Journal of Pediatric Psychology, Vol. 25, No. 6, 2000, pp. 381–391

The Parent-Form Child Health Questionnaire in Australia: Comparison of Reliability, Validity, Structure, and Norms

Elizabeth Waters,[1,2,3] MPH, ***Louisa Salmon,***[1] BA, and ***Melissa Wake,***[1] MBChB, MD
[1]Centre for Community Child Health, Royal Children's Hospital, [2]University of Melbourne, and [3]University of Oxford

Objective: **To improve the ability to describe and compare child health within and between countries, using standardized multidimensional child health measures.**

Methods: **Data on population-specific psychometrics, the measurement structure, and norms are a vital prerequisite. These properties for the Child Health Questionnaire (CHQ) were examined for an Australian population and compared with the originating U.S. data. The CHQ 50-item parent-report was completed by 5,414 parents of children aged 5–18 years. Multi-item/multi-trait analysis tested convergent and discriminatory validity. Construct validity, test-retest reliability, comparative population mean scale scores, and the summary score factor structure were examined.**

Results: **Item and scale internal consistency and item discriminant validity results were good to excellent, and construct (concurrent) validity was supported. Australian children had higher scores than U.S. children except for Family Activities and Physical Functioning. The factor structure of the two summary scores for American children was not replicated in the normative sample but held for a subsample of children with one or more health conditions.**

Conclusions: **The CHQ PF50 performed well in Australia at item and scale level. However, the physical and psychosocial summary scores are not supported for population-level analyses but may be of value for subgroups of children with health problems.**

Key words: *child health questionnaire; Australia; pediatric psychology.*

Rapid growth in the development of comprehensive questionnaires measuring children's functional health status and quality of life means that choice of instrument is already becoming a challenging task (Eisen, Ware, Donald, & Brook, 1979; Koopman et al., 1997; Landgraf, Abetz, & Ware, 1996; Lindstrom & Eriksson, 1993; Nelson, Wasson, Johnson, & Hays, 1996; Spencer & Coe, 1996; Stein & Jessop, 1990; Vogels et al., 1998). The user must decide whether to use a general or condition-specific measure or a combination of the two. Ideally, one must find a measure with appropriate content and sound

All correspondence should be sent to Elizabeth Waters, Centre for Community Child Health, Royal Children's Hospital, Flemington Road, Parkville, Victoria, Australia 3052. E-mail: elizabeth.waters@cryptic.rch.unimelb.edu.au.

psychometric properties. The instrument of choice needs to be practical, reliable, valid, discriminative or sensitive to change, and, preferably, have normative data available for the country in which it will be used (Jenkinson, 1998).

The Child Health Questionnaire (CHQ; Landgraf et al., 1996) is a multidimensional generic health status questionnaire developed for clinicians and researchers interested in measuring children's functional health and well-being. It is available as a parent/proxy report for children aged 5–18 years (the 50-item CHQ PF50, which is the focus of this article, and the short form 28-item CHQ PF28) and as a corresponding self-report for adolescents. The CHQ PF50 includes 13 single and multi-item scales that tap concepts contributing to overall functioning and well-being for children in the context of their family and social environments. One of the purported advantages of the CHQ PF50 is the availability of two summary scores (psychosocial and physical), which may be used in the evaluation of outcomes when information at the scale level is not practical. The existing literature, primarily published by the developer, supports the validity of the CHQ in assessing the overall burden imposed by health problems on functional health and well-being of children, and U.S. population norms are available (Landgraf et al., 1996).

Over the past 2 years we have adapted and evaluated the CHQ PF50 for the Australian population in preparation for its wider application in Australian hospital, outpatient, and primary care settings. The first stage included examination of face validity and linguistic comprehensibility using structured parent interviews and Australian forward and backward translation, followed by the preliminary study of psychometrics and procedural methodologies in a school-based study (Waters, Wright, Wake, Landgraf, & Salmon, 1999). The second stage aimed to comprehensively evaluate the psychometrics of the CHQ and collect normative data.

Summaries of item internal reliability, item discriminant validity, and scale internal consistency are currently being published with norms by age, gender, parental socioeconomic status, and measures of construct validity (Waters, Salmon, Wake, Hesketh, & Wright, 2000). The goals of this study were to compare U.S. and Australian results to assess whether the instrument performance is repeated in Australia. We aimed to (1) extend the psychometric evaluation to include specific results of item internal consistency reliability and discriminant validity, scale internal consistency; criterion validity; test-retest reliability at 2 weeks and 6–8 weeks; (2) calculate and compare CHQ PF50 data at the scale level to assess any significant differences; and (3) analyze and compare the CHQ PF50 factor structure at the item, scale, and summary score level.

Method

Sample

Australian Sample. Australian normative data were drawn from data collected by the Health of Young Victorians Study (HOYVS), a school-based epidemiological study of the health and well-being of children ages 5–18 years conducted between July and November 1997 across Victoria (a state of Australia with a population of 4,689,800 [1998] and an annual birth cohort of 61,143 [1996]). Ethics approval was obtained from the Royal Children's Hospital Ethics in Human Research Committee. A stratified two-stage sample design was employed to sample children representative of the population by age and school education sector (state government, Catholic, and independent). Only schools with total student enrollments over 100 (Australian Bureau of Statistics [ABS], 1996) were included in the sampling. At the second stage, one intact class at each year level was randomly selected, unless a school had less than 30 students in each year band (approximating a total enrollment of less than 240 primary or 210 secondary students), in which case the entire school population was sampled. Participants were parents/proxy respondents of children aged 5–18 years who consented by returning the questionnaire. Information to parents and the questionnaire were written in English of grade 6 equivalent. In total, 7,533 questionnaires were sent home via the selected child. The representativeness of the sample was assessed by comparing proportions of key characteristics with population data (ABS, 1998) and schools census data (ABS, 1997).

U.S. Sample. The U.S. normative data were collected between October and December 1994 as part of the cross-sectional National Survey of Functional Health Status, a cross-sectional survey that included the CHQ. Respondents were drawn from participants in the 1989 and 1990 General Social Survey, which has surveyed the noninstitutionalized adult U.S. population annually over the last 20 years. Of the 2,243 households to be surveyed, 632 included

families with children. Of these, 572 (91%) were located and mailed the CHQ. For families with more than one child, parents (one respondent) were asked to respond for the child with the most recent birthday. Sociodemographic characteristics (age, sex, ethnicity) of the sample from which the normative sample was drawn were compared to those of the U.S. general population reported by the National Center for Health Statistics and support the representativeness of the sample used to obtain U.S. normative data (Landgraf et al., 1996).

Sample Size and Power Calculations

The U.S. CHQ manual (Landgraf et al., 1996) recommends sample sizes necessary to detect small to large group differences in mean CHQ PF50 scale and summary scores. These sample sizes were calculated using variance estimates from the general U.S. normative sample and the six clinical samples (total sample size = 954). All sample sizes were calculated assuming a nondirectional hypothesis (two-tailed distribution) with a "false rejection rate of 5% and with a statistical power of 80%." Five-point differences have been published to be useful for clinical and social differences. Based on this criteria, the Australian normative study aimed to sample 7,500 parents, which, with a response rate of approximately 70% (considered as the likely response from a school-based sample), would achieve approximately 400 questionnaires per year of age. Sample sizes necessary to detect a 5-point difference range from 119 (Mental Health) to 294 (Parental Impact-Emotional) (Landgraf et al., 1996).

Outcome Measures

Parents completed the 11-page, double-sided written questionnaire that included the Aust CHQ PF50 (Authorized Australian Adaptation). The CHQ PF50 consists of 13 health concepts including 11 multi-item and 2 single-item scales. Scales measure Physical Functioning, Role/Social-Emotional/Behavioral, Role/Social-Physical, Bodily Pain, Behavior, Mental Health, Self-Esteem, General Health, Parental Impact-Time, Parental Impact-Emotional, Family Activities, Family Cohesion, and Change in Health (for number of items per scale see tables). For each scale, except Change in Health, item responses are scored, summed, and transformed to a scale from 0 (worst possible health state measured by the CHQ) to 100 (best possible health). The Change in Health item is scored on a continuum of 0–5. The multi-item scales, with the exception of the Family Activities scale, are used to calculate the Psychosocial and Physical summary scores. All questions are based on a retrospective recall of health status over the preceding 4 weeks except for Change in Health, which assesses changes in health over the previous year. The Australian translation and adaptation process resulted in minor wording changes in relation to common children's activities (available from the authors on request).

Statistical Analysis

Multi-trait analysis was used to test the hypothesized item groupings for the CHQ scales by examining internal consistency reliability of the items (Landgraf et al., 1998). The Revised Multi-trait Analysis Program (MAP-R) derived from ANLITH (Analysis of Item-test Homogeneity Program) (Hayashi & Hays, 1987) was used to test item internal consistency, item discriminant validity, and scale level internal consistency reliability. Internal consistency and principal components factor analysis were derived from MAP-R, SPSS (SPSS Inc, 1989–1995) and STATA (StataCorp, 1997).

Internal Consistency Reliability. Item internal consistency was used to test the assumption that the item is linearly related to the underlying concept being measured. Pearson correlations between items and scales (correcting for overlap) and between item means and standard deviations were calculated. Item internal consistency is considered satisfactory if the Pearson correlation between an item and its hypothesized scale is greater than 0.4.

Scale internal consistency was assessed using Cronbach's alpha coefficient (α), which is based on the number of items in a scale and the item homogeneity. It has a correlation of 0–1 with higher values indicating a closer correlation, thus suggesting that the scale is assessing a single domain within the questionnaire. Coefficients above 0.7 and less than 0.9 are recommended (Carmines & Zeller, 1979; Nunnally, 1978; Streiner & Norman, 1995). Using SPSS, we calculated reliability at scale level using the alpha "if item deleted" option.

Item Discriminant Validity. Item discriminant validity assesses the "success" of an item to correlate more strongly with its hypothesized scale than with any other scale within the questionnaire. Success rates are determined using percentage of success; that is, to obtain a high success rate (closer to 100%)

the item-scale correlations must be < 2 standard errors with their hypothesized scale than with other scales (Howard & Forehand, 1962). It is a confirmatory analysis of the factor structure of items within the scales of a questionnaire. For this study, item discriminant validity was measured by assessing (1) the percentage of item-scale correlations with at least 2 standard errors greater than the correlation of the item to other scales and (2) the percentage of item scale correlations greater than the correlation of the item to other scales, but not necessarily by 2 standard errors. The Steiger's *t* test statistic for dependent correlations was used with 95% confidence intervals, equivalent to 2 standard errors, of the correlation coefficient. Item discriminant validity does not directly support the overall ability of an instrument to distinguish or discriminate across conditions but provides evidence of the conceptual logic for placing an item within a particular scale relative to other scales within the instrument.

Content Validity. Concurrent validity of the scales did not involve the use of gold standard instruments or diagnostic criteria against which to verify each childhood condition due to questionnaire length and the absence of a gold standard instrument for many of the health domains included in the CHQ. However, we attempted to examine it with a separate health condition question in the HOYV questionnaire: "Does your child have any of the following conditions?" The prevalence of common childhood conditions in the sample based on this question was similar to current population estimates (Victorian Department Human Services, 1998) and provided some validation for this question. Children were divided into those whose parents reported a condition versus those without a condition. Although this is conventionally used to assess concurrent validity, it remains contentious not only because it is reported by parents usually in response to diagnosis from a health professional but also because of the assumption that a diagnosis or presence of a condition is expected to result in poorer health and well-being on the CHQ, which may not necessarily be demonstrated.

Concurrent Validity

Concurrent validity is assessed by examining the relationship between scores on test/questionnaire and "criterion" scores. We tested this for the CHQ by correlating scores on certain scales of the CHQ with reported health conditions of a similar nature and, for example, would expect scores on the behavior scale to be correlated with reports of behavior-related health conditions (behavior problems, anxiety, depression, and sleep disturbance). For this study, we assessed the concurrent validity of the CHQ by examining the relationships between mental health scale scores and reporting of behavior problems and depression and between the behavior scale and a factored/latent variable combining children with reported behavior problems, anxiety, depression, or sleep disturbance.

Test-Retest Reliability

Test-retest data are not available for the CHQ in the United States. Test-retest reliability for the Australian data was examined using data from samples collected at 2 weeks and 6–8weeks after the initial collection, with two schools in each sample. Identical questionnaire administration methods were used; parents completed a second CHQ PF50 and were asked to report on any major events/happenings that may have affected their child's health and well-being that had occurred since completing the first questionnaire. Results were calculated separately for children who were reported to have experienced an event from those who did not; data were examined for those persons who completed the questionnaire at both time points.

Intra-class correlations (ICC) and Spearman correlations were used to examine differences in scale scores between first and second administration of the questionnaire. An ICC of 0.8 or greater indicates a highly reliable scale (Streiner & Norman, 1995); however, low test-retest correlations may reflect actual change and not necessarily indicate that the instrument has poor reliability (Bowling, 1997). At 2 weeks an overlap period exists between the first and second administrations due to the retrospective 4-week recall, while at 6–8 weeks there is a clear 2–4-week interval in which change may have occurred that is not confounded by the retrospective recall period.

Comparison of U.S. and Australian Normative Data

Mean scales scores between U.S. and Australian children were compared using *t* tests with values of $p < .05$ considered statistically significant.

Table I. Characteristics of U.S. and Australian Normative Samples

Sample details/ characteristics	Australia					United States				
	Method	*M* (SD)	Range	Male	Female	Method	*M* (SD)	Range	Male	Female
Sampling method	Two-stage stratified random sample (child school and class) of children (one parent respondent)					Two-stage stratified random sample (statistical area; race and income) of families (with children) (one parent respondent)				
Achieved sample	5,414					380				
Response rate, %	72					68				
Age, in years										
Parents		40.33 (5.94)	18.89–72.01				40 (7.14)	12–60		
Children		11.58 (3.52)	5.03–18.94				11.5 (3.66)	5–18		
Gender (%)										
Parents				14	85.5				35.2	64.8
Children				50.4	49.6				54.5	45.5

J. Landgraf reports that one "parent" aged 12 years was a sibling of the respective child.

Factor Analysis and Summary Score Coefficients

We used exploratory factor analysis to evaluate and confirm the item-scale and scale-summary score factor structure of the CHQ, to assess whether the items and scales systematically grouped into their hypothesized scales and summary scores, respectively. Two random subsamples were separately drawn, each accounting for approximately 50% of the cases in the complete dataset. Principal components factor analysis with Varimax rotation was used to assess the scale-summary score factor structure for each subsample. Results were used to cross validate the factor structure for the complete dataset. Coefficients derived from the factor score coefficient matrix, obtained from the scale-summary factor analysis, were used to calculate the CHQ summary scores and subsequently compared to U.S. data.

At the scale level, the Australian factor score coefficient matrix was compared with the U.S. factor score coefficient matrix published in the CHQ manual (Landgraf et al., 1996). This matrix combines the U.S. normative data with data derived from children in clinics, thus precluding comparison of factor score coefficients between the two normative samples. The rotated factor matrix for the U.S. population sample was subsequently located (personal communication, Jeanne Landgraf) and compared with results from the current study. In addition, the U.S. factor score coefficient matrix was compared to a number of subsamples from the Australian sample including children with or without illnesses. This analysis was carried out to identify whether the factor score coefficient matrix could be replicated in children with health conditions, akin to the aims of the U.S. developers.

Results

Sample characteristics for each country are shown in Table I. In Australia, 5,414 parents responded (72%). The achieved sample mirrored Victorian Census data (ABS, 1998) for children by age distribution, gender, ethnicity (parental country of birth), and proportion of indigenous persons (Australian Aboriginal and Torres Strait Islanders). Ninety-eight percent (97.6%) were biological parents with step-parents, guardians/foster parents, adoptive parents, and others constituting 0.4%, 0.4%, 0.3%, and 0.8%, respectively. One percent (1.3%) (194) were excluded from psychometric analyses of the individual scales because >50% of the data for that scale was missing. In the United States, 420 families responded to the request for participation (73%), but 29 were elimi-

Table II. Summary of CHQ Item and Scale Internal Consistency, Item Discriminant Validity of CHQ PF50 in Australian Population Sample (n = 5,223) and U.S. sample (n = 380)

Multi-item scales	*k*	Aust *r*	U.S. *r*	Aust α	U.S. α	Aust % + (discriminant validity)	Aust %− (discriminant validity)	Aust total time-scale correlation higher with own scale: %	US total item-scale correlation higher with own scale: %
Physical functioning (PF)	6	0.67–0.84	0.76–0.89	0.91	0.94	100	.0	100	100
Role/social-behavior (REB)	3	0.82–0.85	0.69–0.81	0.92	0.88	100	.0	100	100
Role/social-physical (RP)	2	0.87	0.85	0.93	0.92	100	.0	100	100
Bodily pain (BP)	2	0.78	0.80	0.88	0.89	100	.0	100	100
Behavior (BE)	6	0.35–0.64; (1 k = 0.35, 5 k > 0.57)	0.36–0.67 (1 k = 0.36, 5 k > 0.55)	0.80	0.81	100	.0	100	99
Mental health (MH)	5	0.24–0.59 (1 k = 0.24, 4 k > 0.46)	0.31–0.59 (1 k = 0.31, 4 k > 0.51)	0.71	0.75	90.9	3.6	94.5	91
Self-esteem (SE)	6	0.53–0.73	0.50–0.75	0.83	0.84	100	.0	100	100
General health (GH)	6	0.19–0.50; (1 k = 0.19, 3 k < 0.38, 2 k = 0.50)	0.25–0.55 (1 k = 0.25, 2 k < 0.37, 3 k > 0.47)	0.60	0.66	100	.0	100	100
Parental impact-emotional (PE)	3	0.40–0.58; (1 k = 0.40, 2 k > 0.53)	0.42–0.59	0.68	0.70	90.9	6.1	93.9	82
Parental impact-time (PT)	3	0.50–0.66	0.56–0.71	0.75	0.80	93.9	.0	97	97
Family activities (FA)	6	0.62–0.72	0.78–0.86	0.87	0.93	100	.0	100	100
Total	48	NA	NA	NA	NA	98.1	.8	98.9	NA

r = Pearson's correlation of item-scale internal consistency, α = Cronbach's alpha correlation of scale internal consistency reliability; scaling % + = high item-scale correlations (2 *SE*), %− low item-scale correlations (2 *SE*); k = number of items; NA = not available.
The family cohesion and change in health scales have been omitted from this table because these tests are not applicable for single item scales.

nated because there were no children aged between 5 and 18 years, resulting in a sample of 391 (68% response of located families).

Item and Scale Internal Consistency Reliability

In Australia, the CHQ PF50 demonstrated very good internal consistency across the majority of items and scales (Table II), with the vast majority exceeding the scaling criteria. Only six items had item-scale internal consistency values lower than 0.4, and two scales had α coefficients less than 0.7 (General Health, α = 0.60 and Parental Impact-Emotional, α = 0.68). Alpha coefficients changed marginally across the 2-year age groups. The lowest α was observed on the General Health scale. Its lowest value was for children 13–15 years (α = 0.57) and highest value for children 8–10 years (α = 0.64); with values varying between 0.0–0.07 between each scale. The alpha coefficients differed <0.06 between Australian and U.S. data for all scales. Scale reliability increased marginally (α < 0.07) in five of the eleven multi-item scales if the poorest item was deleted/omitted from the scale (observed for Behavior, Mental Health, General Health, Parental Impact-Emotional, Parental Impact-Time).

Item Discriminant Validity

The scaling results for the tests of item discriminant validity are shown in Table II. Overall success rates were very high, with perfect results attained for eight of the eleven multi-item scales. Mental

Health, Parental Impact-Emotional, and Parental Impact-Time contained items that could, psychometrically, locate within alternative scales. The lowest success rate for both Australian and U.S. samples was for Parental Impact-Time scale (82%, 93.9%).

Concurrent Validity

Statistically significant correlations were found between the Mental Health scale and anxiety problems (Pearson's $r = -.35$, $p < .00$) and depression (Pearson's $r = -.31$, $p < .00$). The Behavior scale also correlated significantly with a separate question about behavioral problems (Pearson's $r = -.50$, $p < .00$), and with a factored composition of anxiety problems, behavior problems, depression, and sleep disturbance (Pearson's $r = -.40$, $p < .00$). Similarly significant correlations were found for physically related items (further data available from authors).

Test-Retest Reliability

For the 2-week test-retest reliability, 17/158 (10.76%) of children in school 1 and 18/191 (9.42%) in school 2 were reported to have experienced a significant event. For the 6–8-week test-retest reliability 22/113 (19.47%) of children in school 3 and 5/139 (3.6%) in school 4 were reported to have experienced a significant event. Schools 1 and 2 were combined for analysis, as were data for schools 3 and 4.

Two-Week Test-Retest Reliability

Where children were not reported to have experienced a significant event, test-retest reliability coefficients for all scales were positive, moderately high, and even (ICC range: 0.49–0.78; Spearman range: 0.54–0.73). Where children were reported to have experienced a significant event, results were similar though negative correlation values were found for Physical Functioning and Role/Social-Physical (ICC range: 0.08–0.77, Spearman range: 0.18–0.77).

Six–Eight-Week Test-Retest Reliability

The coefficient values were moderately high across all scales for children who were not reported to have experienced an event, with similar results to the 2-week test-retest (ICC range: 0.47–0.82, Spearman range: 0.53–0.78, except for Physical Functioning (ICC: 0.05), and Role/Social-Physical (ICC: 0.08), with Spearman Physical Functioning (0.29) and Spearman Role/Social-Physical (0.42). Where children were reported to have experienced an event, only four scales had ICC correlations greater than 0.60 (Behavior, Bodily Pain, Family Activities, Family Cohesion). Five scales had weak (ICC $\leq$ 0.30: Physical Functioning, Role/Social-Physical, Parental Impact-Time, Parental Impact-Emotional and Mental Health) and in some cases, weak and negative ICC correlations (Physical Functioning, Role/Social-Physical, Role/Social-Emotional-Behavior).

Comparisons of Scale Scores With U.S. and Australian Normative Samples

Mean scores and *t* tests between Australian and U.S. normative samples and by gender are shown in Table III. Statistically significant ($p < .05$) differences were found on nine of the twelve single and multi-item scales (Change in Health was excluded from the analysis). Australian children had statistically significant higher scores (i.e., better health) on all scales except for Physical Functioning, Bodily Pain, and Family Activities, although these differences were lower than values considered to be socially or clinically meaningful (>5 points) (Landgraf et al., 1996). By gender, Australian boys had higher levels of health reported especially for General Health (>5 points), but poorer Physical Functioning (not statistically significant) and lower Family Activities ($t = -12.99$). In contrast, no overall trends by country were observed for girls, although Australian girls scored 6 points higher on one scale (Family Cohesion, $t = 12.6$).

Factor Analysis

The exploratory factor analysis at item level produced 11 factors, 8 of which were complete CHQ PF50 scales in which the items systematically grouped into their hypothesized scales (Physical Functioning, Role/Social-Emotional/Behavior, Role/Social-Physical, Bodily Pain, Mental Health, Parental Impact-Emotional, Parental Impact-Time, and Family Activities). The remaining factors were conceptually sound, with items measuring related concepts grouping into separate factors. (The Behavior scale was missing one item.)

Two-Factor Structure of Summary Scores

The two random sample factor analyses extracted from the complete Australian data set produced

Table III. CHQ Scale Mean Score Differences between Australia and U.S. Normative Data

	Entire sample			Males			Females		
Scale	Australia	US	*t* test	Australia	US	*t* test	Australia	US	*t* test
PF	94.63	96.1	−7.11**	94.66	94.9	ns	94.6	97.9	−11.57**
REB	93.78	92.5	5.5**	93.13	91.4	4.98**	94.45	93.9	ns
RP	94.26	93.6	2.81**	94.43	92.5	5.92**	94.08	95.1	−3.02**
BP	81.7	82.44	2.92**	82.97	82.8	ns	81.9	80.3	4.36**
BE	77.38	75.6	8.59**	75.68	74.1	5.16**	79.13	77.5	5.96**
MH	80.13	78.5	9.73**	80.65	79.6	4.49**	79.6	77.1	10.37**
SE	79.98	79.8	ns	80.07	79.6	ns	79.9	79.9	ns
GH	76.93	73	18.06**	77.09	71.6	18.33**	76.75	74.7	6.51**
PE	80.67	80.3	ns	79.35	78	3.53**	82.04	82.9	−2.23*
PT	91.47	87.8	16.31**	90.58	85.8	14.32**	92.39	90.7	5.65**
FA	85.44	89.7	−18.74**	84.57	88.9	−12.99**	86.35	91	−15.18**
FC	76.35	72.3	14.45**	76.62	73.4	8.41**	76.07	70.9	12.6**

ns = p value not statistically significant to .05.
*$p < .05$.
**$p < .01$.

nearly identical factor structures. For the first random sample selected, the total explained variance was 39.81% for factor 1 and 15.24% for factor 2 (total 55.04% explained variance); for the second random sample extracted, the explained variance for factor 1 was 39.49% and 15.47% for factor 2 (total 54.96% explained variance). Table IV provides the coefficients derived from the factor score coefficient matrix for the Australian normative data and the U.S. coefficients, which reveal stark differences. Further, a reanalysis comparing the rotated factor matrix for the U.S. normative data with the Australian data revealed a similar factor structure (Table V). Similarly, as expected, the scale-summary rotated factor matrix for the U.S. data was replicated using a subsample of the Australian data, selecting only those children who had one or more parent-reported health conditions, with the two factors accounting for 52.0% of the total explained variance.

Table IV. Factor Score Coefficients of CHQ Scales for Two Factors: Physical and Psychosocial Health

	Australian norms (n = 5,223)		U.S. norms (n = 380) + clinics (n = 534)	
Scale	Physical	Psychosocial	Physical	Psychosocial
PF	0.38775	−0.17568	0.37138	−.09243
REB	0.24749	0.00246	−0.00178	.21155
RP	0.38393	−0.15991	0.34493	−.06973
BP	0.10283	0.06160	0.27883	−.05514
BE	−0.10526	0.32211	−0.12675	.27911
MH	−0.07213	0.31115	−0.08263	.25335
SE	−0.13685	0.32745	−0.09480	.24792
GH	0.10246	0.09340	0.29460	−.05547
PE	0.02490	0.25167	0.06063	.19823
PT	0.17454	0.08774	0.09113	.16944

U.S. data on summary scores only published as factor score coefficients. Only scales used to compute the summary scores in the U.S. manual.
Family activities and family cohesion are not included.

Discussion

This article is the first to publish a cross-country comparison of the psychometric performance, mean scale scores, and scale and scale-summary score factor structure for the parent-reported CHQ PF50, using a representative population sample. The Australian study has used a large epidemiological design and a representative population sample to compare results with the much smaller U.S. normative sample and the published U.S. summary score results.

Internal consistency reliability of items and scales is generally high. Eight of the eleven multi-item scales had higher correlation values for the placement of their items in their original scale rather than with any other scale, with higher values than the United States.

We compared test-retest reliability of the CHQ PF50 at two time intervals for children whose parents reported whether a significant event had occurred or not. Correlations were generally high between baseline and both the 2- and 6-week reapplications. However, for children who had experienced a significant event, intra-class correlations were particularly weak at 6 to 8 weeks, especially for Physical Functioning, Role/Social-Physical, and Bodily Pain. Changes in reliabilities could be attrib-

Table V. Varimax Rotated Factor Coefficients of CHQ Scales for Two Factors: Summaries of Physical and Psychosocial Health

	Australian children (normative sample) (n = 5,223)		U.S. children (normative sample) (n = 380)		Australian children with one or more health conditions (n = 3,123)	
Scale	Physical	Psychosocial	Physical	Psychosocial	Physical	Psychosocial
PF	0.85923	−0.02279	0.81461	−0.04278	.85563	.03397
REB	0.68406	0.29969	0.52555	0.51234	.47075	.50109
RP	0.86737	0.01591	0.79199	0.02677	.86303	.05730
BP	0.35592	0.29064	0.30261	0.23296	.53609	.13428
BE	0.09156	0.75872	0.22070	0.80432	.02581	.78644
MH	0.16978	0.76788	0.16102	0.75625	.14480	.77067
SE	0.01095	0.73597	0.16623	0.74989	.04117	.70183
GH	0.39256	0.37740	0.47327	0.36629	.45174	.30013
PE	0.36642	0.71962	0.37738	0.68985	.30569	.74980
PT	0.58423	0.44720	0.52572	0.54919	.46096	.58469

Similarity in scale loading on two-factor summary scores of physical and psychosocial health between Australian and U.S. normative data, whereas factor structure for children with health conditions reflects U.S. published summary score factor structure.

uted to familiarity, or resolution of responses to items that have been confronted before. However, they may also be attributable to a change in health status either as a result of a true change in health or from other significant events in a child's life that affect their functioning, health, or well-being. The usefulness of test-retest reliability for instruments that are measuring new and dynamic concepts in a social and family context continues to be debated.

Item discriminant validity testing adds support to the existing structure, with excellent scaling success rates showing that more than 98% of items correlate more highly with their own scale than any other. The results of the exploratory and forced 11-factor analysis conceptually support the existing scale structure of the CHQ PF50, although the psychometric structure diverges for some items. Content validity assessment using a simple corroborating question demonstrates strong correlations between the scale domains and items aiming to tap similar concepts.

As the field of research in health status and quality of life instruments grows, we need to continue to critically analyze the construct and concurrent validity of instruments with gold standard assessments, where possible. This study had the advantage of being large and representative in its sampling, enabling evaluation of age, gender, and demographic effects. However, the administrative methodology and the length of the survey instrument necessitated difficult trade-off decisions between a larger questionnaire that included additional assessments, and efficiencies of data collection within schools. Available instruments measuring aspects of functional status, well-being, and quality of life such as the FSIIR (Stein & Jessop, 1990) and the CHIP (Starfield et al., 1995) were developed for different purposes and aims, and the TACQOL (Vogels et al., 1998) was not publicly available at commencement of the study.

To stringently assess the criterion and construct validity at the level of each scale for an instrument such as the CHQ PF50, many additional measures would have been required. In many of these conceptual domains a gold standard measure does not exist, or where available, would substantially lengthen the questionnaire. Researchers who obtain population samples of parents and children through the school environment need to consider the time required for completion of the instrument and support for the process required of both school staff and parents. Both of these are likely to influence the response rate and the reflection required for completion of the measure. As these measures review the parents' perspective of their child's health over a 4-week period, a pediatric or general clinical assessment would not be measuring similar concepts and is therefore inappropriate for both the time period and the content of the instrument. Clinical studies that employ the CHQ in parallel with clinical assessments are the appropriate study design to evaluate its clinical validity.

Overall, Australian children scored more highly on all but one scale than U.S. children. Unfortunately in this analysis, we were unable to access raw U.S. data to adjust for possible confounding variables such as socioeconomic status, age, or gender,

and recognize that these may influence the differences in mean scale scores. This is, however, a common limitation of cross-country comparisons of any subjective or objective data where local sociodemographic conditions vary in their definition and measurement.

The exploration of the factor structure for the physical and psychosocial summary scores using population normative data has stimulated further discussion of country-specific weights and the appropriateness of summary scores of scales for population data. Initial investigation of the summary scores in Australia was unable to replicate the factor structure of the two summary scores using the population data. In particular, Role/Social-Emotional/Behavioral and Parent Impact/Emotional did not sit within the psychosocial summary score, differing from the published U.S. CHQ summary score structure. These results also differ from those achieved with the adult functional health status Short Form 36 (SF36), whose summary scores were derived from the population sample (Ware & Sherbourne, 1992) and confirmed in other countries such as the United Kingdom and Australia. Upon further discussion with the developer of the CHQ, sufficient variability in the scales to devise two factored summary scores on domains of physical and psychosocial health was only achieved with the addition of clinical sample data to the normative data in the United States. In Australia, we subsequently reanalyzed the factor structure using only a subsample of children who experienced one or more health problems. For this subgroup of children, the two-factor structure mirrors the psychosocial and physical summary scores of the United States.

These results present a quandary for the applicability of the CHQ PF50 summary scores in Australia, or any other country. Given that we have achieved sufficient variability in one subgroup of our population to be able to replicate the summary scores (i.e., children with a health problem), what meaning does this have to their application? Our results do not support the use of two summary scores of psychosocial and physical health in a population sample. With caution, we suggest that summary scores may be relevant or useful for groups of children with more severe illnesses or problems than the normal population, where fewer than 13 scale scores are more useful. We fully recommend that the full range of scales be used for children in clinical and population contexts to observe the impact of a broader range of domains than the summaries provide. However, the nature of clinical work invariably requires clinicians to employ briefer instruments, and for this context it appears that the summary scores are supported as a valid way of summarizing the physical and psychosocial components of health and well-being, thereby providing a two-component summary of the 13 scales.

We have demonstrated that the CHQ PF50 exceeds standard criteria for the evaluation of psychometrics with similar Australian results to those of the United States. As such, it provides an excellent example of a measure that retains its psychometric characteristics within another country and can be considered reliable and sound at the scale level. Nonetheless, we believe that child health researchers using the CHQ (and other standardized instruments) should be mindful of measurement and interpretation pitfalls. For example, a population of parents with predominantly healthy children, or children with a particular diagnosis (for condition specific measures), should be represented in the conceptual development or early evaluation of an instrument. Ideally, the instrument should be developed using the perspectives and research from the population group in which it will be used. This means that for a generic health status measure for use in population and clinical applications, both healthy children and those with common or prevalent conditions should be involved in the development of concepts, appearance, review, and testing prior to more widespread application.

Future research needs to consider the relative value of summarizing children's health into simple scores that may not reflect the dynamic interplay between illness, functioning, health and well-being, and quality of life, but may suit the purposes of clinical outcome evaluation.

Acknowledgments

The study team would like to thank the parents and children who participated; Kylie Hesketh and Dr. Martin Wright of the HOYVS research team, Dr. Rory Wolfe for statistical assistance, Ms. Jeanne Landgraf for additional unpublished data and additional information request, Dr. Sarah Stewart Brown, and Professor Ray Fitzpatrick.

Received March 1, 1999; revisions received May 1, 1999; accepted November 1, 1999

References

Australian Bureau of Statistics. (1996). *Schools Australia 1995*. ABS Catalogue No. 4221.0. Commonwealth of Australia.

Australian Bureau of Statistics. (1997). *Census of population and housing: Selected social and housing characteristics for statistical local areas, Victoria*. ABS Catalogue No. 2015.2. Commonwealth of Australia.

Australian Bureau of Statistics. (1998). *1998 Victorian Year Book*. ABS Catalogue No. 1301.2. Commonwealth of Australia.

Australian Bureau of Statistics. (1998). *Schools Australia 1997*. ABS Catalogue No. 4221.0. Commonwealth of Australia.

Bowling, A. (1997). Measuring health. A review of quality of life measurement scales. 2nd ed. Buckingham: Open University Park.

Carmines, E., & Zeller, R. (1979) *Reliability and validity assessment*. Newbury Park, CA: Sage.

Eisen, M., Ware, J. E., Donald C. A., & Brook R. H. (1979) Measuring components of children's health status. *Medical Care, 17,* 902–921.

Hayashi, T., & Hays, R. D. (1987) A microcomputer program for analysing multitrait-multimethod matrices. *Behaviour Research Methods, Instruments, and Computers, 19,* 345–348.

Howard, K. L., & Forehand, G. C. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Education and Psychological Measurement, 22,* 731.

Jenkinson, C., & McGee, H. (1998). *Health status measurement: A brief but critical introduction*. Oxford: Radcliffe Press.

Koopman, H. M., Kamphuis, R. P., Verrips, G. H., Vogels, A. G. C., Theunissen, N. C. M., Verloove Vanhorick, S. P., & Wit, J. M. (1997). The DUCATQOL: A global measure of quality of life of school-aged children (Abstract). *Quality of Life Research, 6,* 428.

Landgraf, J. M., Abetz, L., & Ware, J. E. (1996). *Child Health Questionnaire (CHQ): A user's manual*. Boston, MA: The Health Institute, New England Medical Center.

Landgraf, J. M., Maunsell, E., Speechley, K. N., Bullinger, M., Campbell, S., Abetz, L., & Ware, J. E. (1998). Canadian-French, German and UK versions of the Child Health Questionnaire: Methodology and preliminary item scaling results. *Quality of Life Research, 7,* 433–445.

Lindstrom, B., & Eriksson, B. (1993). Quality of life among children in the Nordic countries. *Quality of Life Research, 2,* 23–32.

Nelson, E. C., Wasson, J. H., Johnson, D. J, & Hays, R. D. (1996). Dartmouth COOP Functional Health Assessment Charts: Brief measures for clinical practice. In B. Spilker (Ed.), *Quality of life and pharmaco-economics in clinical trials* (pp. 161–169). Philadelphia: Lippincott-Raven.

Nunnally, J. C. (1978). *Psychometric theory.* 2nd ed. New York: McGraw-Hill.

Spencer, N.J., & Coe, C. (1996). The development and validation of a measure of parent-reported child health and morbidity: The Warwick Child Health and Morbidity Profile. *Child: Care, Health and Development, 22,* 367–379.

SPSS for Windows: Release 6.1.3. (1995). Standard version copyright, SPSS Inc. 1989–1995 US.

Starfield, B., Riley, A. W., Green, B. F., Ensminger, M. E., Ryan, S. A., Kelleher, K., Kim-Harris, S., Johnston, D., Vogel, K. (1995). The adolescent child health and illness profile: A population-based measure of health. *Medical Care, 33,* 553–566.

StataCorp. (1997). *Stata Statistical Software*: Release 5. 0. College Station, TX: Stata Corporation.

Stein, R. E., & Jessop, D. J. (1990). Functional Status II®: A measure of child health status. *Medical Care, 28,* 1041–1055.

Streiner, D. L., & Norman, G. R. (1995). Health measurement scales. A practical guide to their development and use. 2nd ed. Oxford: Oxford University Press.

Victorian Department of Human Services. (1998). *The health of young Victorians*. Melbourne: Child Health Unit, Public Health Branch, Victorian Government Department of Human Services.

Vogels, T., Verrips, G. H., Verloove Vanhorick, S. P, Fekkes, M., Kamphuis, R. P., Koopman, H. M., Theunissen, N. C., & Wit, J. M. (1998). Measuring health-related quality of life in children: The development of the TACQOL parent form. *Quality of Life Research, 7,* 457–465.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36). *Medical Care, 30,* 473–483.

Waters, E., Salmon, L., Wake, M., Hesketh, K., & Wright, M. (2000). The Child Health Questionnaire in Australia: Reliability, validity and population means. *Australian and New Zealand Journal of Public Health, 24,* 207–210.

Waters, E., Wright, M., Wake, M., Landgraf, J., & Salmon, L. (1999). Measuring the health and well-being of children and adolescents: A preliminary comparative evaluation of the Child Health Questionnaire. *Ambulatory Child Health, 5,* 131–141.