19]. Each article in the census had been assigned a unique identifier, and a random number generator was used without replacement to select the 30 studies. We did not carry out a formal calculation of sample size, but with two raters, a sample of 30 "subjects" is sufficient to detect a difference in reliability of 0.3 or greater (with $\alpha = 0.05$ and $\beta = 0.20$) [20].

All 30 articles were evaluated independently by two of the authors (F.A. and S.P.) using each of the two instruments. We thus have four scores for each article and four comparisons (two sets of scores for inter-rater reliability—one for each instrument, and two sets of scores for inter-instrument, or parallel forms agreement—one for each rater). Each rater's second evaluation of the same article (using the alternate instrument) occurred 12 months after the first evaluation to reduce contamination from memory effects.

The two instruments are each scored differently, and so some manipulation of the scores was required to facilitate their comparison. The QHES provides a single weighted score out of 100. The PQAQ provides separate domain scores but not a single summary score. Nevertheless, the developers of the PQAQ did sum the 46 quantitative items to assess the test–retest reliability of the instrument [5], and following their lead, we have performed the same, transforming the result into a score out of 100 to match the QHES scale. As a further check, we also transformed the QHES score into a six-category Likert score (where a score 0–17 was recoded as 6, 18–33 was recoded as 5, and so on), and compared this to the results of the global rating taken from the last item on the PQAQ checklist (question 57).

An intraclass correlation coefficient (ICC) was used to examine the inter-rater reliability of both instruments, as well as the agreement between the instruments. ICC estimates were derived from an ANOVA [21]. This follows the "classical" concept of reliability and reflects the amount of error, random, and systematic, inherent in any measurement. In each case, a two-way random-effects model was used, with rater and articles regarded as random effects when testing inter-rater reliability, and instruments and articles being regarded as random in the comparison of the two appraisal instruments [22,23]. A weighted kappa score (using quadratic weights) was used to test for differences between the PQAQ overall score and the transformed (categorical) QHES score.

Table 1 Inter-rater reliability

Instrument	Intraclass correlations (95% confidence intervals)
QHES	0.81 (0.64–0.91)
PQAQ	0.80 (0.63–0.90)

QHES, Quality of Health Economic Studies; PQAQ, Pediatric Quality Appraisal Questionnaire.

Finally, we also used Cronbach's generalizability theory to apportion the variation in the quality scores to its sources: articles (a), raters (r), measurement instrument (i), and measurement error [24]. Because each rater evaluated each article using both instruments, we have a fully crossed (a × r × i) design. Cronbach's approach has the advantage over the classical approach to reliability assessment using the ICC, in that it considers all sources of measurement error at the same time.

The analysis was conducted using SPSS version 14.0 and STATA version 8.

Results

For inter-rater reliability, the ICC was 0.81 (95% confidence interval [CI] 0.64–0.91) for the QHES instrument and 0.80 for the PQAQ (95% CI 0.63–0.90) (Table 1). The ICC between the two instruments was 0.77 for rater 1 (95% CI 0.57–0.88) and 0.56 for rater 2 (95% CI 0.26–0.76) (Table 2). Weighted kappa scores for agreement between the overall scores provided by each instrument were 0.74 for rater 1 and 0.83 for rater 2 (Table 2).

Turning to the generalizability assessment, the biggest source of variation in the scores assigned to each article was systematic differences in the quality of the articles themselves, representing 56% of the variance (Table 3). The proportion of the variance explained by the use of two raters (<0.1%) or by the use of two instruments (1.8%) is very small. So too is the share of the variance explained by the two-way interaction terms: article and rater (a × r); article and instrument (a × i); and rater and instrument (r × i). Thus, each rater scores the different articles consistently, each rater uses each instrument consistently, and the performance of each instrument is not affected by the type of article being appraised. This suggests that studies vary in their quality, and that the two instruments pick up the differences.

Table 2 Agreement between instruments

Rater	Whole scale Intraclass correlations (95% confidence intervals)	Summary question Weighted Kappa (95% confidence intervals)
Rater 1	0.77 (0.57–0.88)	0.74 (0.59–0.77)
Rater 2	0.56 (0.28–0.76)	0.83 (0.82–0.85)

The residual variance (a × r × i, e) contributes nearly 31% of the total variation and is the second biggest source of variation behind differences in the quality of the studies. The residual is made up of two things: first, a three-way interaction effect between article, rater, and instrument (ari); and second, conventional measurement error or unidentified sources of variation (e). Generalizability theory is unable to distinguish between these [25].

Discussion

There is no objective way of interpreting the ICCs, thereby assessing the degree of reliability. Landis and Koch have suggested arbitrary thresholds, with an ICC (or kappa coefficient) above 0.8 indicating excellent agreement [26]. By this convention, our results for the full questionnaires show excellent levels of inter-rater reliability and acceptable to high levels of agreement between the two instruments. Results for the overall assessment (the PQAQ question 57 vs. the transformed QHES scores) were similar.

For the PQAQ, inter-rater reliability was higher than that reported by the developers of the instrument (0.85 vs. 0.75) [5]. The confidence intervals are quite wide though, and in no case does the lower bound exceed the minimum threshold of 0.75 suggested by Lee and colleagues [27]. The intervals we report here are those generated by the SPSS program and are based on the method of Shrout and Fleiss [21]. This method has a tendency to produce liberal confidence intervals, especially at the lower bound [28]. Nevertheless, we reworked the intervals using the method suggested by Rousson and colleagues [28] and found no improvement.

Most of the observed variance in the critical appraisal scores arises from differences in the quality of the studies. There is very little systematic variation in the assessment of quality arising from either choice of rater or choice of quality appraisal instrument. The residual variation is high, which is indicative of measurement error or a three-way interactive effect. An interaction between articles, raters, and instrument is unlikely, however, given the absence of any two-way interaction effects, and so this is most likely simple measurement error. (One source of measurement error is the number of items in each scale. With only 16 items in the QHES, each question carries an average of 6% of the total score. For the PQAQ, each question carries 2% of the total score. A small misclassification by one or other rater therefore has a bigger effect on measurement error for the QHES.)

A more substantive criticism can be made of our use of an unweighted sum of the scored items on the PQAQ to compare it to the QHES. This ignores the summary rating (question 57) and treats each domain of the PQAQ as being of equal importance. The

Table 3 Variance components for crossed (a × r × i) design

Source of variation	Sum of squares	*df*	Mean squares	Estimated variance components	Percentage of total variance
Article (a)	22,317.04	29	769.55	158.65	56.1
Reader (r)	24,223.08	58	417.64	0.00	<0.1
Instrument (i)	24,441.88	58	421.41	2.41	0.9
a × r	5,929.71	61	97.21	5.08	1.8
r × i	5,710.91	61	124.82	18.88	6.7
r × i	1,221.69	3	407.23	10.67	3.8
a × r × i, e	7,835.75	90	87.06	87.06	30.8

developers of the PQAQ warn against summing items across domains. They did not explore the consequences of this empirically however, and so we cannot tell whether their concerns are well founded. They did find that the evaluations were better in some domains relative to others and there was variation in the strength of the relationship between each domain score and the summary assessment of quality (question 57). The developers of the QHES did examine this issue by comparing the generic rating given to a subsample of studies by their expert panel with both weighted and unweighted scores on the QHES. Correlations between all three methods were very high, suggesting that weighting and/or summing items did not make a large difference to the overall assessment of quality. Our own results support this. Agreement between the instruments was high and remained so whether we compared the weighted QHES to the summed PQAQ or the transformed categorical QHES to the global PQAQ question.

Our results show that both instruments perform well in terms of inter-rater reliability. Choice between them can therefore be made according to their other qualities. The QHES is shorter, and is easier and quicker to use. It contains less information about each study than the PQAQ, however. Thus, if speed is of the essence and only a summary score is needed, then the QHES is adequate. If there is more time for the appraisal, and the value of the extra information that will be documented justifies the extra cost, then the PQAQ may be a better choice. There is, of course, nothing to stop one scoring the quality of the study with the QHES and recording any additional descriptive information about each study separately using the types of additional questions featured in the PQAQ. The disaggregated, domain scores on the PQAQ point more precisely to where there are problems with a particular study, and so the PQAQ will be more useful when such a detailed critique is necessary. This may be important in helping the end user to interpret the quality score and in preventing premature dismissal of studies that do not score highly but contain useful information [13].

We thank Gisela Engels for advice on the statistical analysis and Stuart Peacock for providing very helpful comments on an earlier draft. The views expressed in this article and any remaining errors are those of the authors alone.

Source of financial support: This research was funded through an establishment grant awarded to Alan Shiell by the Alberta Heritage Foundation for Medical Research (AHFMR). Alan Shiell is also supported by an AHFMR Senior Health Scholarship.

Supplementary Material

Supplementary material for this article can be found at: http://www.ispor.org/valueinhealth_index.asp.

References

1 Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ 1996;313:275–83.
2 Gerard K, Seymour J, Smoker I. A tool to improve quality of reporting published economic analyses. Int J Technol Assess Health Care 2000;16:100–10.
3 Gonzalez-Perez JG. Developing a scoring system to quality assess economic evaluations. Eur J Health Econ 2002;3:131–6.
4 Sacristan JA, Soto J, Galende I. Evaluation of pharmacoeconomic studies: utilization of a checklist. Ann Pharmacother 1993;27:1126–33.
5 Ungar WJ, Santos MT. The Pediatric Quality Appraisal Questionnaire: an instrument for evaluation of the pediatric health economics literature. Value Health 2003;6:584–94.
6 Evers S, Goossens M, de VH, et al. Criteria list for assessment of methodological quality of economic evaluations: consensus on health economic criteria. Int J Technol Assess Health Care 2005;21:240–5.
7 Chiou CF, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care 2003;41:32–44.
8 Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford: Oxford Medical Publications, 1998.
9 Thompson B. Score Reliability: Contemporary Thinking on Reliability Issues. Thousand Oaks, CA: Sage Publications, 2002.
10 Curtiss FR. Quality of Health Economic Studies (QHES)—tool or mask? J Manag Care Pharm 2003;9:93.

11 Motheral B. Assessing the value of the Quality of Health Economic Studies (QHES). J Manag Care Pharm 2003;9:86–7.
12 Ofman JJ, Sullivan SD, Neumann PJ, et al. Examining the value and quality of health economic analyses: implications of utilizing the QHES. J Manag Care Pharm 2003;9:53–61.
13 Stearns SC, Drummond M. Grading systems for cost-effectiveness studies: is the whole greater than the sum of the parts? Med Care 2003;41:1–3.
14 Spiegel BM, Targownik LE, Kanwal F, et al. The quality of published health economic analyses in digestive diseases: a systematic review and quantitative appraisal. Gastroenterology 2004;127:403–11.
15 Ungar WJ, Santos MT. The Pediatric Economic Database Evaluation (PEDE) Project. Ottawa: Canadian Coordinating Office for Health Technology Assessment, 2002.
16 Ungar WJ, Santos MT. The Pediatric Economic Database Evaluation (PEDE) Project: establishing a database to study trends in pediatric economic evaluation. Med Care 2003;41:1142–52.
17 Ungar WJ, Santos MT. Trends in paediatric health economic evaluation: 1980 to 1999. Arch Dis Child 2004;89:26–9.
18 Ungar WJ, Santos MT. Quality appraisal of pediatric health economic evaluations. Int J Technol Assess Health Care 2005;21:203–10.
19 Rush B, Shiell A, Hawe P. A census of economic evaluations in health promotion. Health Educ Res 2004;19:707–19.
20 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17:101–10.
21 Shrout PE, Fleiss JL. Intra-class correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
22 Muller R, Buttner P. A critical discussion of intra-class correlation coefficients. Stat Med 1994;13:2465–76.
23 McGraw KO, Wong SP. Forming inferences about some intra-class correlation coefficients. Psychol Methods 1996;1:30–46.
24 Cronbach LJ, Rajaratnam K, Gleser GC. Theory of generalizability: a liberation of reliability theory. Br J Stat Psychol 1963;16:137–63.
25 Gleser GC, Cronbach LJ, Rajaratnam N. Generalizability of scores influenced by multiple sources of variance. Psychometrika 1965;30:395–418.
26 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
27 Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989;19:61–70.
28 Rousson V, Gasser T, Seifert B. Confidence intervals for intra-class correlation in inter-rater reliability. Scand J Stat 2003;30:617–24.