

**The content validity of a List of Indicators describing typically developing children's support needs**

**THE PROPERTIES OF A LIST OF INDICATORS**

Antonio M. Amor<sup>a\*</sup>, M. A. Verdugo<sup>a</sup>, B. Arias<sup>b</sup>, V. Aguayo<sup>a</sup>, and M. Fernández<sup>a</sup>

*<sup>a</sup>Institute on Community Integration, University of Salamanca, Salamanca, Spain;*

*<sup>b</sup>Institute on Community Integration, University of Valladolid, Valladolid, Spain*

Corresponding author:

Antonio M. Amor, MA

[aamor@usal.es](mailto:aamor@usal.es)

Institute on Community Integration. University of Salamanca

Room 132

Avda. De la Merced, 109-131

Salamanca (Salamanca), 37005, Spain.

## The content validity of a List of Indicators describing typically developing children's support needs

Word count: 6,970

**Background:** The List of Indicators is training material used for the Supports Intensity Scale—Children's Version (SIS-C). It is aimed at helping interviewers distinguish between extraordinary and age-related typical support needs in children with intellectual disability when implementing the SIS-C. **Aim:** To adapt and test the List of Indicators' content validity and rating scale's functioning in Spain. **Method:** 222 teachers reported their agreement with each indicator's description using a 5-point rating scale. **Results:** 353 of 366 indicators showed evidence of content validity, whereas analyses of the rating scale highlighted the necessity of subsuming one of the scale's categories within another.

**Conclusions:** The need for developing research-based training materials to develop training programs on the use of the SIS-C and the relevance of using the latest methodological approaches available when required are discussed.

**Keywords:** intellectual disability; education; children; adolescents; age-related typical support needs; extraordinary support needs; training material; professional training

Embedded in a socio-ecological approach and strengths-based perspective, the supports paradigm understands intellectual disability (ID) as a mismatch between persons with ID capabilities and the environmental demands of the contexts in which they participate (Schalock et al., 2010). This mismatch creates support needs, which can be understood as “a psychological construct referring to the pattern and intensity of supports necessary for a person to participate in activities linked with normative human functioning” (Thompson et al., 2009, p. 135).

From the perspective of the supports paradigm, the main difference between persons with and without ID concerns their support needs. Because those with ID

experience a persistent mismatch, their support needs are extraordinary and extend beyond what most typically functioning people require (Thompson et al., 2009). Embracing this paradigm has entailed a change in professional practices pertaining to ID and has emphasised the support planning and implementation process. This process begins by identifying the desired life experiences and goals of the person with ID along with his or her support needs to provide individualised support plans to enhance the person's functioning and quality of life (QoL; Schalock, 2018).

The importance of the support needs construct within this paradigm has motivated the development of tools for its measurement. Although different approaches to measure support needs exist, efforts are being made to develop standardised measures of the extraordinary support needs of people with ID based on the supports paradigm (Thompson & Viriyagkura, 2013). Nevertheless, performing a support needs assessment with these instruments poses challenges that are yet to be addressed.

One challenge relates to the nuances in the support needs construct in the case of children. In childhood, support needs are strongly correlated with age. Hence, children (with and without ID) present higher levels of support needs the younger they are, and as they age, their support needs decrease (Shogren et al., 2015). Then, the practical problem concerning support needs assessment for children with ID is to distinguish whether the support needs experienced by a child with ID are linked to his or her age (i.e., age-related typical support needs—support needs that typically developing same-age peers also possess) or whether, on the contrary, they are extraordinary (i.e., connected to the ID). This distinction is important for not only support needs assessment but also support planning because children with ID (and extraordinary support needs) require extraordinary supports (Thompson et al., 2009).

This age-related concern has been considered in the development of the Supports Intensity Scale—Children’s Version (SIS-C; Thompson et al., 2016), the first standardised support needs assessment measure for children with ID. The SIS-C is designed to assess support needs in children with ID aged 5 to 16 years old. Considering that support needs would be confounded by children’s age, Thompson et al. (2016) stratified the standardisation sample to develop norms according to age cohorts (i.e., 5–6, 7–8, 9–10, 11–12, 13–14, and 15–16 years) and, further, levels of intellectual functioning within each age cohort (Shogren et al., 2015).

The SIS-C is organised into two sections: Part I, Exceptional Medical and Behavioral Needs, and Part II, the Support Needs Scale. Part II focuses on support needs assessment in 61 daily life activities across seven domains: home life (HLA), community and neighborhood (CNA), school participation (SPA), school learning (SLA), health and safety (HSA), social activities (SA) and advocacy (AA). To determine extraordinary support needs, each activity is rated across three dimensions: frequency, time and type of extraordinary support. The SIS-C is administered by a qualified interviewer to at least two respondents. Observers reporting the support needs of a child with ID must know the child well and must have recently observed the child in different contexts (Thompson et al., 2016).

This tool is being internationally adapted and validated, and several studies have provided evidence of SIS-C validity and reliability (for detailed information, see Thompson, Schalock, & Tassé, 2018). One country that has been particularly involved in SIS-C validation is Spain because it uses two versions of the tool: the SIS-C Spanish and SIS-C Catalan translations (Thompson et al., 2018). A growing emphasis is placed on using the SIS-C in Spain to address the needs of students with ID (Amor, Verdugo, Calvo, Navas, & Aguayo, 2018). This emphasis is motivated by the necessity of

evolving practices from a deficit-based perspective towards the supports paradigm. The SIS-C is considered an opportunity to develop context-based individualised educational plans (IEPs) that enhance inclusion opportunities of students with ID and attain personal desired outcomes (Verdugo, Amor, Fernández, Navas, & Calvo, 2018).

However, using the SIS-C in practice necessitates addressing the aforementioned challenge concerning the nature of the support needs of children with ID. Despite the efforts of SIS-C research to illuminate the distinction between extraordinary versus age-related typical support needs, this concern is an applied problem that involves decision-making by the interviewer—the person who implements and scores the SIS-C. In this respect, when implementing the SIS-C, the interviewer is the first person to face the challenge of discerning the nature of the support needs of a child with ID based on the information reported by observers. Hence, the interviewer's knowledge on this issue will influence how the information reported by the observers is interpreted and how the decision concerning the type of support needs is made.

Given the importance of training for implementing the SIS-C, the American Association on Intellectual and Developmental Disabilities (AAIDD) has developed and distributed different training materials to the countries participating in SIS-C validation. These materials aim to train interviewers in implementing and using the SIS-C. To help identify the type of support needs of a child with ID, the AAIDD developed a list of indicators (hereafter, List of Indicators) based on a teachers' survey. The List of Indicators describes the support needs of typically developing children for the same activities, domains and age cohorts used in the SIS-C. Through these descriptions, this training material seeks to support interviewers by providing qualitative information about the age-related typical support needs for each SIS-C item and thus help

interviewers make decisions concerning the nature of support needs of a child with ID based on information provided by the observers (AAIDD, n.d.).

Implementing the SIS-C in Spain to develop IEPs requires addressing the practical challenge of discerning the nature of support needs in children with ID. In this respect, the availability of the Spanish SIS-C task force of the List of Indicators may help interviewers address this challenge with the descriptions provided by the list. However, this list remains unexplored in Spain; hence, it is necessary to adapt it and test its appropriateness in this context prior to using it to train interviewers. Considering this requirement, the purpose of this study is twofold: to present the translation and adaptation of the List of Indicators in Spain and to furnish evidence of its appropriateness. The research questions guiding the analyses of the appropriateness of the List of Indicators are:

- Can the indicators included in the List of Indicators be considered valid sources for accurate descriptions of typically developing children's support needs for the same activities, domains and age cohorts as those used in the SIS-C in Spain?
- Is the List of Indicators an effective survey for collecting teachers' subjective impressions of typically developing children's support needs in the Spanish context (i.e., can the appropriateness of the indicators be ascertained after analysing how the information used to determine their content validity has been gathered)?

## **Material and methods**

### ***Instrument***

The List of Indicators is SIS-C training material based on a teachers' survey (AAIDD, n.d.). It aims to help interviewers administer and use the SIS-C. Given the importance of age for determining the support needs of children, six versions of the List of

Indicators corresponding to the SIS-C age cohorts have been created (i.e., 5–6, 7–8, 9–10, 11–12, 13–14 and 15–16 years). Considered across all the cohorts, the List of Indicators contains a total of 366 descriptions (61 per age band), which are designed to educate interviewers on the support needs of typically developing children (who are expected to embody age-related typical support needs) aged 5 to 16 years old for each SIS-C item. Hence, this list helps interviewers to distinguish, based on the information provided by observers, whether the reported support needs are likely extraordinary (i.e., linked to the ID) or related to age (i.e., age-related typical support needs that typically developing same-age peers also experience).

Each indicator represents a daily life activity in a given domain for a certain age band, followed by a description of exemplary activities, a description of the possible support needs that typically developing children may have to pursue the corresponding activity and the rating scale's categories. Professionals express their agreement with the support needs described for each indicator by choosing the category that they believe best describes typically developing children's support needs. The 5-point Likert rating scale used to express agreement from 0 to 4 has the following categories: 0 = *Strongly disagree. Students need far less support than described*; 1 = *Disagree. Students need less support than described*; 2 = *Agree*; 3 = *Disagree. Students need more support than described*; and 4 = *Strongly disagree. Students need far more support than described*.

### ***Procedure***

The following steps were followed: (a) translation and adaptation of the List of Indicators, (b) data collection of teachers' subjective impressions of the indicators' descriptions and (c) data analysis.

First, the indicators were translated and adapted using Tassé and Craig's (1999) guidelines for effectively adapting items to different contexts from the original context: (a) translation/adaptation, (b) consolidation of translation/adaptation, (c) validation of

preliminary translation, (d) revision/adjustments, (e) pilot testing, (f) revision/adjustments and (g) field test validation.

All indicators were independently translated by two of the authors who possess accredited English language knowledge. Only certain exemplary activities were changed for cultural reasons (e.g., watching a baseball game became watching a soccer game). Because none of the research team members was an English native, the research team included another step in Tassé and Craig's guidelines, and the translated indicators were sent to a native English speaker, who translated them back into English. Finally, the entire team ensured that the meaning of the indicators remained unchanged.

Once translated, the indicators were sent to different researchers for feedback and suggestions on improving the indicators. Minor corrections were made, and consequently, the instrument was ready for use. Thereupon, the research team contacted different schools to share the research goal and request teachers' collaboration.

After schools had agreed to collaborate, the first author visited the schools and organised a two-hour seminar with the teachers who were willing to participate. During the seminar, the author explained the supports paradigm and how to complete the task using the List of Indicators. Teachers were required to select the version of the List of Indicators that matched the age groups they taught (e.g., a teacher working with 16-year-old students needed to select the 15–16 version) and show their agreement with each indicator's description using the rating scale. After the seminar, teachers were given a two-week period to complete the tool. Once the instruments were completed, they were collected for data analysis. All procedures were in accordance with the ethical standards on data protection in Spain and the 1964 Helsinki declaration and its amendments.

### ***Participants***



A total of 222 teachers with a mean age of 40.82 years ( $SD = 9.59$ ) and an experience of 16.66 years ( $SD = 9.73$ ) were consulted as experts on typically developing children's support needs owing to their daily experience with the students and their potential role as observers and interviewers for the SIS-C. Table 1 summarises participants' information.

<Table 1>

### ***Data analysis***

Bangdiwala's weighted statistic for ordinal data ( $B^W_N$ ) and the Bangdiwala's agreement chart (Bangdiwala, 1987) were calculated for each indicator to study content validity to determine how well the indicators reflect typically developing children's support needs. Analyses of the rating scale using the many-facet Rasch measurement (MFRM) model were performed to assess the scale's appropriateness (Sick, 2009), thus addressing the second research question. The software R v.3.4.2 (R Core Team, 2017) and Facets v.3.71.3 (Linacre, 2015) were used.

### **Results**

#### ***Research Question 1 – Content validity analyses***

One test of content validity involves analysing the agreement level among judges. In the study, teachers served as judges, showing their agreement or disagreement with each indicator. To test content validity, the  $B^W_N$  and charts for ordinal data were calculated.

The  $B^W_N$  (Bangdiwala, 1987) allows calculation of the agreement level among judges for each indicator to study the judges' agreement strength. In other words, the study focused not on the agreement between judges but on the agreement size among judges regarding the events to categorise (e.g., a perfect agreement between judges can be found for a category different from *Agree*, which would indicate weak evidence of content validity for a given indicator). This statistic expresses agreement strength on a scale from 0 to 1, with 0 indicating the absence of agreement and 1, the strongest

agreement possible. Agreement strength can be poor (.000 to .200), weak (.201 to .400), moderate (.401 to .600), good (.601 to .800) and very good (.801 to 1) (Bangdiwala, 1987).

One advantage of the  $B^W_N$  is its graphical approach, allowing researchers to represent the distribution of agreement to complement  $B^W_N$ . Bangdiwala's agreement chart provides a representation of the agreement among judges based on a contingency table. The chart is built as a square,  $n \times n$ , where  $n$  is the total sample size. The black squares, each one measuring  $n_{ii} \times n_{ii}$ , show the observed agreement. The black squares are within larger rectangles, each one sized  $n_{i+} \times n_{+i}$ . These rectangles show the maximum possible agreement, given the marginal totals. Partial agreement is determined by including a weighted contribution from the cells outside the diagonal and is represented in the chart with shaded rectangles, whose size are proportional to the sum of the frequencies of the cells (Bangdiwala, 1987).

$$B^W_N = 1 - \frac{\sum_i^k [n_{i+}n_{+i} - n_{ii}^2 - \sum_{b=1}^q w_b A_{bi}]}{\sum_i^k n_{i+}n_{+i}} \quad (1)$$

The  $B^W_N$  and charts were calculated for each indicator within each age cohort, totalling 366 calculations to analyse the content validity of the indicators for describing the support needs of typically developing children. Owing to word limits, the  $B^W_N$  results for each indicator alongside its representation cannot be shown, but all data (i.e.,  $B^W_N$  and charts) are available upon request to the first author. The percentages of indicators are presented according to the "agreement" range overall and for each age group. The minimum, maximum and mean of agreement size is shown for each domain, considering the age groups. The indicators that did not show content validity are also reported.

Table 2 provides the agreement size ranges (in percentages) for the entire List of Indicators across all age cohorts. As shown, the agreement size was very good and good for nearly all indicators, thus providing evidence of content validity.

<Table 2>

To examine the indicators in depth, Table 3 summarises the minimum, maximum and mean of the  $B^W_N$ , considering domains and age cohorts. The results also indicated high agreement among judges when categorising activities.

<Table 3>

Although results at the indicator level demonstrated content validity, specifying the indicators that did not show content validity was necessary. For those indicators, either the  $B^W_N$  was low, or the agreement chart was not close to the *Agree* category. Table 4 illustrates these indicators alongside Bangdiwala's agreement charts.

<Table 4>

### ***Research Question 2 – Rating scale assessment***

The Rasch measurement theory refers to a family of models used to assess the quality of tests and construct true interval-scale measures from the raw scores obtained from such instruments. This theory has inspired different Rasch-based models, such as the Rasch-Andrich rating scale model, the Rasch partial credit model and the MFRM model (Sick, 2009).

The MFRM model is commonly used for performances evaluated with subjective ratings (e.g., speaking assessments), permitting researchers to obtain estimates on a common logit scale of the parameters of the components of the facets involved in construct evaluation (Linacre, 2017). In the construct assessments based on judges' evaluations, such as those used in this study, the importance of judges' severity or leniency in determining these evaluation scores, as well as the difficulty of the tasks

evaluated, has been highlighted, with the judges and tasks being treated as facets of the construct assessment (Sick, 2009).

The indicators of the List of Indicators (i.e., hereafter “items”) and the teachers (i.e., henceforward “judges”) were considered facets of construct evaluation along a logit scale representing the “age-related typical support needs” construct. The analysis of the rating scale focused on judges’ assessment of how the rating scale, developed for assessing each item’s accuracy in describing the age-related typical support needs, was useful for the Spanish context.

The aim was to explain whether the 5-category rating scale worked properly using a strong logistic model for assessing the quality of tests (the List of Indicators is a survey that collects subjective ratings). Nevertheless, prior to positing any explanations, it was necessary to ascertain the facets’ adjustment to the MFRM model (depending on the estimates of their parameters on the common scale). Thus, evidence of the facets’ misfit would add noise, and no interpretation of the rating scale should be undertaken (Linacre, 2015). Hence, to assess evidence of the rating scale’s functioning, it was first necessary to analyse the facets’ adjustment to the model, then assess whether the rating scale was working. Information on the facets’ and rating scale’s adjustment to the MFRM model, the graphs of the facets’ distributions along the common logic scale (i.e., Wright’s map) and the probability curves of the rating scale categories were analysed.

The MFRM model is iterative. Thus, if the data (facets and/or rating scale) evidence a poor model adjustment, researchers can test where the problem might be (e.g., if the problem involves judges’ facets, extreme cases can be removed) and conduct additional estimations to test whether data adjustment to the model is possible (Linacre, 2017). Different iterations were necessary to achieve full data adjustment. The iteration processes are presented with the facets’ and rating scale’s estimates. In

iteration 4 (where the data fitted the model), a Wright's map and the probability curves of the rating scale categories were added.

#### *Iteration 1*

The facets' adjustment to the model was assessed prior to interpreting the rating scale's adjustment. To consider the facets as adjusted to the MFRM model, it was necessary to study four estimates: *SD*, *separation*, *strata* and *reliability*. Items' adjustment was indicated by high *SD*, *separation* > 1, *strata* > 2 and *reliability* > .80, whereas judges' adjustment to the model required low *SD* and *separation*, *strata* < 2 and low *reliability* (Linacre, 2017). To analyse the rating scale's adjustment, the *Rasch-Andrich thresholds* ( $\tau$ ) were calculated. In the case of a polytomous rating scale (as in the teachers' survey used in this study),  $\tau$  are understood as local dichotomies between adjacent Likert-scale steps (Sick, 2009). The rating scale's fit to the MFRM model is possible only if the  $\tau$  values exhibit a rising progression or monotonic order (Linacre, 2017). Categories and  $\tau$  values are presented in order (categories from 0 to 4;  $\tau$  from 1 to 4).

Item estimates evidenced a good fit ( $SD_{\text{items}} = .42$ ;  $separation_{\text{items}} = 3.81$ ;  $strata_{\text{items}} = 5.42$ ; and  $reliability_{\text{items}} = .94$ ), whereas judge estimates did not ( $SD_{\text{judges}} = .65$ ;  $separation_{\text{judges}} = 3.06$ ;  $strata_{\text{judges}} = 4.41$ ;  $reliability_{\text{judges}} = .90$ ). Regarding the rating scale's adjustment, the data did not fit the model. Although the categories' average mean values exhibited a rising progression (-1.27; -.53; -.06; .46; .89), the  $\tau$  values did not (-1.70; -2.54; 1.45; 2.80). Upon closer examination of the data, it seemed that the problem lay between  $\tau_1$  (-1.70) and  $\tau_2$  (-2.54), suggesting, from the logistic model, that category 1 was not the most likely along the continuum (Linacre, 2017).

Prior to asserting that this rating scale's category was invalid, it was necessary to identify the reason for the misfit in the case of both the judges and the rating scale. To be orthodox, the authors decided to first remove judges that did not fit the model and then repeat the analyses to determine whether the judges' and rating scale's fit was

possible. Judges whose outfit values were higher than  $3 ZStd$  and lower than  $-3 ZStd$  were removed (see Table 5) because they were considered “extreme” (Linacre, 2017).

<Table 5>

#### *Iterations 2 and 3*

After removing extreme judges, estimations were recalculated for depurating the model fit. Table 6 summarises the facets’ and rating scale’s estimates, identifying extreme judges by their outfit values for each iteration. As shown, the data did not fit the model, and the problem again lay with the judges’ estimates and the transition between  $\tau_1$  and  $\tau_2$ .

<Table 6>

Once the first possibility had been analysed (i.e., misfit caused by extreme judges) and the data still did not fit the model, it was necessary to determine whether the misfit stemmed from the rating scale being ineffective. Thus, the lack of monotonic order of  $\tau$  values in the three iterations with the problem in the transition between  $\tau_1$  and  $\tau_2$  indicated that category 1 was not working. It seemed that this category was ambiguous and that judges might have misunderstood it. To test whether the problem was in category 1 based on the  $\tau$  values found, the authors collapsed category 1 within category 0 and then repeated the analyses and tested the facets’ and rating scale’s (4 categories now) adjustment to the MFRM model (*Agree* was now in category 1 instead of category 2).

#### *Iteration 4*

Two parallel analyses were conducted after collapsing categories: (a) an analysis without the extreme judges and (b) an analysis with all the judges. After collapsing categories, whether the judges previously considered extreme were included in the data

pool was inconsequential because all the data fitted the model (see Table 7), indicating that the previous misfit problem did not concern the judges but the original rating scale.

<Table 7>

Wright's map and the categories' probability distribution (without extreme judges;  $n = 181$ ) are presented in this iteration. Wright's map (Figure 1) represents both the items (based on each item's difficulty in representing the age-related typical support needs) and the judges (ranked by their severity/leniency when assessing each item) in the logit scale situated in the central axis (positive indicates a "high level" of the construct, while negative indicates a "low level"). The figure describes the facets' relationships across the logit continuum and communicates important information. First, regarding the logit scale data, the judges' mean ( $M = .00$ ;  $SD = .32$ ) was slightly higher than the items' mean ( $M = -.54$ ;  $SD = .31$ ). Second, the judges' spread ( $-.87$  to  $.89$  logits) was also higher than the items' spread ( $.01$  to  $-1.23$ ). Third, 78 judges (43.09%) scored above the items' range, whereas no judge scored below it. Finally, the targeting region in the logits between item difficulty and latent construct presence in judges corresponded to more than half of the participants (56.91%), indicating an acceptable relationship between the facets in the logit scale (Linacre, 2017).

<Figure 1>

Figure 2 illustrates the differences between the rating scale categories' probability curves along the logit continuum with respect to item difficulty in iterations 1 and 4. In iteration 1 (left), category 1 was not the most likely category in the common scale, whereas in iteration 4 (category 1 collapsed within category 0), all categories worked.

<Figure 2>

**Discussion**

This article presents evidence of the appropriateness of a SIS-C training material in Spain to support interviewers to discern the nature of support needs of children with ID while implementing the SIS-C. Content validity analyses of the indicators were conducted and the rating scale's appropriateness of the List of Indicators was examined.

Regarding the first research question, the BWN and Bangdiwala's agreement charts were calculated for each indicator. Judges (teachers) exhibited strong agreement when categorising the accurateness of the indicators describing typically developing children's support needs. For 353 indicators, the agreement size was high and around the Agree category, showing evidence of their content validity. Only 13 of the 366 indicators presented difficulties regarding content validity. These indicators were situated mainly within the 9–10, 13–14, and 15–16 age cohorts in the HSA and AA domains, and professionals tended to consider that greater support was required (i.e., agreement concerning the category Disagree. Students need more support than described).

Different explanations may illuminate the results for these indicators. The areas for which the indicators did not function well are related to health, self-determination and social relationships for children aged 9–10 years and adolescents aged 13–16 years. Before further research is undertaken, developmental psychology can provide insights into these results. A constant in human development research is that as people grow and reach certain stages of development, developmental milestones become increasingly complex (Sigelman & Rider, 2015). Therefore, milestones can identify particular difficulties during adolescence that are due to risk-taking behaviours (Romer, 2010; Tymula et al., 2012) related to the HSA domain and social, cognitive, emotional and behavioural changes and competencies (Booker & Dunsmore, 2016; Kilford, Garrett, &



Blakemore, 2016) linked to interactions with others (which involve the SA and AA domains). Thus, as people grow, they face new challenges that may demand greater support from others, and teachers—perhaps aware of this—have considered that typically developing children require more support needs than are described for those indicators.

Concerning the rating scale's assessment analyses, 222 judges assessed how the rating scale works while assessing indicators describing age-related typical support needs. These analyses show that category 1 (*Disagree. Students need less support than described*) seemed not to have been understood by judges in Spain, as evidenced by data from multiple iterations that tested the facets' and rating scale's adjustment to the MFRM model. Adjustment was achieved only after collapsing categories, whereas all other categories (including *Agree*) showed a good fit, indicating that the judges understood them.

The results of the rating scale's analyses highlight the most important finding of this work, particularly when considered alongside the results of content validity. The fact that one of the rating scale's categories did not work in Spain implies that although evidence on content validity was found for 353 indicators, the indicators should not be used. Determining whether the indicators work is impossible, since the rating scale used to gather the information used for testing their content validity did not fit the logistic model. Hence, additional research is required before using the List of Indicators to train interviewers in Spain.

The lack of international studies furnishing evidence of the appropriateness of this List of Indicators hinders the generation of discussion regarding our findings. Nevertheless, the main finding reported in this study in relation to this material in Spain has important implications for researchers working on SIS-C validation who have

access to SIS-C training materials. If training materials associated with the SIS-C (regardless of their purpose) are to be used, then rather than assume they are valid, the appropriateness of those materials must be analysed. If the gathered evidence suggests that the material requires additional research (as in this case), there is no methodological justification for its use. However, without analysing these materials, whether their use is justified cannot be known. Given that interviewers must be qualified to administer the SIS-C (Thompson et al., 2016), and that this qualification is provided through training, offering interviewers training based on materials whose appropriateness is unproven could bias the training. This bias may distort information gathering through SIS-C use, providing a poor basis for support planning, which, instead of enhancing children's functioning and QoL, could hinder their development and inclusion. Hence, the lack of studies that have analysed SIS-C training materials and the List of Indicators is troubling because these items are closely related to the use of the SIS-C, a tool intended for international use in areas such health, social services and education. Thus, additional studies on this topic are required to generate discussion.

Another implication of this work is that the latest available approaches are preferable to address a research concern, when necessary. In this study, not only did we conduct analyses of content validity but also we performed analyses of how the information used for that purpose was gathered (i.e., rating scale's assessment). In this case, if the information provided by teachers had been used only for content validity analyses, the main finding of this work would have been evidence of content validity for nearly all the indicators. However, as discussed, the MFRM model analyses indicated the need for additional research prior to use of the List of Indicators in Spain.

The present research has several strengths. First, it foregrounds the SIS-C training materials as the object of study. This study is the first to contribute evidence

concerning the List of Indicators, which aims to help interviewers address challenges concerning SIS-C use, like discerning the nature of support needs in children with ID. Second, this work offers researchers who have access to SIS-C training materials a methodological framework for gathering evidence on the List of Indicators to generate discussion. Third, this work has been parsimonious, and thus the content validity of each indicator was studied. Finally, to our knowledge, this study is the only work to assess the appropriateness of a survey's rating scale (i.e., in this case, the List of Indicators), adopting the MFRM model using a large number of judges (N = 222).

However, this work also has limitations. First, the study used an incidental sample, which does not assure representativeness and affects the generalisability of results. Considering this, a bootstrapping strategy was adopted to generate different versions of the same data pool. Second, regarding the rating scale's assessment, the judges were all teachers, so testing the extent to which their expertise influenced the results was impossible. Finally, the List of Indicators (training material), the study design (contributing evidence on the list's appropriateness) and the results (additional research is required before using this material in Spain) highlight limited yet important practical implications of this research.

Thus, although additional research is required before using the List of Indicators in Spain, the significance of a valid List of Indicators is worth stressing, given its role in supporting the use of the SIS-C to distinguish between extraordinary and age-related typical support needs in children with ID. In this sense, the importance of training for SIS-C implementation and scoring (Thompson et al., 2016) necessitates the development of training programs with different goals (e.g., discerning the nature of the support needs of children with ID). The significance of offering evidence concerning the appropriateness of SIS-C training material is that it guarantees an adequate starting

point to develop such training programs. Once developed, it will be necessary to investigate the efficacy of the training programs, given their purpose. Thus, this work has an applied relevance on which to base the development of training programs concerning the SIS-C.

Finally, the limitations highlighted serve as a starting point for future research. Regarding the rating scale's analyses, participants from different areas (e.g., social work or psychology) should be included, and the ratings provided by them should be compared with those presented in this study to analyse the presence or absence of biases depending on each professional's expertise. If the data again show that the rating scale is ineffective, then re-defining the categories would be necessary, as the MFRM model shows. Thereafter, analyses of the content validity of the indicators should be conducted. If the data suggest that certain indicators do not show evidence of content validity, a qualitative study should be conducted addressing which support needs, in the participants' opinions, typically developing children might require to pursue the activities corresponding to the indicator, in order to improve the indicators' accurateness.

#### **Declaration of interest statement**

There was no conflict of interest

#### **References**

American Association on Intellectual and Developmental Disabilities (n.d.).

*Descriptions of Sample Activities and Support Needs for Typically Functioning Children in Relation to SIS-C Items* [unpublished training materials].

Washington, D.C.: Author. Translated and adapted version retrieved from:

<http://inico.usal.es/563/informacion-divulgacion-innovacion/recursos-online/lista-de-indicadores-ni-os-con-desarrollo-tipico-list-of-indicators-typically-developing-children.aspx>

- Amor, A. M., Verdugo, M. A., Calvo, I., Navas, P., & Aguayo, V. (2018). Psychoeducational assessment of students with intellectual disability: Professional-action framework analysis. *Psicothema*, *30*(1), 39-45. doi: 10.7334/psicothema2017.175
- Bangdiwala, K. (1987). Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS Users Group International Conference*, *12*, 1083-1088.
- Booker, J. A., & Dunsmore, J. C. (2016). Affective social competence in adolescence: Current findings and future directions. *Social Development*, *26*(1), 3-20. doi: 10.1111/sode.12193
- Kilford, E. J., Garrett, E., & Blakemore, S. (2016). The development of social cognition in adolescence: An integrated perspective. *Neuroscience & Biobehavioral Reviews*, *70*, 106-120. doi: <https://doi.org/10.1016/j.neubiorev.2016.08.016>
- Linacre, J. M. (2015). *Facet Rasch Measurement computer program* (version 3.71.3) [Computer program]. Chicago: Winsteps.com.
- Linacre, J. M. (2017). *A User's Guide to FACETS. Rasch-Model Computer Programs*. Chicago: Winsteps.com.
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing.
- Romer, D. (2010). Adolescent risk taking, impulsivity, and brain development: Implications for prevention. *Developmental Psychobiology*, *52*(3), 263-276.
- Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. L., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., ... Yeager, M. H. (2010). *Intellectual Disability. Definition, Classification, and System of Supports. 11th Edition*. Washington, D.C.: American Association on Intellectual and Developmental Disabilities.
- Schalock, R. L., (2018). Six ideas that are changing the IDD field internationally. *Siglo Cero*, *49*(1), 21-33. doi: <http://dx.doi.org/10.14201/scero20184912133>.
- Shogren, K. A., Seo, H., Wehmeyer, M. L., Palmer, S. B., Thompson, J. R., Hughes, C., & Little, T. (2015). Support needs of children with intellectual and developmental disabilities: Age-related implications for assessment. *Psychology in the Schools*, *59*(9), 874-891. doi: 10.1002/pits.21863
- Sick, J. (2009). Rasch Measurement in Language Education Part 3: The family of Rasch Models. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *13*(1), 4-10.

- Sigelman, C. K., & Rider, E. A. (2015). *Life-Span Human Development* (8<sup>th</sup> Ed.). Stanford: Cengage Learning.
- Tassé, M. J., & Craig, E. M. (1999). Critical issues in the cross-cultural assessment of adaptive behavior. In R. L. Schalock (Ed.), *Adaptive behavior and its measurement: Implications for the field of mental retardation* (pp. 161-184). Washington D.C.: American Association on Mental Retardation.
- Thompson, J. R., Schalock, R. L., & Tassé, M. J. (2018). *Evidence for the Reliability and Validity of the Supports Intensity Scales* [White paper]. Retrieved from: [https://aaid.org/docs/default-source/sis-docs/evidence-for-the-reliabilityandvalidity-of-the-sis.pdf?sfvrsn=7ed3021\\_0](https://aaid.org/docs/default-source/sis-docs/evidence-for-the-reliabilityandvalidity-of-the-sis.pdf?sfvrsn=7ed3021_0)
- Thompson, J. R., Valerie, J. B., Buntinx, W. H. E., Schalock, R. L., Shogren, K. A., Snell, M. E.,... Yeager, M. H. (2009). Conceptualizing supports and the support needs of people with intellectual disability. *Intellectual and Developmental Disabilities, 47*(2), 135-146.
- Thompson, J. R., & Viriyangkura, Y. (2013). Supports and support needs. In M. L. Wehmeyer (Ed.), *The Oxford handbook of positive psychology and disability* (pp. 317–337). New York: Oxford University Press.
- Thompson, J. R., Wehmeyer, M. L., Hughes, C., Shogren, K. A., Little, T. D., Copeland, S. R., ... Tassé, M. J. (2016). *Supports Intensity Scale-Children's version: User's Manual*. Washington D.C.: American Association on Intellectual and Developmental Disabilities.
- Tymula, A., Rosenberg, L. A., Roy, A. K., Ruderman, L., Manson, K., Glimcher, P. W., & Levy, I. (2012). Adolescents' risk-taking behavior is driven by tolerance to ambiguity. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), 17135-17140. doi: 10.1073/pnas.1207144109
- Verdugo, M. A., Amor, A. M., Fernández, M., Navas, P., & Calvo, I. (2018). La regulación de la inclusión educativa del alumnado con discapacidad intelectual: una reforma pendiente [The regulation of inclusive education of students with intellectual disability: a pending reform]. *Siglo Cero, 49*(2), 27-58. doi: 10.14201/scero20184922758

Table 1. Participants' demographic characteristics

<b>Variable</b>	<b><i>n</i></b>	<b>%</b>
Gender		
Male	69	31.10
Female	147	66.22
Missing	6	2.68
Age cohort		
5-6	37	16.67
7-8	35	15.76
9-10	35	15.76
11-12	37	16.67
13-14	43	19.38
15-16	35	15.76
Schooling		
Private school	115	51.80
Public school	107	48.20

Table 2. Judges' agreement size (age cohorts)

Age cohort	$B^w_N$ ranges (% of indicators)				
	Poor	Weak	Moderate	Good	Very Good
	.000 - .200	.201 - .400	.401 - .600	.601 - .800	.801 - 1
5-6	0	0	4.92	14.75	80.33
7-8	0	1.64	3.28	24.59	70.49
9-10	0	0	0	44.26	55.74
11-12	0	0	3.27	36.07	60.66
13-14	0	3.28	11.47	21.31	63.94
15-16	0	1.64	3.28	36.07	59.01
General	0	1.09	4.37	29.51	65.03

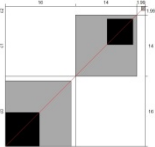
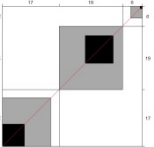
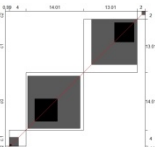
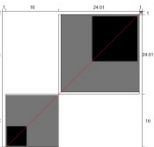
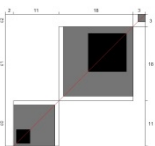
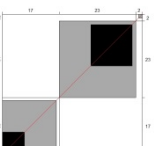
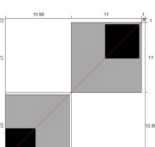
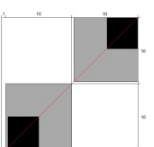
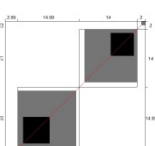
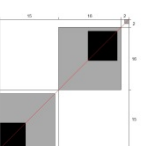
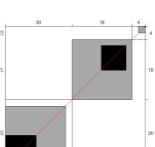
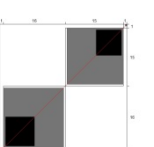
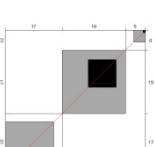


Table 3. Agreement size among judges (domains and age cohorts)

Age cohort	5-6			7-8			9-10			11-12			13-14			15-16		
	$(B^w_N)$			$(B^w_N)$			$(B^w_N)$			$(B^w_N)$			$(B^w_N)$			$(B^w_N)$		
Domain	m	M	<i>M</i>	m	M	<i>M</i>	m	M	<i>M</i>	m	M	<i>M</i>	m	M	<i>M</i>	m	M	<i>M</i>
HLA	.75	.92	.86	.49	.89	.81	.64	1	.77	.65	.89	.78	.44	.96	.82	.54	.95	.84
CNA	.84	.92	.88	.65	.87	.81	.64	.83	.77	.58	.85	.76	.71	.95	.87	.76	.87	.79
SPA	.72	.95	.83	.79	.88	.84	.73	.91	.81	.76	.94	.84	.73	.92	.77	.57	.92	.84
SLA	.82	.92	.87	.39	.92	.77	.74	.89	.83	.76	.91	.82	.31	.93	.80	.70	1	.84
HSA	.45	.97	.71	.50	.97	.83	.65	.91	.84	.49	.90	.73	.31	.88	.67	.22	.89	.71
SA	.72	.94	.87	.62	.95	.84	.70	.95	.85	.82	.87	.85	.52	.89	.75	.71	.86	.79
AA	.45	.96	.80	.62	.96	.85	.76	.89	.81	.76	.92	.85	.57	.91	.81	.66	.87	.79

Note.  $B^w_N$  = Bangdiwala's weighted statistic; HLA = Home Life; CNA = Community and Neighborhood; SPA = School Participation; SLA = School Learning; HAS = Health and Safety; SA = Social Activities; AA = Advocacy

Table 4. Indicators showing weak evidence on content validity (age cohorts and domains)

Domain/Indicator (age cohort)	$B^W_N$	Chart	Domain/Indicator (age cohort)	$B^W_N$	Chart
SLA/04 (7-8)	.72		HSA/07 (13-14)	.68	
CNA/07 (9-10)	.71		HSA/08 (13-14)	.87	
SLA/03 (9-10)	.78		AA/03 (13-14)	.78	
AA/02 (9-10)	.78		HSA/08 (15-16)	.22	
AA/09 (9-10)	.76		SA/02 (15-16)	.75	
CNA/05 (13-14)	.71		AA/04 (15-16)	.85	
HSA/06 (13-14)	.68				

Note.  $B^W_N$  = Bangdiwala’s weighted statistic; HLA = Home Life; CNA = Community and Neighborhood; SPA= School Participation; SLA = School Learning; HSA = Health and Safety; SA = Social Activities; AA = Advocacy

Table 5. Maladjusted judges to MFRM model ( $n = 37$ )

<b>Judge</b>	<b>Logit(SE) severity/leniency</b>	<b>Outfit (ZStd)</b>	<b>Judge</b>	<b>Logit(SE) severity/leniency</b>	<b>Outfit (ZStd)</b>
<b>221</b>	-1.21(.20)	5.7	032	.29(.22)	-5.5
<b>008</b>	-1.17(.20)	3.5	052	.29(.22)	-4.8
<b>022</b>	-1.13(.20)	3.2	157	.29(.22)	-4.3
<b>140</b>	-1.09(.20)	3.3	018	.34(.22)	-6.6
<b>219</b>	-1.05(.20)	4.2	016	.38(.21)	-6.0
<b>056</b>	-1.01(.20)	5.8	019	.43(.21)	-5.6
<b>220</b>	-1.01(.20)	3.9	020	.43(.21)	-4.4
<b>207</b>	-.89(.20)	4.2	217	.43(.21)	-4.1
<b>097</b>	-.72(.21)	3.4	033	.47(.21)	-4.9
<b>096</b>	-.67(.21)	3.1	036	.47(.21)	-2.5
<b>208</b>	-.67(.21)	3.7	072	.47(.21)	-3.2
<b>201</b>	-.39(.22)	3.3	069	.52(.21)	-4.9
<b>172</b>	-.25(.22)	3.1	035	.56(.21)	-4.6
<b>118</b>	.00(.22)	3.1	062	1.48(.16)	4.00
<b>039</b>	.05(.22)	6.9	004	1.73(.15)	3.7
<b>030</b>	.10(.22)	-3.4	109	1.78(.15)	3.2
<b>013</b>	.20(.22)	-3.9	206	2.71(.17)	6.2
<b>049</b>	.20(.22)	-3.5	210	2.83(.17)	7.1
<b>121</b>	.20(.22)	-3.7			

Note. SE = Standard error

Table 6. Data fit (iterations 2 and 3)

	Facets								Judge	Maladjusted Judges Logit(SE) severity/Leniency	Outfit (ZStd)	Rating Scale's Categories	
	Items				Judges							Avg. Meas.	$\tau$
	SD	Separation	Strata	Reliability	SD	Separation	Strata	Reliability					
Iteration 2	.44	3.54	5.05	.93	.59	2.70	3.94	.88	174	-.44(.22)	3.3	0 = -.97	$\tau_1 = -2.02$
									095	-.34(.23)	3.2	1 = -.58	$\tau_2 = -2.49$
									101	-.23(.23)	3.3	2 = -.02	$\tau_3 = 1.46$
									139	1.08(.19)	3.1	3 = .50	$\tau_4 = 3.04$
Iteration 3	.44	3.47	4.96	.92	.61	2.73	3.98	.88	NONE	NONE	NONE	0 = -1.00	$\tau_1 = -2.07$
												1 = -.60	$\tau_2 = -2.50$
												2 = -.01	$\tau_3 = 1.49$
												3 = .52	$\tau_4 = 3.08$
												4 = .87	

Note. SD = Standard Deviation; SE = Standard error; Avg. Meas. = Average measure;  $\tau$  = Rasch-Andrich threshold; 0 = *Strongly disagree. Students need far less support than described*; 1 = *Disagree. Students need less support than described*; 2 = *Agree*; 3 = *Disagree. Students need more support than described*; 4 = *Strongly disagree. Students need far more support than described*

Table 7. Data fit to the MFRM model after collapsing categories

	Facets								Rating Scale's Categories	
	Items				Judges				Average Measure	$\tau$
	SD	Separation	Strata	Reliability	SD	Separation	Strata	Reliability		
<i>Without extreme judges</i>	.28	2.22	3.29	.83	.23	1.08	1.77	.54	0 = -.80 1 = -.55 2 = -.40 3 = -.28	$\tau_1 = -2.53$ $\tau_2 = .77$ $\tau_3 = 1.75$
<i>All judges</i>	.30	2.60	3.80	.87	.22	1.02	1.70	.51	0 = -.84 1 = -.57 2 = -.42 3 = -.31	$\tau_1 = -2.55$ $\tau_2 = .75$ $\tau_3 = 1.80$

Note. SD = Standard Deviation;  $\tau$  = Rasch-Andrich threshold; 0 = collapsed category representing *less support than described*; 1 = *Agree*; 2 = *Disagree*. *Students need more support than described*; 3 = *Strongly Disagree*. *Students need far more support than describe*

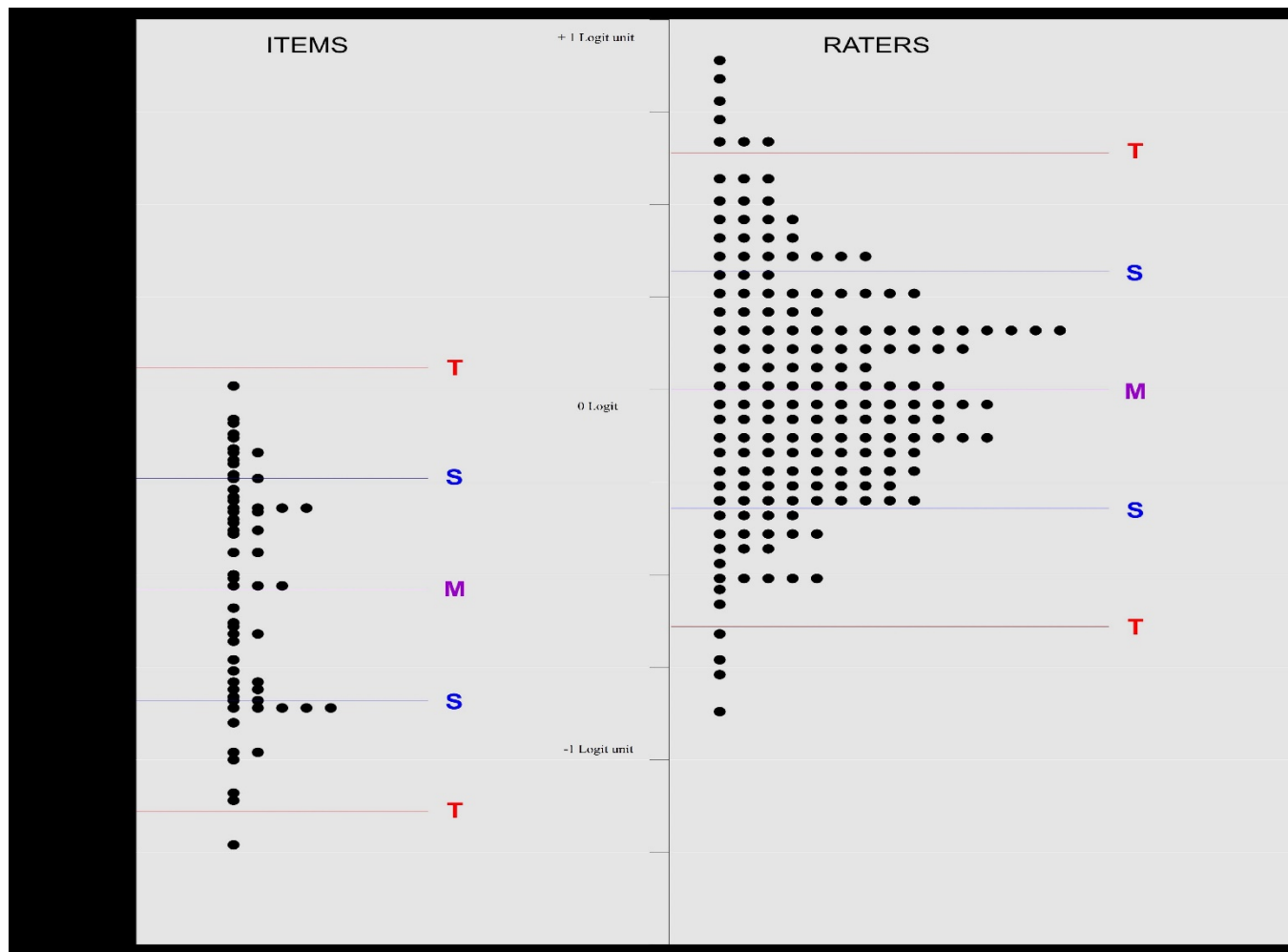


Figure 1. Wright's map. The central line represents the underlying construct. The right half of the figure represents the raters (i.e., teachers) who are ranked by their severity/leniency when assessing each item. Left side of the image represent the items (ordered by their difficulty in representing age-related typical support needs)

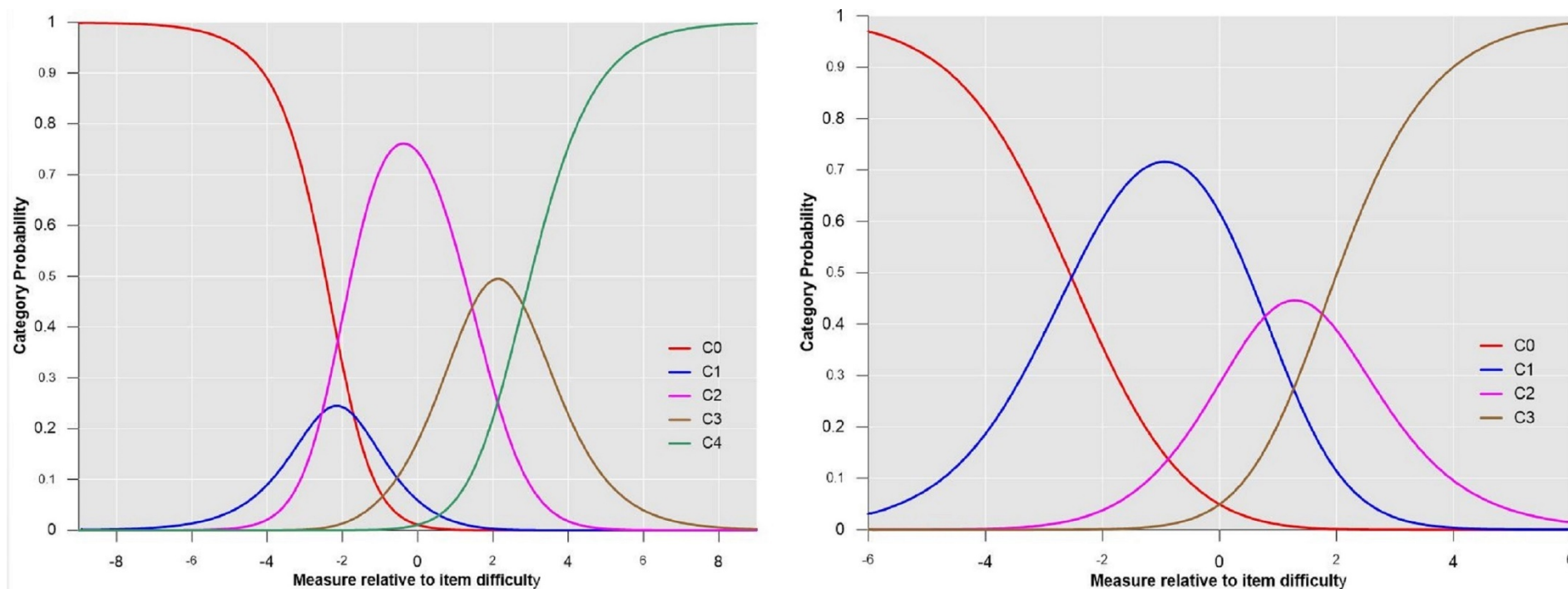


Figure 2. Probability curves of the rating scale categories (iterations 1 vs. 4). The Figure illustrates the differences between the rating scale categories' probability curves along the logit continuum with respect to the item difficulty in iterations 1 and 4. In iteration 1 (left), category 1 was not the most likely category in the common scale, whereas in iteration 4 (category 1 collapsed within category 0), all categories worked