

Improving the Likert Scale of the Children's Social Skills Test by Means of Rasch Model

Daniel Bartholomeu, Marjorie Cristina Rocha da Silva, José Maria Montiel

Department of Educational Psychology, UNIFIEO, Osasco, Brasil
Email: d_bartholomeu@yahoo.com.br

Received 5 April 2016; accepted 11 June 2016; published 14 June 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This study addressed the psychometric properties of the Children's Social Skills Test (THAS-C). Results were divided into two parts. Initial studies investigated 257 children attending Grades 2 to 4 in public schools of two cities in the countryside of the state of São Paulo. Study 2 assessed 1381 children and adolescents aged 7 to 15 years from public and private schools in cities located in the countryside of the state of São Paulo. Infit and outfit indexes showed adequate fit for both items and people. A progression analysis using Rasch measurement model showed greater people's ability in function of response categories. On the other hand, the outfit index suggested good fit in all categories. These data show relevance for the structure of three analysis categories. Few items showed DIF. Precision data obtained with these two methods showed satisfactory indexes. These data enable validity evidences based on the internal structure. Such findings may facilitate the assessment process and planning of interventions that are more focused on individual needs.

Keywords

Item Response Theory, Social Skills, Children, Rating Scale

1. Introduction

A rating scale with several levels of response serves primarily to three functions. In the first instance, it allows the researcher to emphasize relevant areas for research to be conducted. Also, the format of the responses provides a number of possible answers to each question, allowing opt for an intermediate response. Finally, it requires that respondents use the same stimulus when formulating their responses. However, some problems may occur. Respondents can not use a scale as would be expected, choosing socially acceptable answers for example. In addition, if the items are ambiguous or not all are definitions; or when there is a neutral category (in the middle of the response scale) or too many of them, a noise may be introduced in the response obtained by a respondent,

affecting the test results and its validity (Smith, Wakely, de Kruif, & Swartz, 2003).

Parts of the literature regarding the Classical Tests Theory have investigated the effect of the proper number of categories in several tests of different psychological variables such as self-efficacy e.g. (Wewers & Lowe, 1990; Pajares et al., 2000, among others). In general, the methods of analysis used in these cases are factor analysis, internal consistency measures like Cronbach's alpha or regression analysis that assumes the use of interval data, even though the observable are at ordinal level.

The Rasch model can be used in the process of optimization of the number of points and categories of items without the need for administration of different versions of a same scale. This model assumes additivity of data, defined as units of measurement that has the same size in the continuous (interval data) if the data is well adjusted to the model. Such units are called "logits" and are a linear function of the probability of getting a certain score for a person with a given skill level. These interval measures may be used in parametric statistical analyzes, depending, however, of how the data fits to Rasch model. Thus, these parameters are estimated and used to determine the patterns of response expected for each item. The adjustment is derived from a comparison of these with the observed patterns. Such an assessment would provide validity evidence. In turn, the standard errors associated with each item calibration and ability estimates of persons are used to calculate precision in this model. These errors can be used to describe the confidence interval where the real difficulty of the items and ability of people is (Wright & Stone, 1988).

Bond and Fox (2001), Shanks & Lopez (1996) and Wright and Masters (1982) also agree that in the likert scales optimization process, when compared various forms of polytomic scales, beyond diagnostic categories other indicators such as the precision rates, separation and validity through InFit and outfit measures or even differential item functioning can be observed, providing additional evidences to different forms of likert scales of items. All procedures were used in this study and will be analyzed for each item format.

Although there are innumerable procedures to assess social skills, measurement of this kind of skill is still very controversial among researchers as most of the procedures have not been investigated thoroughly to confirm their quality. According to Hersen and Bellack (1977), this happens due to the doubtful nature of such procedures and may be related to the nature of social behavior, to the difficulty in finding an external validation criterion and/or to the lack of agreement concerning definition. Rosema, Crowe, and Anderson (2012) have described that limited literature, to build on children and adolescent's social results, and on the understanding of potential impairment due to the lack of social skills, suggests that this topic be investigated in more detail.

Excluding procedures such as interviews, controlled observations and self-reporting measures, some standardized instruments used to assess social skills should be referred as to Rathus Assertiveness Schedule (RAS), created by Rathus (1975), the first test systematically developed for this purpose. Another frequently-used instrument is the College Self-Expression Scale (CSES) developed by Galassi, Delo, Galassi & Bastien (1974).

The Social Performance Survey Schedule (SPSS) is another measurement to be considered. It was developed by Lowe and Cautela (1978). Caballo (1987) developed the Social Expression Multidimensional Scale (EMES). Trower, Bryant and Argyle (1978) created the Social Situations Inventory (SSI). In a study conducted by Lim, Rodger and Brown (2011), the Child Behavior Rating Scales were used (CBRSs).

In Brazil, Del Prette and Del Prette (2001) developed the Social Skills Inventory (IHS-Del-Prette). There was only one multimedia assessment system proposed by Del Prette and Del Prette (2005) to measure social skills in children.

Given the importance of assessing social skills in the teaching and learning environment, and based on the need to take psychometric properties into account, the Social Skills Test was created for school children and adolescents. One of the advantages of THAS-C specified by Bartholomeu, Silva and Montiel (2011) is providing distinguished information. By creating different levels of difficulty for each item, the test enables the identification of behaviors that are easily found in the social interaction context and, then, planning of interventions for people with poor social ability.

For a successful assessment, in addition to adequate tests, it is important to consider the scientific quality of these tests. There are different validation methods which are selected based on the aspects they will assess (American Educational Research Association—AERA, American Psychological Association—APA, National Council on Measurement in Education—NCME, 1999; Urbina, 2007).

In this study, the aim is showing validity evidence based on the internal structure for the Children's Social Skills Test (THAS-C). The initial study of THAS-C psychometric properties is based on an analysis of items, DIF analysis and IRT categories. Subsequently, new studies involving THAS-C internal structure analysis will

be presented to point out the advantages of this new tool.

2. Study 1—Adjustment to Rasch Model, Differential Item Functioning by Sex and Scales Optimization

2.1. Method

2.1.1. Participants

Initial studies investigated 257 children attending Grades 2 to 4 in public schools of two cities in the countryside of the state of São Paulo. Participants were 8 to 11 years old, mean age of 9 years ($SD = 0.77$), 58.8% were girls and 40.9% were boys. Studies included seven schools in one city and one school in another. The authors chose to include public schools in less (five schools) and more (three schools) socioeconomically favored neighborhoods.

2.1.2. Instrument

Children's Social Skills Test (THAS-C, Bartholomeu, Silva, & Montiel, 2011)

The items to be answered in a three-point Likert scale, corresponding to three dimensions, civility and altruism (13 items); resourcefulness and self-control in social situations (6 items); and assertiveness with confronting (4 items).

Some item samples for the three factors respectively are presented below.

I am kind to my room friends.

I'm ashamed to talk to the entire room.

I show my displeasure or annoyance to my colleagues without offending them.

2.1.3. Procedures

The instrument was applied collectively in classrooms only to children whose parents had previously granted authorization through an informed consent form (Ethics Committee approval number 424-2010). Children were first explained that they were participating in a research study and were given instructions on how to answer the instrument. Questions were read one by one by the researcher who waited for the answer before writing down the information provided.

2.2. Results

THAS-C indicators were investigated as to their fit to Rasch model in the Item Response Theory. The precision offered by this model provided a 0.96 coefficient for items and a 0.84 coefficient for people, which supports the idea that people have provided more information on the items than items have provided on their behaviors, although there is indication of high precision both for items and people. The average measurement error for items and people was 0.06 ($SD = 0.01$) and 0.10 ($SD = 0.01$), respectively.

Results have indicated an outfit mean of 1.00 ($SD = 0.19$), generally showing good item fit in this aspect. A more detailed analysis of these data shows that the outfit range was 1.47 and -0.65 , thus suggesting that not all items were within the model fit parameters. No item showed indexes above 1.50.

The infit mean was 1.00 ($SD = 0.17$), indicating that most of the items were within the expected parameters. The variation range was 1.38 to -0.64 , indicating that some items did not fit. Again, no item showed indexes above 1.50. The number of items that did not fit can be considered to be small, especially if we think it is a scale under development.

Generally, for people, infit and outfit indexes followed the good fit standard suggested by Bond and Fox (2001), with averages of 1.02 ($SD = 0.37$) and 1.01 ($SD = 0.37$), respectively. In summary, total infit discrepancies were 41.22% and outfit were 40.40%. This is a reasonable amount that suggests that almost half the subjects in question faced problems to solve the items, being necessary to take into account the extent to which these children actually understood what was being asked of them.

Children showed average social ability of 0.35 (Rasch measurement) ($SD = 0.28$) and a θ variation range of 1.33 to -0.40 . The mean difficulty of items in logit was 0.00 ($SD = 0.34$), with a range of -0.83 to -0.89 , suggesting that items have captured the mean social ability. In other words, this instrument was easily solved by the study children as they showed to fully agree with most of the sentences. Additionally, some of the items

were too easy because they assessed poorer social competency and did not have the right number of children to provide a good estimate.

Regarding gender-based differential item functioning, two methods have been used for analysis (Table 1).

Out of the 99 initial items found in the scale, only five (5.05%) had differences identified based on Draba criteria (1977), with t-values above 2.40. Using Mantel-Haenszel method, the same indicators favored one group or the other. Therefore, approximately 10% of the items showed gender bias. The authors decided to consider the five items presented in both criteria. This amount is actually small and is not indicative of major problems in this sense.

It was checked whether the number of categories in each item would be representative of the latent trait (social ability). Thus, as a person continuously progresses in terms of skills, each one of the subsequent points in scale become the most likely response (Linacre, 1999; Smith, Wakely, de Kruif, & Swartz, 2003).

To delimit the number of analysis categories necessary to adequately represent the construct, respondents' skill means were first analyzed for each category; threshold parameters were analyzed to determine which categories are not effective in measurement; and, lastly, an outfit analysis (mean square) was conducted to detect random uses of the analysis categories.

Although this analysis provides information on categories with potential problems, the final decision to merge or eliminate a certain category should be made not based on statistical criteria only, but also on expected assumptions concerning the variable targeted by the investigation. Additionally, scale improvement depends on the sample being studied and should be tested once again with a new sample of the same population (Smith, Wakely, de Kruif, & Swartz, 2003), which we will present later on in final scale studies.

A progression analysis using Rasch measurement model (Table 2) showed greater people's ability in function of response categories. On the other hand, the outfit index suggested good fit in all categories. By analyzing the threshold of "sometimes" and "usually" categories, a discontinuity was observed where the "sometimes" category increases progression until around 0, while the "usually" category starts reducing at this point. Among the solutions presented, the authors chose to merge the "sometimes" and "usually" categories that have similar thresholds (results show in Table 3 and Figure 1). These data show relevance for the structure of three analysis categories. Additionally, the "seldom" category was renamed as "never" to test the new structure regarding instrument replication. This new version provided more satisfactory results.

Table 1. Gender-based DIF measurements for items with significant difference (N = 257).

Item	Gender	Added DIF	Gender	Added DIF	Change	t
Items more frequent among the male gender						
Hs27	1	0.49	2	1.07	-0.58	-4.38
Hs75	1	0.38	2	0.75	-0.37	-2.91
Hs87	1	0.25	2	0.62	0.37	2.95
Items more frequent among the female gender						
Hs29	1	0.23	2	-0.30	0.53	4.11
Hs68	1	-0.45	2	-0.96	0.51	3.00

Table 2. Statistical data for THAS-C scale response categories (N = 257).

Category label	Score	Infit Mnsq	Outfit Mnsq	Category measure	Structure calibration	Standard error
Seldom	0	1.05	1.05	-1.62	-	-
Sometimes	1	0.93	0.89	-0.42	-0.07	0.02
Usually	2	1.05	0.99	0.43	0.10	0.01
Frequently	3	0.99	1.01	1.59	-0.02	0.01

3.2. Results

After this first analysis, the authors tried to replicate Rasch analyses conducted in Study 1 to identify: 1) whether the items would fit the model when applied to another sample; 2) whether some items would show gender-based differential functioning; and 3) whether the Likert assessment structure would suffer any changes.

According to Rasch analysis statistics, no items exceeded the maximum limit in terms of infit and outfit indexes (1.5), and all items are within acceptable fit parameters. For people, around 4% of infit and outfit lack of adjustment was observed. In the gender-based differential functioning, three items still showed DIF in this last version (6—I'm kind to my classmates; 20—I feel embarrassed to address the whole class; and 23—I feel embarrassed when I meet a new group and no one is familiar to me), the first and last ones favoring men, and item 20 favoring women. As in the first application, the items assessed median levels of social skills. People's skills changed both to higher and lower levels, thus revealing a gap in the development of new items to measure such ability levels.

Finally, Likert categories analysis showed relevance of three levels in this scale, with adequate progression in ability levels in function of levels and lack of outfit non-adjustments. Consequently, the structure initially studied and proposed in Study 1 was empirically confirmed in Study 2.

It's interesting to note that the test information function was better to the three-point likert scale format contrariwise to the four-point likert scale. Hence, the three-point likert scale is a better way to assess children's social skills as they can represent better their perceptions within this structure (Figure 2).

4. Discussion

This study was proposed considering the lack of evidence based on the Rasch model in social skills measures. The lack this kind of study, particularly that used these procedures in optimizing levels of items analysis in social skills scales in children restrict the discussion of the data.

Regarding the identification of difficulty levels to comply with the latent trait in a predominantly median way, it is necessary to emphasize the importance of future studies based on an adaptive testing approach that covers items from several measurement levels. In this sense, such approach would facilitate the assessment process and planning of interventions that are more focused on individual needs. By ranking the difficulty of the items, identifying which behaviors are easier to perform in the context of social interaction, enables to plan interventions for people who have low social skills to move on to more complex behaviors allowing an accurate assessment of the behavioral repertoire of the child and a behavioral schedule which calls for new researches. In other words, the data provided by this instrument not only indicates what specific behaviors of social skills are being measured but provides an individual's social capacity index, with the items loaded in each of factors also include the

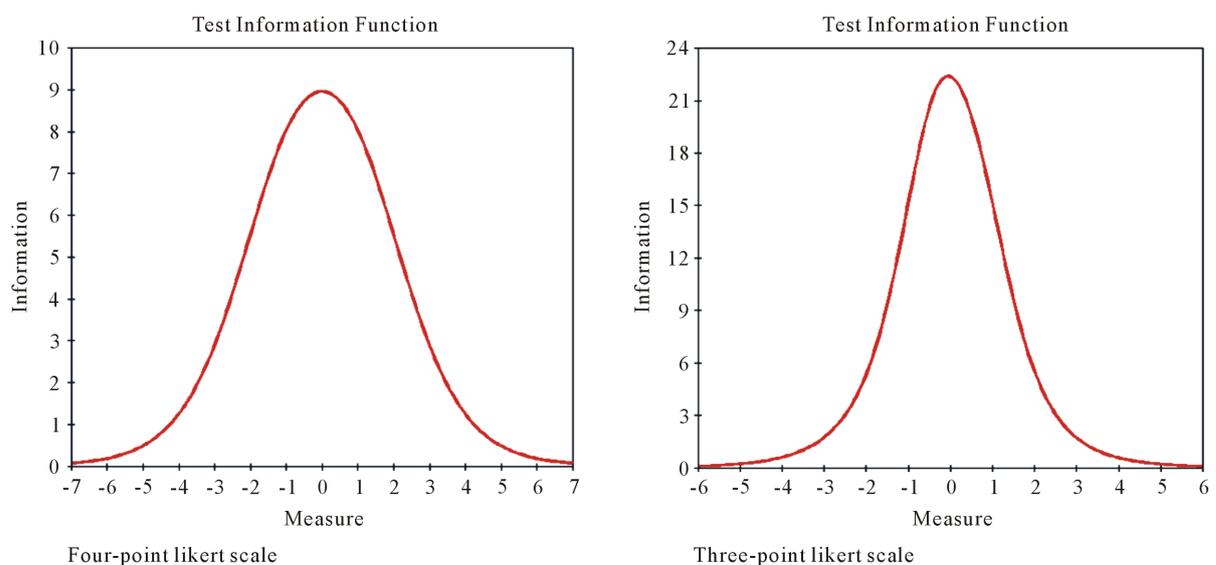


Figure 2. Test information curves with three and four-point likert scale.

behavioral categories mentioned in the literature. At the same time, identifies the interpersonal behaviors that are not identified in the other instruments in Brazil (Caballo, 2003; Trower, 1982).

In Study 1, a four-level structure of items was tested and refused and a three-point structure has been a good format in the assessment of children's social skills. Initially, it was believed that more levels would give more options to students and could provide more accurate information to their repertoire of social skills.

Indeed, data obtained in Study 1 and 2 refuted this hypothesis, since most of the items in four categories structure did not provide good fits to the Rasch model, in addition to the categories themselves presented problems. Analyzing the results of Study 1, the four categories caused more confusion in the responses of children and also provided less information of the social skills of children. By the way, future work with this scale could establish the three-point scale format.

Regarding the difficulty level of social skills items, most of the indicators in Study 1 who had lower levels of agreement (difficulty level of the items) were thus identified in Study 2. In fact, the subject's ability estimates showed differences from one study to another. As the items difficulty did not differ in these two studies and schools in both studies were similar, it could be assumed that the type of categories analysis did not affected the ability estimate of people.

Also, the precision of people and items was higher in Study 2 and the items misfits were higher in Study 1 concerning the InFit and outfit. For people the amount of misfits in Study 2 was lower.

Additionally, differential item functioning studies have to be conducted in more detail to allow clarification of identifiable gender bias linked to social skills. Gender differences in social skills measures are commonly observed in the literature (Del Prette & Del Prette, 2005; Bartholomeu, Silva, & Montiel, 2011). Despite, item bias by gender suggests that other latent constructs or measurement errors are being assessed in the test and that gender differences in the measure can be arising and inflated by item bias. Hence the exclusion of these items from the final scale can be a solution to avoid such measurement problem.

The items precision in Study 2 was superior compared with the first study. However, both indices are acceptable. The InFit range was higher in Study 2 also, although no item has been outside acceptable limits to that extent.

As for the outfit, the range of variation was again higher in Study 2, however, the amount of mismatches that was superior also in detriment of Study 1. However, the first study presented at the discretion of 1.5 violations, limit proposed by Linacre to accept an item, which was not observed in Study 2. In these terms, it may be suggested that the evaluation with three response levels would improve item fit, indicating fewer problems of this nature.

In addition, maximize the discrepancies, since the items that showed maladjustment problems, had worse rates compared to the Study 2. This data could be explained by the fact that those students calling for this discrepancy (in other words, that have unexpected agreements to their ability) with low social skills that would present a unexpected high concordance with items with higher level of difficulty, when assessed four possible points, it allows a greater variability to manifest their agreement. By reducing the amount of categories, these children have less likely to express such agreement, responding at the ends and not in the intermediate classes. This increases the values of outfit (InFit could also be changed for the same reason in this case was the outfit).

Another aspect that should be considered in explaining this result is that the outfit measure is based on the conventional sum of squared standardized residuals of people, extracting their average at the end (so mean square). In turn, the InFit is calculated based on a sum of the amount of information supplied by items weights. The amount of information of an item is its variance, the squared estimate of the standard deviation that is greater for central observation (the variability is highest for values around the average) than for outliers. The calculation is done by adding the residual values multiplied by the variance of the items to only then be added and extracted their average. Thus, let the differential effects of weight variances of each item, or the quantity of information. As a result, discrepancies in this indicator are more important. In this particular discrepancies were identified which is a good indicator for both modes of response to items (five and three levels) (Bond & Fox, 2001).

For persons, the accuracy in Study 1 was lower compared to the second. Thus, the amount of information provided to the items by people in Study 2 was superior. One possible explanation is the fact that in this study, the ability of people (level of social skills) was higher at the expense of the first. Thus, these students would provide better information about the items.

As for InFit and outfit, their respective ranges were higher in Study 2 but the total mismatches in both measures

was higher in Study 1. Violations of the 1.5 criterion in outfit were also higher in Study 1. Summarizing the information, in Study 1 items had fewer adjustment problems to the Rasch model, while in the second study, people had fewer problems of this nature. In this context, considering that the format of three categories showed better statistical results (as already mentioned and shown), this should be maintained. However, new studies could compare people with skill levels (motivation) and similar check whether the amount of mismatches outfit remain.

5. Limitations and Practical Implications

In the article describing initial studies for a rating scale of social skills in children, Bartholomeu, Silva and Montiel (2011) suggested a hierarchy, and social skills training with the Social Skills Test for children and adolescents of school age (THAS-C) by Rasch IRT and suggest that the planning of interventions based on the difficulty of the items can facilitate the process of learning these skills in order to insert new social behavior in the individual repertoire. In this study, a better refinement has been made to assure not only data structure but the likert scale and data bias by gender. These aspects could bias any intervention procedure based on this test results (that is, in the item's difficulty level). Hence, besides the validity evidences for THAS-C, the use of Rasch model in Social Skills tests can be very useful to indicate what kind of social behaviors are harder or easier to perform in the social context and this information can be used in planning interventions in such aspect.

This test, in its final form can be used in psychoeducational assessment to evaluate children's social skills in the dimensions presented and with a three point likert scale. Also, this test provides information regarding what children's are able or not to do in the school social environment and suggests points to interventions.

Finally, it remains to note that the reliability coefficients were satisfactory. Integrating this information with others already mentioned, allows attest to the good quality made instrument, enabling him to use in research. Further investigations can be conducted in order to verify that the factor structure of the instrument is kept under other conditions or even new precision studies as test-retest and discriminant validity. Further studies with other countries and other Brazilian states should be done as we have only data from São Paulo state. Also, in this study, we employ the Rasch model that shares the principle of invariance TRI parameters, would be expected in other samples of children, these behaviors were maintained with similar difficulty levels. Nevertheless, this fact must be properly investigated since the theories of social skills, they change because of the context due to their expectations and values. Thus it is reasonable to point out that the same social behavior appropriate in one context may not be in another, this fact calls for new studies and research. Still, further research could be done in order to establish difficulty levels in different stages of life and contexts for social skills, conduct which, by definition, would be expected to be different.

References

- AERA, APA, & NCME (1999). *The Standards for Educational and Psychological Testing*. Washington: AERA Publications Sales.
- Bartholomeu, D., Silva, M. C., & Montiel, J. M. (2011). Children's Social Skills Test: Psychometric Evidences of the Initial Version. *Psico-USF*, 16, 33-43. <http://dx.doi.org/10.1590/S1413-82712011000100005>
- Bartholomeu, D., Silva, M. C., & Montiel, J. M. (2014). *Social Skills Test for Children and Adolescents in School Situation-THAS-C*. São Paulo: Memnon.
- Bond, T., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah NJ: Lawrence Erlbaum Associates.
- Caballo, V. (1987). *Teoría, evaluación y entrenamiento de las habilidades sociales*. Valencia: Promolibro.
- Caballo, V. E. (2003). *Handbook of Assessment and Training of Social Skills*. São Paulo: Santos.
- Del Prette, A., & Del Prette, Z. A. (2001). *Psychology of Interpersonal Relationships: Experiences for Group Work*. Petrópolis: Vozes.
- Del Prette, Z. A., & Del Prette, A. (2005). *Psychology of Social Skills in Childhood: Therapy and Education*. Petrópolis: Vozes.
- Draba, R. E. (1977). *The Identification and Interpretation of Item Bias (Research Memorandum No. 25)*. Chicago, IL: The University of Chicago, Department of Education, Education Statistics Laboratory.
- Galassi, J. P., Delo, J. S., Galassi, M. D., & Bastien, S. (1974). The College Self-Expression Scale: A Measure of Assertiveness. *Behavior Therapy*, 5, 165-171. [http://dx.doi.org/10.1016/S0005-7894\(74\)80131-0](http://dx.doi.org/10.1016/S0005-7894(74)80131-0)

- Hersen, M., & Bellack, A. S. (1977). Assessment of Social Skills. In A. R. Cininero, & K. S. Calhoun (Eds.), *Handbook of Behavior Assessment* (pp. 509-554). New York: Wiley.
- Lim, S. M., Rodger, S., & Brown, T. (2011). Validity of Child Behavior Rating Scale in Singapore (Part 1): Rasch Analysis. *Hong Kong Journal of Occupational Therapy*, 20, 52-62. [http://dx.doi.org/10.1016/S1569-1861\(11\)70004-3](http://dx.doi.org/10.1016/S1569-1861(11)70004-3)
- Linacre, J. M. (1999). Investigating Rating Scale Category Utility. *Journal of Outcome Measurement*, 3, 103-122.
- Lowe, M. R., & Cautela, J. R. (1978). A Self-Report Measure of Social Skill. *Behavior Therapy*, 9, 535-544. [http://dx.doi.org/10.1016/S0005-7894\(78\)80126-9](http://dx.doi.org/10.1016/S0005-7894(78)80126-9)
- Pajares, F., Britner, S. L., & Valiante, G. (2000). Relation between Achievement Goals and Self-Beliefs of Middle School Students in Writing and Science. *Contemporary Educational Psychology*, 25, 406-422. <http://dx.doi.org/10.1006/ceps.1999.1027>
- Rathus, S. A. (1975). Principles and Practices of Assertive Training: An Eclectic Overview. *The Counseling Psychologist*, 5, 9-20. <http://dx.doi.org/10.1177/001100007500500403>
- Rosema, S., Crowe, L., & Anderson, V. (2012). Social Function in Children and Adolescents after Traumatic Brain Injury: A Systematic Review 1989-2011. *Journal of Neurotrauma*, 29, 1277-1291. <http://dx.doi.org/10.1089/neu.2011.2144>
- Shanks, D. R., & Lopez, F. J. (1996). Causal Order Does Not Affect Cue Selection in Human Associative Learning. *Memory & Cognition*, 24, 511-522. <http://dx.doi.org/10.3758/BF03200939>
- Smith, E. V., Wakely, M. B., De Kruijff, R. E., & Swartz, C. W. (2003). Optimizing Rating Scales for Self-Efficacy (and Other) Research. *Educational and Psychological Measurement*, 63, 369-391. <http://dx.doi.org/10.1177/0013164403063003002>
- Trower, P. (1982). Toward a Generative Model of Social Skills: A Critique and Synthesis. In J. P. Curran, & P. M. Monti (Eds.), *Social Skills Training: A Practical Handbook for Assessment and Treatment* (pp. 427-599). New York: Guilford Press.
- Trower, P., Bryant, B., & Argyle, M. (1978). *Social Skills and Mental Health*. London: Methuen.
- Urbina, S. (2007). *Foundations of Psychological Assessment*. Porto Alegre: Artmed.
- Wewers, M. E., & Lowe, N. K. (1990). A Critical Review of Visual Analogue Scales in the Measurement of Clinical Phenomena. *Research in Nursing & Health*, 13, 227-236. <http://dx.doi.org/10.1002/nur.4770130405>
- Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.
- Wright, B. D., & Stone, M. H. (1988). *Validity in Rasch Measurement*. Research Memorandum 54. MESA. Chicago, IL: University of Chicago.