

Validity Issues in the Use of Pictorial Likert Scales

Laura Reynolds-Keefer, University of Michigan-Dearborn, lrkeefer@umd.umich.edu

Robert Johnson, University of South Carolina, RJOHNSON@sc.edu

Tammie Dickenson, University of South Carolina, TDICKENSEN@sc.edu

Laura McFadden, University of South Carolina, RLJOHNSON@sc.edu

Abstract

Likert scale assessments using pictures as anchor (e.g., smiley faces) are often used in the assessment of young children or non-readers; however, teachers and researchers seldom consider the potential importance of the type of picture selected. Because data from such assessments may inform instruction or serve as the basis for research findings, validity issues surrounding the importance of the picture selected may be informative. In this study, 136 first and third grade elementary school students were randomly assigned to 1 of 3 versions of a short reading attitudinal instrument with response scales indicated by differing pictures (i.e., sad/happy face, clouds/sun, and NO/YES). An ANOVA analysis was used to explore validity type of picture, grade, and gender. The findings of this study indicated that Likert scales using different pictures did not show variability in student responses.

This article has been peer-reviewed and accepted for publication in *SLEID*, an international journal of scholarship and research that supports emerging scholars and the development of evidence-based practice in education.

© Copyright of articles is retained by authors. As an open access journal, articles are free to use, with proper attribution, in educational and other non-commercial settings.
ISSN 1832-2050

Introduction

As noted by McMillan (2000), the skill of classroom teachers and administrators regarding assessment practices and principles has become an area of concern (Cizek, 1997; McMillan, 2001). An important part of the creation of quality assessments is an understanding of the principle of validity and the importance of validity in making accurate inferences (McMillan 2000). One type of assessment that has become a favourite of classroom teachers is the Likert-style survey using pictures in place of text in communicating levels of choice, with the most common set of pictures being the smiley face: ☹ ☺ ☺. Teachers and administrators often select an existing Likert instrument with pictures, add pictures to an existing instrument using clip art, or draw the images by hand. The preference for a specific image, however, could introduce a systematic bias that results in inaccurate measurement of a construct. In an effort to explore this issue, this study considered validity issues relating to the use of Likert scale assessments with pictures or images to depict selection choice.

Pictures in Likert Scales

Likert assessments that have pictures associated with the choices are created by teachers but are also commercially available. New and popularised images are now commonly used to represent scale options, underscoring the need to consider the impact of such images. *The Elementary Reading Attitude Survey* (McKenna & Kear, 1990) and the *Writing Attitude Survey* (WAS) (Kear, Coffman, McKenna, & Ambrosio, 2000) both use images of the cartoon character Garfield. Each scale asks children to select the picture of Garfield that most closely matches how they “feel” regarding reading or writing. The scale depicts Garfield in states that range from “angry” to “happy.” Neither study, however, investigates whether the pictures influence student responses and systematically bias the scores.

Clinicians in health care also use Likert scales with pictures, such as the *Faces Pain Scale* (FPS). The FPS combines pictures with a Likert scales to assess the level of pain experienced by young children. Chambers and Craig (1998) varied the images used in the FPS and reported that children’s ratings of pain varied according to the characteristics of the images. More specifically, the ratings of children varied with the type of expression appearing on the human faces. This highlights the potential role of picture selection in relation to the accuracy of the resulting data.

The current assessment environment, prevalence of high-stakes testing, and the emphasis on data-driven decision making, increases the importance of creating high quality assessment tools. Preparing teachers and administrators to create and use high quality assessments has become increasingly important, as has examining the validity of the types of instruments popular in classrooms (McMillan, 2000). This study considers whether the results of the Likert-style assessments using pictures as categorical descriptors vary based on the type of image used. This study hopes to begin a discussion regarding the utility and adequacy of this type of assessment.

Methodology

To explore the impact of different anchors in Likert assessments, a convenience sample of 69 first grade students and 67 third grade students at a suburban elementary school in the South-eastern United States was undertaken. Participating students completed reading attitude assessments differing only in the three pictorial anchors assigned to the response scale (see Figure 1). First and third grades were used in this sample in order to consider effects due to age (Wing & Cheng, 2000). Four classrooms participated at each grade level, and the assessments were completed in one day. Students absent on the day of the assessment were eliminated from the sample. Each student was randomly assigned to one of the three versions of the instrument: smile, sun, or text. Children completed the instruments in groups with other children assigned to the same version of the assessment.

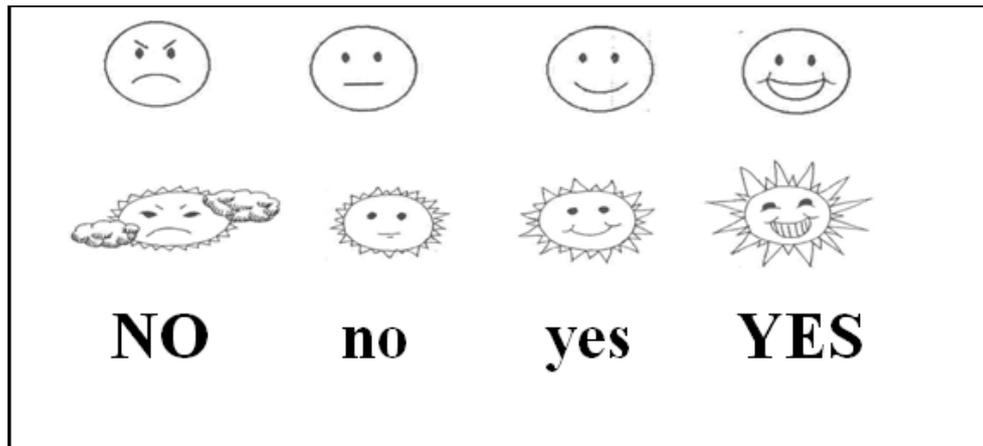


Figure 1. Three separate scales used on three versions of the survey. .

The survey consisted of 11 activities relating to students' attitude toward reading, and asked the students to what extent they thought these statements applied to their own feeling about reading. The school district was given the opportunity to assist in item development to tailor the survey to their program evaluation needs regarding reading. The questions formulated by the school district were integrated into the instrument used in this study. The questions focused on students' "like" or "dislike" of reading both at school and at home (Stahl & Yaden, 2004). Responses were recorded as a "1" for the lowest response (i.e., a frown face, cloud-covered sun, or NO) to a "4" for the highest response (i.e., a smiley face, a bright sun, or YES). A total score for each student was generated by summing his or her responses.

The eleven statements describe activities related to reading, such as, "I like to read at home." The same 11 statements appear in the three versions of the instrument. Figure 1 illustrates the three different scales, each scale appearing on a separate version of the instrument. As shown in Figure 1, the first version of the instrument the response scale uses "unhappy – happy" faces as scale anchors, the second instrument uses "sunshine – clouds" as anchors, and the third instrument uses the words "Yes" and "No" in large and small font as anchors. The version of the scale using words (i.e. "Yes" and "No") were used to explore the merit of using images instead of the more standard text.

Before the administration of each instrument to each group of participants, I read a detailed set of instructions to the students. The intended meaning of the pictorial options was explained in the directions (e.g., AERA, APA, & NCME, 1996; Kear, Coffman, McKenna, & Ambrosio). For example, for the group using the instrument with the image of the sun, students were told that if they disagreed with a statement, they would select the picture of the sun with clouds. If they agreed with the statement a great deal, they should select the sun figure on the far right with the largest smile. After completing the instructions, each item in the instrument was read to the students to facilitate participation of non-readers. If students did not understand an item that described a reading activity, it was read to them again (e.g., AERA, APA, & NCME, 1996).

Results

As shown in Table 1, the means and standard deviations of students for the three instruments were similar, and means fell between 32.3 and 34.3 with a total mean

of 33.6. Although the text condition shows the lowest mean, it differs from the highest mean (i.e., the Smile version) by approximately 2 points of a 44-point scale. Cronbach’s alpha for the reading attitudinal instrument was .72. Alpha for the smile version was .78, for the sun version was .76, and for the text version (i.e., YES/NO) was .55.

Table 1 Mean Scores for Survey Types

Condition	Mean	N	SD
Smile	34.3	48	6.4
Sun	34.2	43	5.4
Text	32.3	43	4.9
Total	33.6	134	5.7

When student performance on the three scales is examined descriptively by grade level, some differences can be seen. Table 2 and Figure 2 display the mean survey scores for each picture type by grade level. The biggest difference between grade levels occurs for the smile picture (2.8) and the smallest difference occurs for the text (Yes/No) (1.1). As shown in Table 2 and Figure 2, first grade students had the highest mean score on the scale with the smile pictures (M=35.7) and the lowest mean on the scale with the text (Yes/No)(M=33.1). The third grade scale with sun pictures resulted in the highest mean (M=35.3) and the lowest mean was associated with the text (Yes/“No”) (M=31.7).

Table 2 Mean of 1st and 3rd Grade Groups by Picture Type

	Smile		Sun		Text		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Grade 1	35.7	6.6	33.1	6.5	32.8	5.8	34.0	6.4
Grade 3	32.9	6.0	35.3	3.6	31.7	4.0	33.3	4.9

N=134

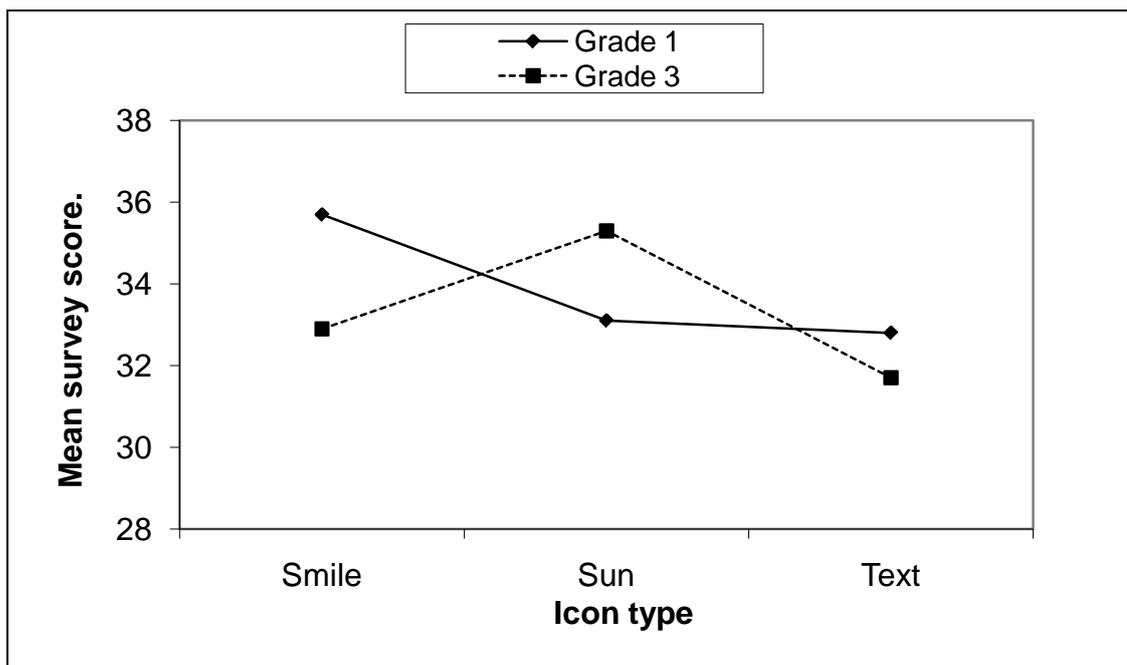


Figure 2. Plot displaying the sample means

Table 3 and Figure 3 display statistics on survey scores by picture type and gender. A comparison of means of boys and girls under the three instrument conditions shows a difference in total means of 1.1, and a difference in standard deviations of 1.2. Male and female mean scores on the smile instrument were within .2 of a point with each other, and standard deviations were the same. The means of boys and girls taking the sun instrument showed a 1.5 difference in mean scores, as well as a substantial difference in standard deviations. Girls taking the sun instrument had a standard deviation of 3.4, whereas, boys had a standard deviation of 7.1. The mean score of girls using the text scales (i.e., NO-YES) was 2.3 higher than boys, with both groups having a standard deviation of 4.9. Figure 3 shows that the mean survey scores difference between boys and female were least for the smile picture and greatest for the text (Yes/No).

Table 3 Mean of Boys and Girls by Picture Type

	Smile		Sun		Text		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Boys	34.2	6.5	33.4	7.1	30.9	4.9	33.0	6.3
Girls	34.4	6.5	34.9	3.4	33.2	4.9	34.1	5.1

N=134

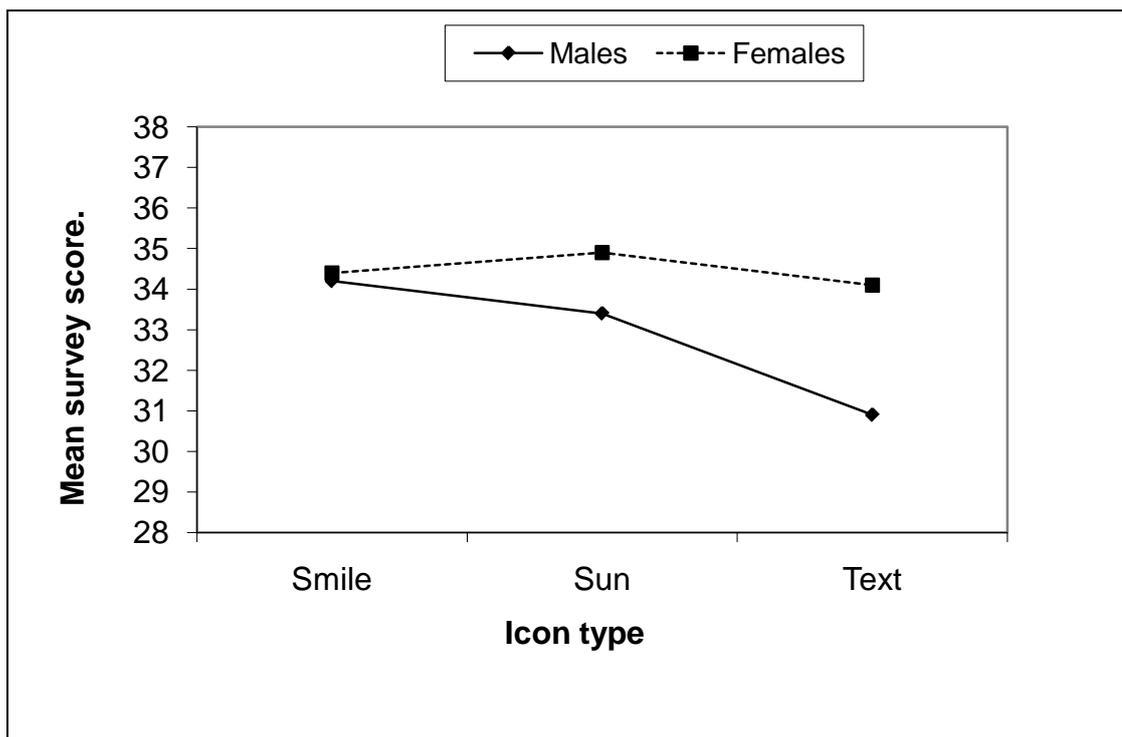


Figure 3. Plot displaying the sample means by picture type and gender

When we examined the potential influence of gender and grade level on the validity of children’s responses to the items (see Table 4), we found that first grade boys and girls had the greatest difference in means with the sun condition (5.4 points), and third grade boys and girls were the most different in scores on text instrument (3.7 points). Female means did not differ greatly between grades, with differences of 2.2, 1.4, and 0.4 for the smile, sun, and text conditions, respectively. Male students, however, had a notably higher mean for the sun condition in the third grade (difference of 6.7 points), and a lower mean under the text condition (difference of 2.7 points).

Table 4 Mean Scores for Survey Types, Gender, and Grade Level

	Smile		Sun		Text		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
<u>Grade 1</u>								
Boys	35.7	4.3	30.2	8.0	32.1	5.2	33.0	6.1
Girls	35.7	9.3	35.6	3.8	33.5	6.5	34.9	6.6
<u>Grade 3</u>								
Boys	31.9	8.6	36.9	4.0	29.4	4.2	32.9	6.6
Girls	33.5	3.9	34.2	3.0	33.1	3.4	33.5	3.4

N=134

The descriptive statistics suggest that first and third grade boys react differently to the sun pictures than to the other two picture types. Figure 4 shows that boys had greater grade level differences in mean scores than girls, particularly for the sun picture. As seen in Figure 4, the mean survey score on the sun picture scale is 6.7 points lower for first grade boys than for third grade boys. The order is reversed for the smile and text (Yes/No) scales with first grade boys having a mean score 3.8 points higher for the smile pictures and 2.7 points higher for the text (Yes/No) as compared with the third grade boys. The mean survey score for each picture type is higher for first grade girls than for third grade girls.

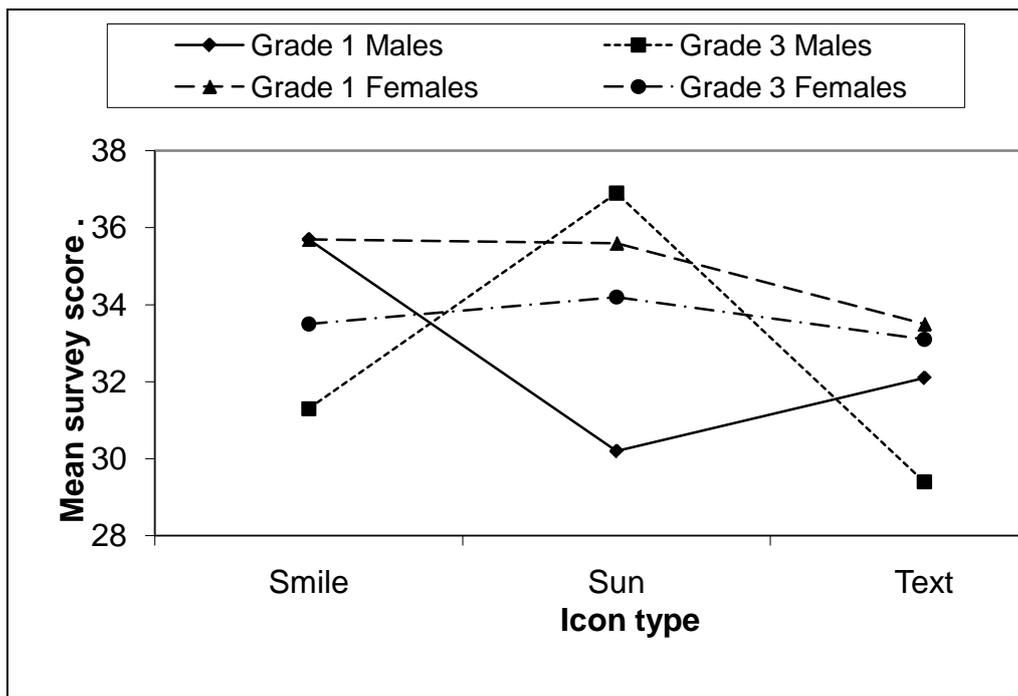


Figure 4. Plot of means for the three types of pictures by grade and gender combination

We performed a three-way analysis of variance (ANOVA) significance test using the survey score as the dependent variable and picture type (i.e., pictorial anchor), grade, and gender as the independent variables. The three-way ANOVA test determines if there are significant differences among the means of the individual independent variables, interactions between each pair of independent variables, and interaction among all three independent variables. As shown in Table 5, among the individual independent variables, there were no significant differences in mean survey scores. There was a significant two-way interaction between picture type and grade ($p=0.049$) and the three-way interaction among picture type, grade, and gender was not significant at the .05 level ($p=0.055$). The p -value for the two-way interaction between grade level and picture type gives evidence that the mean score differs depending on the combination of grade level and picture type.

Table 5 Summary Table for Three-Way Analysis of Variance

Source	<i>df</i>	<i>F</i>	<i>P</i>	<i>Partial Eta Squared</i>
Picture	2	2.239	.111	.035
Grade	1	.447	.505	.004
Gender	1	2.538	.114	.020
Picture * Grade	2	3.082	.049	.048
Picture * Gender	2	.280	.756	.005
Grade * Gender	1	.513	.475	.004
Picture * Grade * Gender	2	2.972	.055	.046
Error	122			
Total	134			

A measure of effect size, partial eta squared (η^2), was also calculated as a means of determining the magnitude of the results found in the present study. The results from this analysis are included in Table 5. Partial Eta squared represents the fraction of the total variation due to the factor of interest. An Eta-squared of .05 indicates the two-way interaction of Picture \times Grade and the three-way interaction of Picture \times Grade \times Gender have a minor effect on scores.

Because the interaction between grade level and picture type was found to be statistically significant, post-hoc analyses were conducted on contrasts involving those interaction terms. Confidence intervals were computed for six contrasts to determine which, if any, had significant differences between mean scores. The upper and lower limits of these intervals are displayed in Table 6. These contrasts include the mean difference in scores between smile and sun pictures, smile and text (Yes/No), and sun and text (Yes/No) for each grade level (grade 1 and grade 3). The Bonferroni method was used to control the Type I error rate at .05 for all six comparisons. The confidence intervals indicate whether there are significant pair wise differences in mean scores between picture types at each grade level.

Table 6 Confidence Intervals for the Difference in Mean Scores between Grades 1 and 3 by Picture Type

Gender	Picture type differences	Lower Limit	Upper Limit
Grade 1	Smile – Sun	-1.4	6.5
	Smile – Text	-1.1	6.9
	Sun – Text	-3.8	4.4
Grade 3	Smile – Sun	-6.5	1.6
	Smile – Text	-2.8	5.1
	Sun – Text	-0.5	7.7

Zero is contained in all confidence intervals, which indicates there were no statistically significant differences between the means considered in this set of contrasts. When interpreting these confidence intervals, the positive upper limit indicates how much greater the mean score for the first type of picture could be than the second type. The negative lower limit indicates how much greater the mean score for the second type of picture could be than the first. With 95% confidence across all comparisons, there is no evidence that the mean scores differ

between sun and text (Yes/No) for grade 3 students. If the mean score for grade 3 students is higher with the sun picture, it would be by at most 7.7 points higher. If the mean score for grade 3 students is higher with the text (Yes/No), it would be by at most 0.5 points higher. While this interval includes zero, the relatively large value of the upper limit compared with that of the lower limit is suggestive that the sun picture may have a greater mean than the text (Yes/No) among third graders. Further investigation with a larger sample would provide more power to detect whether the difference observed in this study exists in the population and in regards to gender.

Implications and Significance

The utility of Likert scales with pictures in classroom assessment relies on their ability to accurately reflect the attitudes of children toward the construct being measured. It appears that for first grade students and third grade students that scale pictures and gender contribute little to mean differences in scores on an attitudinal scale or survey. Although the variables of picture type and grade level are significant when involved with a two-way interaction, the actual effect size was minor. This study offers some evidence that the picture used in Likert scales might not affect the measurement of attitudes of young children. If this finding holds in future studies, then teachers may use pictorial and text-based scales interchangeably. However, additional studies with different pictures and on larger samples should be conducted to see if these findings generalize to other uses of pictorial scales. This study serves only to begin a discussion regarding the use of this type of assessment tool.

Future studies should also examine whether the pictorial scales are interchangeable across ethnic and English language learner groups. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) note that, whereas, differences in scores on an instrument might reflect differences in groups on the construct, such as reading attitude, score differences for subgroups might reflect extraneous variables and require study. The *Standards* recommend the conduct of validity studies for each subgroup.

References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Cizek, G.J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G.D. Phye (Ed.), *Handbook of classroom assessment: Learning, adjustment, and achievement* (pp. 1-32). New York, NY: Academic Press.
- Chambers, C. T., & Craig, K. D. (1998). An intrusive impact of anchors in children's faces pain scales. *Pain*, 78, 27-37.
- Chan, J. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, 51, 531-540.

- Griffin, R. S., & Gross, A. M. (2004). Childhood bullying: Current empirical findings and future directions for research. *Aggression and Violent Behavior, 9*, 379-400.
- Kear, D., Coffman, G., McKenna, M., & Ambrosio, A. (2000). Measuring attitude toward writing: A new tool for teachers. *The Reading Teacher, 54*, 10-23.
- Lawford, J., Volavka, N., & Eiser, C. (2001). A generic measure of quality of life for children aged 3–8 years: Results of two preliminary studies. *Pediatric Rehabilitation, 4*, 197–207.
- McKenna, M., & Kear, D. (1999). Measuring attitude toward reading: A new tool for teachers. In S. Barrentine (Ed.), *Reading assessment: Principles and practices for elementary teachers* (pp. 199–214). Newark, DE: International Reading Association.
- McMillan, J.H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation, 7*(8). Retrieved from <http://PAREonline.net/getvn.asp?v=7&n=8>
- McMillan, J.H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Press.
- Rigby, K. (1997). *The peer relations assessment questionnaire*. Point Lonsdale: Professional Reading Guide.
- Rigby, K., & Slee, P. T. (1993). *The peer relations questionnaire*. Point Lonsdale: Professional Reading Guide.
- Stahl, S., & Yaden, D. (2004). The development of literacy in preschool and primary grades: Work by the Center for the Improvement of Early Reading Achievement. *The Elementary School Journal, 105*, 141-165.
- Weng, L. J., & Cheng, C. P. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement, 60*, 908-924.