

Psychometric properties of the questionnaire version (ERS-Q) of the Environmental Rating Scale (ERS) for assessment of residential programmes for individuals with autism

Marie Hubel*, Peter Hagell and Bengt Sivberg

Department of Health Sciences, Faculty of Medicine, Lund University, Sweden

(Received 1 January 2009; accepted 10 November 2009)

The Environmental Rating Scale (ERS) is the only autism spectrum disorders (ASD) specific tool for assessment of residential services and treatment models. However, one limitation with the ERS is its dependence on expert observations and interviews, particularly in larger scale studies. The ERS has therefore been adapted into a staff self-report questionnaire (ERS-Q). Here the measurement properties of the ERS-Q were examined according to traditional test theory criteria. Data provided support for summation of raw item scores into total and subscale ERS-Q scores and item-total correlations indicated that items within scales tap a common construct, suggesting that the ERS-Q is useful in survey as well as interventional studies. As such the ERS-Q appears a valuable addition to the current ASD research toolbox.

Keywords: autism spectrum disorders; ERS-Q; questionnaire; validity; reliability

Due to the fact that there are several community-based treatment strategies (Van Bourgondien et al. 1998) a major force in creating public policy is to document and evaluate the daily care and the pedagogical environment for individuals with ASD. Lord et al. (2005) suggests that measures of outcome in ASD should be a high priority in order to document the goal for and effect of treatment. However, research on the effectiveness of treatment models is limited (Van Bourgondien et al. 1998). Such outcomes are typically measured by means of summed rating scales. Valid interpretation of rating scale data requires that certain criteria are met (Ware and Gandek 1998; Ware et al. 1997; Hobart et al. 2004). Measures of outcome of therapeutic interventions must be rigorously evaluated (Hobart et al. 2004) as these results are used to make judgements about the effectiveness of the care and the pedagogical work for vulnerable individuals not able to express their wishes and rights. By confirming the assumptions underlying the construction of summated rating scales, scores can be calculated with confidence that they have their desired properties and that they represent the variables they are assumed to measure.

The Environmental Rating Scale (ERS) was developed based on well-founded theories of autism spectrum disorder (ASD) (Schopler 1997) and appears to be the only ASD specific instrument for assessment of residential ASD services and treatment models (Van Bourgondien et al. 1998; Van Bourgondien, Reichle, and

*Corresponding author. Email: marie.hubel@med.lu.se

Schopler 2003). However, one limitation of the ERS, particular in larger scale studies, is its dependence on expert observations of residential homes and face-to-face interviews with the staff members.

In an attempt to overcome this, the original interview based ERS was adapted into a staff self-administered questionnaire version (the ERS-Q) and pilot tested among 18 residential staff members (Hubel, Hagell, and Sivberg 2007). The results provided general support for the notion that the ERS can be adapted into a questionnaire without substantial loss of conceptual meaning. However, further evaluations in larger samples are needed to document the measurement properties of the ERS-Q. This study describes the psychometric properties of the ERS-Q in such a sample of residential staff members.

Methods

Setting and participants

All group homes ($n = 26$) for individuals with ASD in a west Swedish municipality were invited, of which 19 (73.1%) homes agreed to participate. The residents of the group homes were adults functioning in the moderate to severe ranges of intellectual disability. All group homes were working according to the Treatment and Education of Autistic and related Communication Handicapped CHildren (TEACCH) (Schopler 1997) philosophy. Activities and daily routines of the homes were also very similar and adapted to the individual residents.

All staff members working at the included group homes for more than two years were eligible for participation. A total of 303 of 360 eligible respondents at the 19 group homes gave written informed consent and responded to the questionnaire (response rate, 84%). The mean age of respondents was 38.5 ($SD = 11.46$) years and 82.5% were female (Table 1). Respondents had been working with ASD for 9.47 ($SD = 8.06$) years and they had been working at the current homes for 5.33 ($SD = 3.51$) years, according to the respondents, explained by the high range of age. A majority (63%) stated that their choice of occupation was based on an interest for individuals with ASD. A majority of the respondents (76.5%) had undergone TEACCH training, declared that they worked according to the TEACCH pedagogy (80.1%), had regular staff meetings (94.8%) and discussed pedagogical issues regularly (96.9%) (Table 1).

The ERS-Q

The ERS-Q was adapted from the ERS on an item-by-item basis (Hubel, Hagell, and Sivberg 2007). As such, it consist of 32 items with a possible score range from 1 to 5, that are grouped conceptually into five subscales (Table 2): Communication (five items); Structure (six items); Socialization (six items); Developmental Assessment and Planning (eight items); and Behaviour Management (seven items). These subscales are considered to reflect various aspects of the overall construct, which is represented by a total score. Similar to the original ERS, the ERS-Q is scored using Likert's method of summing scale items into total scores (Likert 1932). Subscale scores are thus created by summing the item scores in each area and the total score is obtained by summing all 32 item scores. Higher scores indicate opportunities for

Table 1. Respondents' demographic background and data (n = 303).

		Missing (n)
Age	38.50 (11.46) ^b	15
Gender		11
Women	241 (82.5) ^a	
Men	51 (17.5) ^a	
Education		28
High school	210 (76.4) ^a	
College	65 (23.6) ^a	
Pedagogical /TEACCH education	225 (76.5) ^a	9
Workload		13
heavy	124 (42.8) ^a	
moderate	154 (53.1) ^a	
mild	12 (4.1) ^a	
Years experience working with ASD	9.47 (8.06) ^b	12
Years worked at present group home	5.33 (3.51) ^b	7
Frequency of continued TEACCH training		19
> Twice/year	51 (18.0) ^a	
About once/year	124 (43.7) ^a	
< Once/year	100 (35.2) ^a	
Never	9 (3.1) ^a	
Consider themselves working according to a special pedagogy	230 (80.1) ^a	16
Regular staff meetings	276 (94.8) ^a	12
Discussing pedagogical issues	281 (96.9) ^a	13
Daily activities at the group home	70 (23.9) ^a	10

Notes: ^an (%),^bmean (SD).

greater environmental adaptation for the individuals with ASD (Van Bourgondien, Reichle, and Schopler 2003).

Procedures

Data were collected during a two month period in 2007. Study participants received the ERS-Q and a demographic questionnaire from their staff manager at the group homes. Each respondent was the contact person for one resident only and worked directly with this individual on a daily basis. The respondents were instructed to complete the ERS-Q individually by referring to this person. All responses remained confidential to other staff members.

Analyses

Data quality was assessed by examining the amount of missing data. Up to 10% missing item responses has been suggested as acceptable (Saris-Baglana et al. 2004).

To evaluate if it is legitimate to sum up items to scale scores within the traditional test theory framework (Likert 1932; Ware et al. 1997; Ware and Gandek 1998), item descriptive statistics were examined. Item means and standard deviations should be roughly equivalent within a scale, otherwise standardization or weighting of item

Table 2. The Environmental Rating Scale.

Item No.	Item content	Subscale	Possible score range
1	Caregiver's language adjusted	1. Communication	5–25
2	Visual systems supplement communication		
3	Training incorporated into daily routines		
4	Information available about communication skills		
5	Directions communicated clearly to resident		
6	Physical organization, facilitate independence	2. Structure	6–30
7	Visual systems for independence		
8	Daily schedule for client visible		
9	Daily schedule for home visible		
10	Full daily schedule with leisure pursuits		
11	Systems to facilitate transitions between activities	3. Socialization	6–30
12	Socialization training in daily routines		
13	Clearly goals for social, leisure and affective skills		
14	Leisure activities are planned individually		
15	Independent skills developed for free time use		
16	Social skills training used in interactions	4. Developmental Assessment and Planning	8–40
17	Social skills are taught in meaningful context		
18	Caregivers awareness of cognitive level		
19	Caregivers awareness of social level		
20	Training based on assessment information		
21	Training based on individual abilities	5. Behaviour Management	7–35
22	Emerging skills incorporated into training		
23	Functional needs are incorporated into training		
24	Training activities are rethought if necessary		
25	Efforts are made to generalize skills		
26	Limits and rules are clear for the client	5. Behaviour Management	7–35
27	Consistence maintain behavioural limits		
28	Problem behaviour are analyzed		
29	Client reinforced for positive behaviours		
30	More positive strategies than punitive		
31	Less intrusive than intrusive approaches		
32	Documentation of behaviour programmes		7–35
Total score			32–160

scores will be needed. Furthermore, each item should contribute about equally to the total score. This was examined by inspection of the corrected item-total correlations, i.e., the correlation between each item's score and the total score of the other items in its scale, which should exceed 0.30 (Ware et al. 1997).

To support the validity of summing item responses into scale scores there should also be evidence that items in the scale tap a common underlying variable and that they appear correctly grouped into their respective subscales. To assess this, corrected item-total correlations were examined. Different criteria have been suggested, ranging from >0.20 (Streiner and Norman 2000) to >0.40 (Ware and Gandek 1998). To support the hypothesized grouping of items, items should also correlate stronger with the scale they are hypothesized to represent (corrected item-total correlation) than with the other scales. This criterion has been considered satisfied when at least 80% of the hypothesized item-scale correlations are stronger than the alternative item-scale correlations (Saris-Baglama et al. 2004). This aspect can also be more rigorously assessed by determining the number of instances that corrected item-total correlations is significantly stronger than the alternative item-total correlations, as determined by 95% confidence interval around correlations (Ware et al. 1997). Significantly and non-significantly stronger item-total correlations are considered definite and probable scaling success, respectively (Ware et al. 1997). Similarly, significantly and non-significantly weaker item-to-own scale correlations are considered definite and probable scaling failure (Ware et al. 1997). Some degree of scaling failure can be acceptable when scales within an instrument measure correlated constructs as defined by a strong theory (Ware and Gandek 1998).

Next, the amounts of floor and ceiling effects and score reliabilities were examined. Floor and ceiling effects are the proportions of minimum and maximum scores, respectively, observed for each scale. These should not exceed 15% (McHorney and Tarlov 1995). The reliabilities of scale scores were estimated using Cronbach's coefficient alpha (Cronbach 1951), which should be above 0.70 to be regarded as acceptable and preferably exceed 0.80 (Streiner and Norman 2000).

To assess if the subscales tap distinct aspects, scale-to-scale correlations were computed and compared with each subscale's internal consistency (Cronbach's alpha), of which the latter can be viewed as the correlation between a scale and itself. Therefore, alpha values that exceed correlations with other scales are taken as support for the notion that subscales represent distinct aspects of the overall construct (Ware et al. 1997; Ware and Gandek 1998).

All statistical analyses were performed using *SPSS for Windows*, version 11.0 (SPSS Inc., Chicago, 2003).

Results

There was evidence of good data quality as the proportion of missing item responses ranged from 3.3–7.3% (mean, 5.0%), indicating that the ERS-Q is acceptable to the sample. There was also support for summing up items to generate scale scores as descriptive statistics showed roughly similar item means and standard deviations within the five subscales and for the total score (Table 3). Similarly, the total score and all subscales had corrected item-total correlations exceeding 0.30 ranging between 0.32–0.60. The exception was item 23 which had corrected item-total correlations of 0.22 with its subscale (Table 3).

Table 3. Item descriptive statistics and item-scale correlations.

Scale	Item No.	Mean ^a	SD	Item-scale correlations					
				Total ERS-Q Score ^b	Communication ^c	Structure ^c	Socialization ^c	Developmental Assessment... ^c	Behaviour Management ^c
Communication	1	4.14	1.04	0.37	0.51	0.35	0.19	0.28	0.18
	2	3.74	1.39	0.47	0.67	0.55	0.14	0.18	0.26
	3	3.18	1.49	0.47	0.61	0.52	0.32	0.22	0.23
	4	3.44	1.20	0.41	0.40	0.34	0.30	0.23	0.25
	5	3.92	0.87	0.42	0.35	0.41*	0.31	0.27	0.23
Structure	6	3.83	1.08	0.44	0.34*	0.32	0.38*	0.38*	0.28
	7	3.42	1.30	0.60	0.54	0.61	0.37	0.36	0.37
	8	3.77	1.27	0.49	0.53	0.62	0.17	0.25	0.25
	9	3.44	1.33	0.44	0.33	0.50	0.32	0.24	0.30
	10	3.72	1.10	0.38	0.40	0.58	0.18	0.18	0.23
	11	3.94	0.94	0.55	0.51	0.52	0.30	0.33	0.34
Socialization	12	3.56	1.25	0.50	0.36	0.32	0.59	0.33	0.31
	13	3.30	1.22	0.56	0.33	0.37	0.70	0.49	0.28
	14	4.10	0.91	0.37	0.10	0.21	0.38	0.43*	0.25
	15	3.28	1.21	0.39	0.20	0.24	0.53	0.35	0.25
	16	3.49	1.19	0.46	0.25	0.28	0.70	0.33	0.33
	17	3.80	1.01	0.49	0.21	0.26	0.60	0.54	0.34
Developmental Assessment and Planning	18	3.84	1.02	0.39	0.20	0.23	0.35	0.44	0.23
	19	4.10	0.86	0.39	0.84	0.13	0.22	0.43	0.22
	20	4.13	0.88	0.38	0.16	0.26	0.28	0.54	0.35
	21	4.15	0.81	0.35	0.09	0.18	0.24	0.51	0.34
	22	3.63	1.04	0.57	0.34	0.44	0.47	0.50	0.39
	23	2.85	1.41	0.40	0.32*	0.22*	0.46*	0.22	0.25*
	24	3.81	1.09	0.53	0.32	0.31	0.46	0.51	0.42
	25	3.98	0.93	0.49	0.24	0.25	0.40	0.54	0.43

Table 3 (Continued)

Scale	Item No.	Mean ^a	SD	Item-scale correlations					
				Total ERS-Q Score ^b	Communication ^c	Structure ^c	Socialization ^c	Developmental Assessment... ^c	Behaviour Management ^c
Behaviour Management	26	4.13	0.86	0.45	0.21	0.30	0.29	0.45*	0.43^b
	27	4.29	0.87	0.43	0.14	0.33	0.27	0.43	0.49^b
	28	4.32	0.82	0.44	0.20	0.34	0.22	0.39	0.51^b
	29	4.44	0.74	0.32	0.17	0.14	0.25	0.27	0.47^b
	30	4.37	0.75	0.43	0.25	0.24	0.25	0.31	0.51^b
	31	4.04	1.06	0.46	0.40*	0.36	0.28	0.29	0.38^b
	32	4.40	0.92	0.33	0.13	0.27	0.21	0.28	0.38^b

Notes: ^aPossible item score range, 1–5; ^bCorrected item-total correlation; ^cBolded values are corrected item-total correlations; unbolded values are correlations between each item and other scales. *Item-to-other scale correlation stronger than the corrected item-total correlation.

With few exceptions, corrected item-total correlations also exceeded the 0.40 criterion, suggesting that items within scales tap a common construct (Table 3). The corrected item-total correlations were stronger than the item-to-other subscale correlations in 87.5–95.8% (mean, 91.7%) of instances across the five subscales (Tables 3 and 4). This provides general support for the suggested grouping of items into subscales. When taking statistical criteria into consideration, the definite scaling success rates varied between 32.1% and 75%, and there was no instance of definite scaling failure (Table 4).

The distribution of scores of the ERS-Q spanned the entire subscale range with no notable floor or ceiling effects. Floor effects were virtually absent (0.4% in the Structure scale, otherwise 0% and ceiling effects ranged between 0.8% (ERS-Q total score) and 8.1% (Behaviour Management). Reliability of the total ERS-Q was good (0.90), and 0.73 or above for all subscales (Table 5). The reliability coefficients were substantially greater than inter-scale correlations for all subscales (Table 6). This suggests that the ERS-Q subscales represent related but distinct constructs.

Discussion

This study assessed the psychometric properties of the ERS-Q. The result showed good data quality, general support for assumptions underlying summation of ERS-Q items into subscales and total scores and acceptable score reliabilities. Furthermore, there were no notable floor or ceilings effects, and the five ERS-Q subscales appear to represent related but distinct constructs.

The corrected item-total correlations varied but were above the 0.30 level except for item 23 i.e., item 6 (functional needs are incorporated into training) in the Developmental Assessment and Planning subscale. This item also showed signs of scaling failure in that its correlations with other subscales were stronger or equal to that with its hypothesized subscale, and had a somewhat lower mean score than other ERS-Q items did. This could indicate that this item should be revised or deleted from the ERS-Q. Speaking against deleting the item is the fact that it appears to work well in the context of the total ERS-Q score and that the Developmental Assessment and Planning subscale (involving this item) appears to represent a separate but related construct from the other subscales. This implies that the construct is sufficiently anchored by the subscale items to allow for item 6 to be retained. Possible explanations for the observed problems with this item in the Developmental Assessment and Planning subscale may relate to cultural differences between Sweden and the US (where the original ERS was developed) and/or ambiguities in wording. However, no such indications were found in the pilot study of the ERS-Q (Hubel, Hagell, and Sivberg 2008). Similar arguments also hold for other observations of potential item level problems.

The fact that a small subset of items correlated stronger with other subscales than their own is not surprising. In their factor analysis of the original ERS, Van Bourgondien et al. (1998) found factor (subscale) interrelatedness. This idea is also supported by the theory behind the ERS (Schopler 1997; Van Bourgondien et al. 1998). Taken together, we therefore do not consider the current observations to justify any modifications of the ERS-Q at this stage.

Floor and ceiling effects were absent or trivial for all ERS-Q scores. This is an important observation for two related reasons. First, it indicates that the scale scores are able to reflect the relevant levels of therapeutic environmental qualities among the

Table 4. Frequency and percentage of item-scale correlations at each level of scaling success.

Scale	Definite scaling failure ^a		Probable scaling failure ^b		Probable scaling success ^c		Definite scaling success ^d		Probable/Definite Scale success	
	n	%	n	%	n	%	n	%	n	%
Communication n = 20	0	0	1	5.0	11	55.0	8	40.0	19	95.0
Structure n = 24	0	0	3	12.5	11	45.8	10	41.7	21	87.5
Socialization n = 24	0	0	1	4.1	5	20.8	18	75.0	23	95.8
Developmental Assessment and Planning n = 32	0	0	4	12.5	17	53.1	11	34.4	28	87.5
Behaviour Management n = 28	0	0	2	7.1	17	60.7	9	32.1	26	92.8

Notes: ^aPercentage of occasions when items correlated significantly stronger with other subscales than with their hypothesized subscale (definite scaling failure); ^bPercentage of occasion when items correlated non-significantly stronger with other subscales than with their hypothesized subscale (probable scaling failure); ^cPercentage of occasions when corrected item-total correlations exceeded item-to-other correlations but not significantly (probable scaling success); ^dPercentage of occasions when corrected item-total correlations significantly exceeded item-to-other correlation (definite scaling success)

Table 5. Scale descriptive statistics and internal consistency.

Scale (possible score range)	Md (q1–q3; min–max) ^a	Mean (SD) ^b	Cronbach's alpha	Computable scale scores (%) ^c
Total score (32–160)	125 (112–135; 67–160)	3.84 (0.51)	0.90	79.2
Communication (5–25)	19.0 (16.0–22.0; 6–25)	3.69 (0.84)	0.73	93.7
Structure (6–30)	23.0 (19.0–26.0; 6–30)	3.68 (0.80)	0.77	90.8
Socialization (6–30)	22.0 (18.0–25.0; 8–30)	3.60 (0.82)	0.81	90.8
Developmental Assessment and Planning (8–40)	31.0 (28.0–34.0; 17–40)	3.82 (0.60)	0.77	85.8
Behaviour Management (7–35)	30.0 (28.0–33.0; 16–35)	4.28 (0.53)	0.74	89.1

Notes: ^aData are summed scale median (q1–q3; min–max scores); ^bData are mean (SD) scores across items (range: 0–5); ^cProportion without missing item responses.

assessed group homes. Consequently, this increases the possibilities to detect differences between settings and changes over time (Baron et al. 2006). With large floor/ceiling effects, changes or differences outside the range covered by the scale will be undetected and only changes or differences in one direction can be detected. While floor/ceiling effects were acceptable, there was a tendency for ERS-Q scores in this study to be somewhat skewed towards the better end of therapeutic environments. This implies that the ERS-Q may be less effective at differentiating among group homes performing at the high end of the continuum compared to at the lower end. We do not consider this a major problem since it appears feasible to assume that, in general, it is the less well performing homes that are of most concern.

Table 6. Inter-correlations and internal consistencies among ERS-Q scales.^a

ERS-Q Scale	ERS-Q Scale				
	Communication	Structure	Socialization	Developmental Assessment and Planning	Behaviour Management
Communication	(0.73) ^b				
Structure	0.63	(0.77)			
Socialization	0.35	0.41	(0.81)		
Developmental Assessment and Planning	0.32	0.42	0.58	(0.77)	
Behaviour Management	0.32	0.44	0.41	0.55	(0.74)

Notes: ^aInter-correlations among scales should be substantially less than their respective alpha coefficients to support measurement of distinct constructs; ^bInternal consistency reliability (coefficient alpha) in parentheses.

In addition to targeting of the sample, i.e., the distribution of scores, the ability of scores to detect differences is also influenced by their reliabilities. Specifically, compromised reliability has an adverse impact on the sample size required to detect differences; power calculations traditionally do not take this into account (Fleiss 1986). That is, with compromised reliability, even if exceeding the minimal acceptable criteria, sample size requirements will be underestimated for a certain desired level of statistical power (Fleiss 1986). These concerns are relevant for the ERS and ERS-Q as this and other studies (Van Bourgondien et al. 1998) have shown that their reliabilities are similar but compromised (albeit considered ‘acceptable’). Given the requirement for expert on-site assessments to administer the original ERS and the maintained validity of its version (Hubel, Hagell, and Sivberg 2008), there thus appears to be a good case for the ERS-Q in future ASD studies. That is, the ERS-Q should be able to allow for considerably larger sample sizes without unreasonable funding demands.

While the present observations provide further evidence for the possibility to adapt the ERS into a questionnaire version with acceptable measurement properties, a number of important aspects were not considered in this study. In particular, the study design did not allow us to address test-retest reliability or to directly study its responsiveness. Responsiveness refers to the ability of an instrument to detect small but clinically important changes. As discussed above, while the score distributions observed here suggest that the ERS-Q should not be associated with major responsiveness problems, this needs to be assessed empirically.

In conclusion, we found general support for the measurement properties of the ERS-Q according to traditional test theory criteria, which lends support for its use in ASD residential home surveys and interventional studies. Our observations indicate that the five ERS-Q subscales appear to tap different but related aspects of a single construct, represented by its total score. Whether the total and/or subscale score profile should be used needs to be governed by the objective of the study at hand. While our observations indicate that a few items did not meet the applied criteria, we do not consider any changes to the ERS-Q to be justified at this stage. For such actions to be taken, more rigorous analyses such as Rasch analysis and confirmatory factor analysis are needed, preferably followed by replications in independent samples.

References

- Baron, R., A. Elashaal, T. Germon, and J. Hobart. 2006. Measuring outcomes in cervical spine surgery: Think twice before using the SF-36. *Spine* 31, no. 22: 2575–84.
- Cronbach, L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.
- Fleiss, J.L. 1986. *Design and analysis of clinical experiments*. New York: John Wiley.
- Hobart, J., A. Riazi, DL. Lamping, P. Fitzpatrick, and A.J. Thompson. 2004. *Improving the evaluation of therapeutic interventions in multiple sclerosis. Development of a patient-based measure of outcome*. Health Technology Assessment NHS R6D HTA Programme.
- Hubel, M, P. Hagell, and B. Sivberg. 2008. Brief report: Development and initial testing of a questionnaire version of the Environment Rating Scale (ERS) for assessment of residential programs for individuals with autism. *Journal of Autism and Development Disorders* 38, no. 6: 1178–83.
- Likert, R.A. 1937. A technique for measurement of attitudes. *Archives of Psychology* 140: 1–55.
- Lord, C., A. Wagner, S. Rogers, P. Szatmari, M. Aman, T. Charman, G. Dawson et al. 2005. Challenges in evaluating psychosocial interventions for Autism Spectrum Disorders. *Journal of Autism and Developmental Disorder* 35, no. 5: 695–711.

- McHorney, C.A., and A.R. Tarlov. 1995. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research* 4: 293–307.
- Saris-Baglama, R.N., C.J. Dewey, G.B. Chisholm, M. Koskinski, J.B. Bjorner, and J.E. Ware. 2004. *SF health outcomesTM scoring software user's guide*. Lincoln, RI: QualityMetric, Inc.
- Schopler, E. 1997. Implementation of TEACCH philosophy. In *Handbook of autism and pervasive developmental disorder*, ed. D.J. Cohen and F.R. Volkmar, 767–95. New York: Wiley.
- Streiner, D., and G. Norman. 2000. *Health measurement scales. A practical guide to their development and use*. New York: Oxford University Press.
- Van Bourgondien, M.E., N.C. Reichle, D. Cambell, and G. Mesibov. 1998. The environment rating scale (ERS): A measure of the quality of the residential environment for adults with autism. *Research in Developmental Disabilities* 19, no. 5: 381–94.
- Van Bourgondien, M.E., N.C. Reichle, and E. Schopler. 2003. Effects of a model treatment approach on adults with autism. *Journal of Autism and Developmental Disorders* 33, no. 2: 131–9.
- Ware, J.E., and E. Gandek. 1998. Methods for testing data quality, scaling assumptions, and reliability: The IQOLA project approach. *Journal of Clinical Epidemiology* 51, no. 11: 945–52.
- Ware, J.E., W.J. Harris, B. Gandek, B.W. Rogers, and P.R. Reese. 1997. *MAP-R for windows: multitrait/multiitem analysis program – revised user's guide*. Boston, MA: Health Assessment Laboratory.