Journal of School Psychology 55 (2016) 57–69

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Maximizing measurement efficiency of behavior rating scales using Item Response Theory: An example with the Social Skills Improvement System — Teacher Rating Scale☆

Christopher J. Anthony*, James C. DiPerna, Pui-Wa Lei

The Pennsylvania State University, USA

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form 9 December 2015
Accepted 17 December 2015
Available online 18 January 2016

Keywords:
Social skills
Problem Behaviors
Measurement efficiency
Item Response Theory
Polytomous models

ABSTRACT

Measurement efficiency is an important consideration when developing behavior rating scales for use in research and practice. Although most published scales have been developed within a Classical Test Theory (CTT) framework, Item Response Theory (IRT) offers several advantages for developing scales that maximize measurement efficiency. The current study provides an example of using IRT to maximize rating scale efficiency with the Social Skills Improvement System — Teacher Rating Scale (SSIS — TRS), a measure of student social skills frequently used in practice and research. Based on IRT analyses, 27 items from the Social Skills subscales and 14 items from the Problem Behavior subscales of the SSIS — TRS were identified as maximally efficient. In addition to maintaining similar content coverage to the published version, these sets of maximally efficient items demonstrated similar psychometric properties to the published SSIS — TRS.

1. Introduction

Difficulties in the area of social skills have been linked with numerous problems including juvenile delinquency (Roff, Sell, & Golden, 1972), social isolation (Chung et al., 2007; Matson & Boisjoli, 2007), and school drop-out (Parker & Asher, 1987). In addition, social skills demonstrate moderate relationships with academic achievement (Wentzel, 1993; Malecki & Elliott, 2002). As a result of their relationship with student success, social skills have been identified as an important educational outcome. Currently, all 50 states have explicitly specified standards related to social–emotional functioning, and 96% of states have standards for social–emotional preschool development (DiPerna, Bailey, & Anthony, 2014). Further, learning in schools is laden with social and emotional dimensions due to the necessity that students function within the social environment of a school (Zins, Weissberg, Wang, & Walberg, 2004). As such, efficient and accurate assessment of social skills is an important focus for education professionals and researchers.

Several rating scales have been developed to measure the social skills of children from preschool through high school. Examples include the School Social Behavior Scale-2 (Merrell, 2002) and the Walker–McConnell Scales of Social Competence and School Adjustment (Walker & McConnell, 1995). The Social Skills Rating System (SSRS; Gresham & Elliott, 1990) has been widely used to

☆ The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090438 to The Pennsylvania State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

* Corresponding author at: Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, 125 CEDAR Building, University Park, PA 16802, USA. Tel: +1 406 210 3527. fax: +1 814 865 7066.

E-mail address: cja5171@psu.edu (C.J. Anthony).

Action Editor: Michelle Demaray

http://dx.doi.org/10.1016/j.jsp.2015.12.005

assess children's social skills in both research and practice. Gresham, Elliott, Vance, and Cook (2011) reported that the SSRS was used in 127 studies published in 50 different peer-reviewed journals and 53 doctoral dissertations between 2003 and 2008. The revision of the SSRS, the Social Skills Improvement System — Teacher Rating Scales (SSIS — TRS) features a broader conceptualization of several domains of social skills (Gresham et al., 2011) and scores with psychometric properties similar to those of the original version.

Given the widespread use of behavior rating scales to assess social skills in research and practice, the efficiency with which a rating scale measures social skills is of critical importance. Teachers often are asked to complete multiple assessments as part of a comprehensive evaluation process, which further stretches the limited amount of non-instructional time they have available during the day. Similarly, the time required to complete rating scales is challenging in a research context as well. School-based research participants (often teachers) may have to complete several rating forms per class at several time points throughout the year. Although they may be compensated for these activities, inefficiency of measurement and unnecessary time burdens can add to teacher stress and lead to incomplete data or withdrawal from study participation. As a result, many researchers shorten existing measures to minimize time required of participants (e.g., Duckworth, Quinn, & Tsukayama, 2012). Given the importance of the reliability and validity of these scores for substantive research, development and evaluation of shorter or streamlined versions of social skills measures is crucial.

1.1. Test theory and measurement efficiency

Most behavior rating scales, including the SSIS — TRS, have been developed within a Classical Test Theory (CTT) framework. Briefly, CTT is based on the assumption that the observed score produced by a measure is equal to the sum of two parts: the true score, which reflects a student's ability, and measurement error, which results from any systematic or random factor not related to the construct of interest (Suen, 2008). One commonly examined source of such error is that associated with differences among items. This type of error is quantified and measured indirectly by examining the internal consistency of a measure, which refers to how well items on a measure are related (Barchard, 2010). A major focus of scale development in a CTT context is the maximization of internal consistency (Streiner, 2003). Although internal consistency is undeniably important, overemphasizing it can be problematic for several reasons (Briggs & Cheek, 1986). First, this emphasis encourages inclusion of a larger number of items, which increases Cronbach's α provided items are positively related to the total score. Second, maximizing internal consistency can result in the inclusion of highly similar items within a measure. Indeed, Streiner (2003) noted that Cronbach's α levels that exceed .90, a threshold often used for evaluating technical adequacy for individual decisions (e.g., Salvia, Ysseldyke, & Bolt, 2010) more likely indicate unnecessary redundancy rather than optimal internal consistency. As such, rating scales developed in a CTT framework may display some inefficiency, the elimination of which would result in more parsimonious and rater-friendly measures.

Efficiency of measurement is well addressed by another theoretical framework under which rating scales can be developed, Item Response Theory (IRT; Lord, 1980; Hambleton & Swaminathan, 1985). Broadly, IRT refers to a series of latent trait methods used to assess item functioning and estimate latent trait score. A major advantage of IRT in the context of measurement efficiency is its facilitation of graphical evaluation of item performance (Edelen & Reeve, 2007). As part of IRT procedures, information curves are produced displaying the level of information (akin to precision of measurement) produced by a particular item across varying levels of an underlying ability or trait (e.g., social skills; Hambleton & Swaminathan, 1985). This advantage allows rating scale developers to identify items that provide desired amounts of information at desired levels of an underlying ability or trait. For example, if a researcher is interested in developing a measure to be used primarily to identify children who have low social skills (i.e., a measure in which high precision at low levels of social skills is desired), the researcher could select items that give high information at lower levels of social skills. As a result, fewer items would be required to achieve a desired level of precision at the targeted social skill levels (because items that function well at higher social skill levels would be excluded). Such item evaluation also can be conducted on item parameter estimates produced by IRT, such as an item's discrimination (which is positively related to information) and location on the latent trait scale at which an item is discriminating (often referred to as location or difficulty). As a whole, these procedures allow researchers to identify and eliminate items that do not function as desired relative to the measure's primary purpose, thus maximizing efficiency of measurement.

In addition to the general utility of IRT in improving measurement efficiency, assessing model assumptions can be helpful in identifying sources of measurement inefficiency. Alongside the assumption of unidimensionality, two major assumptions tested within the IRT framework are the assumptions of local independence (Hambleton & Swaminathan, 1985) and functional form (Toland, 2014). Local independence requires items to be mutually independent (or uncorrelated in a weaker form of the assumption) after the latent trait(s) that they purport to measure are controlled. A violation of this assumption indicates that some items share common variance that is not due to the intended latent traits. Among various potential causes of local dependence, item redundancy is of particular interest for streamlining measures. For example, similar or slightly different wordings of essentially the identical items can result in local dependence. Violations can also be due to incorrectly specifying the dimensionality of the scale or other incorrect model specifications, however, so it is important to examine item content for redundancy and examine other potential causes of local dependence.

Furthermore, the assumption of functional form states that the empirical data or distribution roughly follows the function specified by the IRT model chosen to analyze the data (De Ayala, 2009). This assumption is examined through reviewing Option Characteristic Curves (OCCs) as well as calculating item- and model-level fit statistics. Examining OCCs can be especially helpful in the context of increasing measurement efficiency (Toland, 2014). Often, poorly functioning items or response categories can be