Brief Report: Assessing the Impact of Rating Scale Type, Types of Items, and Age on the Measurement of School-Age Children's Self-Reported Quality of Life

Joanne Cremeens,[1]* PhD, Christine Eiser,[2] PhD, and Mark Blades,[2] PhD
[1]*Division of Behavioral Medicine, St. Jude Children's Research Hospital, and*
[2]*Department of Psychology, University of Sheffield*

Objective To examine the effect of age, scale, and item type on the reliability and reproducibility of children's quality-of-life (QOL) responses. **Methods** The TedQL (ability, social, and mood items) was administered to 266 healthy children (age range of 5–6 and 7–9 years) at two time points, comparing three rating scales (circles, faces, and thermometer). Children were given the same (n = 144) or different (n = 122) scales over time. **Results** Reliability for total QOL and ability items was highest for circles and for social items using the faces. Faces and thermometer scales showed highest reproducibility over time. Greater agreement over time across different scales was found between circles and faces (5–6 years) and thermometer and circles (7–9 years). **Conclusions** For maximum internal reliability, circles are recommended for ability items and faces for social items. For maximum reproducibility over time, the thermometer is recommended for 5–6 years and faces for 7–9 years.

Key words children; health-related quality of life; rating scales; self-report.

Children's self-reports about their physical, psychological, and social functioning are often used to help make diagnoses of adjustment and psychopathology, make treatment decisions for chronically ill populations, and guide legal decisions in courts of law (e.g., custody). Many self-report instruments have been developed for young children (Cremeens, Eiser, & Blades, in press). A range of different rating scales have been used in child measures, including Likert, graphic (e.g., three-dimensional), and facial expression scales (Cremeens et al., in press; Van Laerhoven, van der Zaag-Loonen, & Derkx 2004). Graduated circular and facial expression scales have been popular in the measures for children below 8 years (Rebok et al., 2001).

Despite the range of options, there have been few attempts to consider the agreement between children's responses across scales or to assess the impact of scale type on the accuracy or reliability of responses. The few studies that have considered the comparability of children's responses across scales have been conducted with pain ratings. Goodenough et al. (1997) reported 4- to 6-year-old children's pain ratings (for routine immunization) that were comparable across different types of pain rating scales. Chambers et al. (1999) compared venepuncture pain ratings in 5- to 12-year-old children with diabetes across five different facial expression scales. Despite high correlations among pain ratings across all the faces scales, Chambers et al. (1999) reported a significant main effect of scale type (i.e., faces with happy and sad faces vs. faces with neutral and sad faces) on pain ratings. Similar assessment issues arise when considering children's responses to quality-of-life (QOL) items.

A handful of studies have investigated scale choice for child self-report measurement. Rebok et al. (2001) found that the three-quarters of 5- to 11-year-old healthy

*Joanne Cremeens had a name change during the completion of this study from Lawford to Cremeens.

All correspondence concerning this article should be addressed to Joanne Cremeens, Division of Behavioral Medicine, St. Jude Children's Research Hospital, 332 North Lauderdale, Mail Stop No. 740, Memphis, Tennessee 38105-2794. E-mail: joanne.cremeens@stjude.org.

Journal of Pediatric Psychology 32(2) pp. 132–138, 2007
doi:10.1093/jpepsy/jsj119
Advance Access publication April 19, 2006

children preferred graduated-sized circles to a visual analogue scale (VAS). Van Laerhoven et al. (2004) reported that the children aged 6–12 years (at an outpatient clinic) preferred a Likert scale to a VAS. Although preferences are a valid way to make scale choices, preferences can be influenced by extraneous factors such as visual appeal.

Age and item type can also affect children's understanding of scales. Shields, Palermo, Powers, Grewe, and Smith (2003) found that only 42% of healthy children aged 5–7 years could use a VAS accurately, and cognitive ability (estimated intelligence quotient [IQ]) and age were the best predictors of ability to use a VAS. Additionally, Shields, Cohen, Harbeck-Weber, Powers, and Smith (2003) found that the age remained the determining factor of successful VAS use by 5- to 14-year-old children rating pain (presenting to an emergency department with a laceration requiring sutures), even when they were taught how to use the VAS. In contrast, Chambers and Johnston (2002) examined the effect of age, question type, and the number of response choices on the amount of extreme responses produced by 5- to 12-year-old healthy children. Their results showed that regardless of age, children produced more extreme responses when rating emotional items compared with items focused on physical abilities or social interactions.

Thus, further work is needed to compare children's use of a variety of scale types and to explore specifically whether the reliability and reproducibility of children's responses on specific self-report measures are influenced by the choice of rating scale. For example, Rebok et al. (2001) compared children's preferences for different response scales; however, this study did not report the comparison of the reliability of children's responses across scales. In addition, while researchers such as Chambers and Johnston (2002) found that item type influenced the amount of extreme responses in young children, further work is needed to consider whether age or item type affects the reliability and agreement (or reproducibility) of children's responses across scale types.

Such studies are necessary for judging the most appropriate scale for specific child measures, item types, and/or age ranges. To address this, we investigated the impact of scale type (bipolar circles, Harter & Pike, 1984; bipolar faces, Christie, French, Sowden, & West, 1993; linear thermometer, Szyfelbein, Osgood, & Carr, 1985), age (range 5–6 and 7–9 years), and types of items (i.e., ability, social, and mood items) on the reliability, agreement over time, and accuracy of children's responses to items from a generic QOL measure (the TedQL; Lawford, Volavka, & Eiser, 2001).

On the basis of the work of Rebok et al. (2001) and Shields, Palermo, et al. (2003), we expected that the responses for the circles and faces scales would show higher reliability compared with the responses for the thermometer. In addition, we predicted that the agreement (over time) would be influenced by scale type (i.e., higher agreement over time for the circles and faces scales compared with the thermometer scale). Furthermore, following Chambers and Johnston (2002), we predicted that the reliability of responses across scales would be influenced by item type (i.e., higher reliability for physical/social items over mood items).

Methods

Participants

Participants were 266 healthy children (138 girls and 128 boys) between the ages of 5 and 9 years (M = 7.14, SD = 1.07) recruited from U.K. schools. Children were excluded if they were receiving any treatment for a chronic or an acute medical condition and if they had prior diagnoses of special needs or learning difficulties. Children were stratified into two age categories: 5–6 years (n = 130) and 7–9 years (n = 136). Two hundred and forty-seven were Caucasian, six were of Afro-Caribbean origin, and 13 were of Asian origin. They were randomly assigned to either the same rating scale (n = 144; circles n = 49, faces n = 47, and thermometer n = 48) or different scales at time 1 and time 2 (n = 122; circles and faces n = 46, faces and thermometer n = 38, and circles and thermometer n = 38). The order of scale presentation was counterbalanced, and the assignment to scale type was balanced within age groups. Ethics approval was given by the appropriate committee. Written consent from parents and verbal assent from children were obtained.

Procedure

Children were interviewed individually in a room separate to the main classroom. Two identical teddy bears (40 cm high, differentiated by name badges) were placed opposite the children with verbal instructions: "I have two teddy bears: Iggy and Ziggy. They look the same, but they like doing different things. I am going to tell you what Iggy and Ziggy like to do, and then I am going to ask you what you like to do." Children were shown how to answer using three 4-point rating scales (circles bipolar scale using two small and two large circles; faces bipolar scale using four cartoon faces ranging from happy to sad; and thermometer-shaped continuous scale). Children were given practice items to familiarize

them with the rating task and taught to use a "don't know" (DK) option (a question mark) if they did not know the answer or did not understand the question.

Once children demonstrated understanding of the task and rating scales, they were given the TedQL items using the teddy bears. Children were asked to respond in relation to their activities during the last week. The bears were referred to as female or male depending on child gender and counterbalanced to avoid one being seen as particularly "good" or "bad." Children were retested after 1 week. (Four children were excluded because of teacher's indications of family problems between the two time points, reducing the sample to $N = 266$.)

TedQL Measure

This generic child self-report measure includes 30 items (12 ability items including climbing, running, writing, and drawing; 12 social items about how they get on with family/friends; and six mood items about feelings/mood). Full details of the TedQL items are available from the author. The specific content of the TedQL was guided by a review of 53 existing child self-report measures (QOL, self-esteem/concept, and mental health instruments; Cremeens et al., in press) and child interview data, including abilities identified as important to QOL (Lawford et al., 2001). The internal reliability of the TedQL with healthy populations is moderate to good (.60–.75), and there is some evidence for face and construct validity (Lawford et al., 2001).

A forced-choice recognition task was used where the bears are described, then children chose "which bear was most like them" (actual self), and then probed for whether they were "really like this" or "just a little bit." Children were then asked: "what they would like to be like" if they could change how they were at any given activity (ideal self).

Scoring and Statistical Analyses

Children's responses were scored from 0 to +3 for both "actual" and "ideal" QOL, with higher scores representing higher QOL. Actual and ideal scores were used to compute discrepancy scores for each item in the measure. These scores were calculated in the same way as existing discrepancy-based QOL measures (i.e., ideal minus actual QOL scores; Collier, Mackinlay, & Phillips, 2000). For the purposes of the study reported here, actual scores were used in the analyses. The internal reliability (Cronbach's α coefficients) and the agreement (intraclass ρ_1 correlation coefficients) between responses using the same and different scales were compared across item types (ability/social/mood) and age groups. We used a standard of .70 for Cronbach's coefficients (Scientific Advisory Committee of the Medical Outcomes Trust, 2002) and for intraclass coefficients (Ottenbacher, 1995). The mean percentages of DK responses were compared across rating scales and item types.

Results

Internal Reliability Across Rating Scales by Types of Items and Age

Internal consistency coefficients across scales by age and item type are presented in Table I. The total TedQL scores exceeded .70 for children using the circles (.80) and faces (.75) scales. By age, 5- to 6-year-olds' responses showed good reliability (above .70) across all scale types and 7- to 9-year-olds showed good reliability using the circles scale (.78). By item type, there was good reliability for ability items for the circles scale for the total sample (.72) and 5- to 6-year-olds (.85). For social items, there was good reliability across the total sample and both age groups for the faces scale (total sample: .70; 5–6 years: .72; and 7–9 years: .70). Children's responses to the mood items were below .70 across all scales and ages.

Agreement Between Responses Across Rating Scales by Types of Items and Age

Table II presents the intraclass correlation coefficients across scales by types of items and age, for children

Table I. Internal Reliability Coefficients[a] across Rating Scales by Types of Items and Age

	Rating scale type					
	Circles		Faces		Thermometer	
	n	α	*n*	α	*n*	α
Total scale (30 items)						
Total sample	49	0.80	47	0.75	48	0.67
5–6 years	23	0.77	21	0.76	22	0.75
7–9 years	26	0.78	26	0.60	26	0.57
Ability items (12 items)						
Total sample	49	0.72	47	0.67	48	0.56
5–6 years	23	0.85	21	0.51	22	0.57
7–9 years	26	0.59	26	0.58	26	0.57
Social items (12 items)						
Total sample	49	0.45	47	0.70	48	0.56
5–6 years	23	0.47	21	0.72	22	0.64
7–9 years	26	0.60	26	0.70	26	0.41
Mood items (six items)						
Total sample	49	0.33	47	0.24	48	0.20
5–6 years	23	0.39	21	0.25	22	0.18
7–9 years	26	0.25	26	0.20	26	0.35

[a]Cronbach's α, cut-off values >.70 (shaded).

Table II. Correlation Coefficients[a] between Responses using the Same or Different Rating Scales by Types of Items and Age

	Group[b]											
	C-C		F-F		T-T		C-F		F-T		T-C	
	n	ρ_I	*n*	ρ_I	*n*	ρ_I	*n*	ρ_I	*n*	ρ_I	*n*	ρ_I
Total scale												
Total sample	49	0.58*	47	0.75*	48	0.70*	46	0.76*	38	0.40**	38	0.74*
5–6 years	23	0.53*	21	0.70*	26	0.78*	23	0.96*	21	0.33	20	0.70*
7–9 years	26	0.62*	26	0.77*	26	0.62*	23	0.37***	17	0.42**	18	0.86*
Ability items												
Total sample	49	0.59*	47	0.70*	48	0.66*	46	0.84*	38	0.44*	38	0.70*
5–6 years	23	0.60*	21	0.60**	26	0.55**	23	0.93*	21	0.39***	20	0.54**
7–9 years	26	0.50**	26	0.66*	26	0.72*	23	0.76*	17	0.48***	18	0.89*
Social items												
Total sample	49	0.52*	47	0.75*	48	0.76*	46	0.69*	38	0.5*	38	0.75*
5–6 years	23	0.45**	21	0.68*	26	0.86*	23	0.97*	21	0.45***	20	0.66*
7–9 years	26	0.58*	26	0.81*	26	0.64*	23	0.28	17	0.65*	18	0.86*
Mood items												
Total sample	49	0.47*	47	0.49*	48	0.46*	46	0.30**	38	0.44**	38	0.26
5–6 years	23	0.57**	21	0.45**	26	0.40**	23	0.82*	21	0.48**	20	0.20
7–9 years	26	0.41***	26	0.55**	26	0.50**	23	0.20	17	0.42***	18	0.82*

[a]Intraclass ρ_I cut-off values >.70 (shaded).

[b]C-C, F-F, T-T circles/faces/thermometer scales at both times, C-F circles and faces scales, F-T faces and thermometer scales, and T-C thermometer and faces scale over both times.

*$p < .001$. **$p < .01$. ***$p < .05$.

given the same (n = 144) or different (n = 122) scales over time.

Agreement Between Responses Using the Same Scales Over Time

For the total scale, there was good agreement over time (above .70) for the faces and thermometer scales for the total sample and 5- to 6-year-olds. By item type, there was good agreement for the faces (total sample) and for the thermometer scales (7–9 years) for ability items. For social items, there was good agreement for the faces (total sample, 7–9 years) and for the thermometer scales (total sample, 5–6 years). For the mood items, agreement over time was below .70 across all scale types and both age groups. Wilcoxon median testing revealed no significant differences between responses over time when using the same scales, across the total sample, both age groups, and all item types.

Agreement Between Responses Using Different Scales Over Time

For the total scale, agreement was above .70 between the circles and faces and between the thermometer and circles. By age, agreement for 5- to 6-year-olds was above .70 between the circles and faces scales across the total scale and by item types. Agreement for 7–9 years olds was above .70 between the thermometer and circles across the total scale and by item types. Wilcoxon median testing revealed significant differences between ratings on the thermometer and the circles for 5- to 6-year-olds across the total scale, ability, social, and mood items ($p < .05$).

Percentage of DK Responses to Items Across Rating Scale Type

The percentages of DK responses were low across all rating scales, across both age groups, and at both time points (circles: 1.8–2.7%; faces: 1.7–2.5%; and thermometer: 1.2–2.2%) and were not significantly different across scale type or age. The DK responses were evenly spread across items and showed similar patterns across all rating scales.

Discussion

The growing interest and need for valid and reliable child self-report instruments have raised questions about the appropriateness of different rating scales for different measures, ages, and types of items (De Civita et al., 2005). This study contributes to the literature by examining the impact of three rating scales on the reliability of and agreement between children's responses to items in a QOL measure.

Consistent with our predictions, scale type was shown to affect the reliability of children's responses, with responses across the total scale showing the highest

reliability for the circles and the faces scales. However, the impact of scale type showed differences by age, where 7- to 9-year-olds' responses using the circles scale showed highest reliability; however, 5- to 6-year-olds' responses showed similar levels of reliability across all three scales.

The impact of scale type was also influenced by the types of items being rated. Responses to ability items showed higher reliability when the circles scale was used, and responses to social items had higher reliability when the faces scale was used. The reliability of responses to the mood items was below the .70 standard across all scales and ages. This lowered reliability may have been due to there being fewer mood items in the TedQL (n = 6) than ability (n = 12) and social items (n = 12). Our results are consistent with Chambers and Johnston's (2002) findings that the children respond differently to rating physical aspects compared with emotional states.

Scale type, age group, and item type affected the levels of agreement between children's responses using the same scales over time. Overall, the highest agreement was found for the faces and thermometer scales across the total scale. However, agreement was highest for younger children (5–6 years) using the thermometer scale, compared with the faces scale for older children (7–9 years). We had not expected either of the age groups to perform best with the thermometer scale; nonetheless, it was clear in this study that the younger children were more reliable when using the thermometer scale, and this result could be investigated further in other studies to find out whether such a graduated scale is particularly beneficial to younger children.

Furthermore, by item type, agreement for ability items was highest using the faces scale, compared with both the faces and thermometer scales for social items. Some age effects were also shown here, where for ability items agreement was highest for older children using the thermometer scale. In addition for social items, agreement was highest for older children using the faces scale and younger children using the thermometer scale. These age effects on agreement over time seem to show that 5- to 6-year-olds provide more reproducible responses over time when using a continuous (i.e., thermometer) rating scale, compared with a bipolar (i.e., faces) rating scale for 7- to 9-year-olds.

The levels of agreement between responses using different scales over time were influenced by age. For the total scale, agreement was highest between the circles and faces and between the circles and thermometer scales. However, for younger children, the agreement between the circles and faces scales was highest, compared with the circles and thermometer scales for older children. This age effect on agreement over time remained consistent regardless of item type. Our results show that younger children used the circles and faces scales in similar ways (i.e., as bipolar scales), compared with older children who used the circles scale in a similar way to the thermometer scale (i.e., seeing both scales as continuous).

Understanding of items can affect the reliability of children's responses. The percentage of DK responses was generally low across all three scales, and we take this to mean that the children understood and could answer the questions irrespective of scale type.

Only few researchers have compared different scales before (e.g., Chambers et al., 1999) and reported that different scales can have an effect on children's responses. Our results confirm these earlier findings because we also found an effect of scale type. Investigating reliability of responses to any given measure or scale involves repeated questioning. We know from the wider developmental research into the effects of repeated questions that the children who are questioned more than once often change their responses, and this is particularly the case when children are asked questions that require an opinion rather than a factual answer (Krähenbuhl & Blades, in press-a). Therefore, our results correspond to the a wider developmental research on repeated questions and confirm that young children may often change their responses (Krähenbuhl & Blades, in press-b). In the repeated questioning research, children have always been asked for verbal responses, but we asked the children for a response on a scale. The fact that the children were not always consistent (when using the same or different scales at the second questioning) extends the repeated questioning research by demonstrating that, irrespective of response mode, the children will alter their answers. This raises the issue of why children do not give consistent responses to the same question. The nature of questioning children more than once about the items in the TedQL may have made them reconsider their answers to items when they were questioned for the second time, or they may have intended to give the same response the second time but may have been unable to recall their first response accurately. These possibilities could be investigated empirically in future studies.

This study was undertaken as part of a series of studies to develop the TedQL and is therefore limited by its use of a relatively new QOL measure. Further work could also be done using other QOL measures (e.g., PedsQL™; Varni, Seid, & Kurtin, 2001) and other scales (e.g., Likert and VAS) to find out whether the effects we have

reported are consistent across measures and scales. In addition, this study was conducted with healthy children, and further work with populations coping with chronic conditions is needed to investigate the measurement issues addressed here.

The results of this research have implications for the measurement of children's self-reports and the design of new child measures. This study showed that the reliability of children's responses can be affected by scale type and specifically that specific rating scale types may be appropriate for different types of items. For example, on the basis of this study, we would recommend that circles (of two sizes) be used with children when rating items asking about physical abilities and facial expression scales be used with social items asking about friends and family. Our results also highlighted age effects in the use of rating scales, where younger children produce more highly reproducible responses over time using a continuous scale (i.e., a thermometer) compared with older children using a bipolar scale (i.e., happy and sad faces). The discovery that different scales had different effects on the children's responses means that if previous studies had used alternative scales, their results might not have been the same. Therefore, there is a need for further research to explore the extent to which different scales affect children's judgments about the quality of their lives.

Acknowledgments

This study was funded in part by a University of Sheffield grant awarded to the second author.

Received July 11, 2005; revision received October 27, 2005, February 14 2006, March 2, 2006, and March 22, 2006; accepted March 26, 2006

References

Chambers, C. T., Giesbrecht, K., Craig, K. D., Bennett, S. M., & Huntsman, E. (1999). A comparison of faces scales for the measurement of pediatric pain. *Pain*, *83*, 25–35.

Chambers, C. T., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology*, *27*, 27–36.

Christie, M. J., French, D., Sowden, A., & West, A. (1993). Development of child-centred disease-specific questionnaire for children living with asthma. *Psychosomatic Medicine*, *55*, 514–518.

Collier, J., Mackinlay, D., & Phillips, D. (2000). Norm values for the generic children's quality of life measure from a large school-based sample. *Quality of Life Research*, *9*, 617–623.

Cremeens, J., Eiser, C., & Blades, M. (in press). Characteristics of health-related self-report measures for children aged three to eight years: a review. *Quality of Life Research*.

De Civita, M., Regier, D., Alamgir, A. H., Anis, A. H., Fitzgerald, M. J., & Marra, C. A. (2005). Evaluating health-related quality-of-life studies in paediatric populations. *Pharmacoeconomics*, *23*, 659–685.

Goodenough, B., Addicoat, L., Champion, G. D., McInerney, M., Young, B., Juniper, K., et al. (1997). Pain in 4- to 6-year old children receiving intromuscular injections: a comparison of the faces pain scale with other self-report and behavioral measures. *The Clinical Journal of Pain*, *13*, 60–73.

Harter, S., & Pike, K. (1984). The pictorial scale of perceived competence and social acceptance for young children. *Child Development*, *48*, 80–87.

Krähenbuhl, S., & Blades, M. (2006). The effect of question repetition within interviews on young children's eyewitness recall. *Journal of Experimental Child Psychology*, *23*.

Krähenbuhl, S., & Blades, M. (in press-b). The effect of interviewing techniques of young children's responses to questions. *Child: Care, Health and Development*.

Lawford, J., Volavka, N., & Eiser, C. (2001). A generic measure of quality of life for children aged three to eight years. *Pediatric Rehabilitation*, *4*, 197–207.

Ottenbacher, K. J. (1995). An examination of reliability in developmental research. *Developmental and Behavioral Pediatrics*, *16*, 177–182.

Rebok, G., Riley, A., Forrest, C., Starfield, B., Green, B., Robertson, J., et al. (2001). Elementary school-aged children's reports of their health: a cognitive interviewing study. *Quality of Life Research*, *10*, 59–70.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: review criteria. *Quality of Life Research*, *11*, 193–205.

Shields, B. J., Cohen, D. M., Harbeck-Weber, C., Powers, J. D., & Smith, G. A. (2003). Pediatric pain measurement using a visual analogue scale: a comparison of two teaching methods. *Clinical Pediatrics*, *42*, 227–234.

Shields, B. J., Palermo, T. M., Powers, J. D., Grewe, D., & Smith, G. A. (2003). Predictors of a child's ability to use a visual analogue scale. *Child: Care, Health and Development*, *29*, 281–290.

Szyfelbein, S. K., Osgood, P. F., & Carr, D. B. (1985). The assessment of pain and plasma b-endorphin immunoactivity in burned children. *Pain*, *22*, 173–182.

Van Laerhoven, H., van der Zaag-Loonen, H. J., & Derkx, B. H. F. (2004). A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatrica*, *93*, 830–835.

Varni, J. W., Seid, M., & Kurtin, P. S. (2001). Reliability and validity of the paediatric quality of life inventory version 4.0. Generic core scales in healthy and patient populations. *Medical Care*, *39*, 800–812.