ijbd

International Journal of Behavioral Development
2017, Vol. 41(6) 743–750

DOI: 10.1177/0165025416671612
journals.sagepub.com/home/ijbd
SAGE

The benefits of latent variable modeling to develop norms for a translated version of a standardized scale

Hyojeong Seo,[1] Leslie A. Shaw,[2] Karrie A. Shogren,[2] Kyle M. Lang,[3] and Todd D. Little[3]

Abstract

This article demonstrates the use of structural equation modeling to develop norms for a translated version of a standardized scale, the Supports Intensity Scale – Children's Version (SIS-C). The latent variable norming method proposed is useful when the standardization sample for a translated version is relatively small to derive norms independently but the original standardization sample is larger and more robust. Specifically, we leveraged a large, representative US standardization sample ($n = 4{,}015$) to add power and stability to a smaller Spanish ($n = 405$) standardization sample. Using a series of multiple-group mean and covariance structures confirmatory factor analyses using effects-coded scaling constraints, measurement invariance was tested across (a) Spanish only and (b) both US and Spanish age bands (5–6, 7–8, 9–10, 11–12, 13–14, and 15–16). After establishing measurement invariance across the US and Spain, tests for latent means and variance differences within age-bands were only performed for Spanish data; the latent means and variances in the US sample were freely estimated. The study findings suggest that the information in the US data stabilized the overall model parameters, and the inclusion of the US sample did not influence on the norms of the SIS-C Spanish Translation.

Keywords

effects-coded method of identification, international norming, Supports Intensity Scale – Children's Version

Scale norms derived from a standardization sample are essential to provide interpretable assessment scores. The norming process, however, requires a large sample size, particularly given that the standardization sample must be systematically stratified on relevant demographic variables. This sample-size burden is a challenge for researchers engaged in efforts to translate and apply scales to various cultural contexts that differ from the context where the standardization sample was generated. When it comes to international or cross-cultural norming, it is best practice to develop norms specifically for the national or cultural group that will use the scale rather than assuming a single set of norms can apply across countries, because the latter approach can produce statistically-biased norm scores due to cultural differences or problems in translation (Batram, 2008). However, it can be difficult to draw upon large enough sample sizes for stand-alone norming across multiple countries.

Some researchers may choose to develop multinational or aggregated norms (see Meyer, Shaffer, Erdberg, & Horn, 2015). Other researchers have proceeded with the development of norms with samples that may be considered small (see Lappalainen, Savolainen, Kuorelahti, & Epstein, 2009). However, small samples can lead to highly variable estimates (Angoff, 1984). Whatever the sample size, the approach undertaken to develop norms should take into account reliability and validity of scores. Elosua and Iliescu (2012) encourage the use of latent variable methods to evaluate dimensionality, investigate invariance of both the items and the factor structure, and establish criterion validity in models that have separate measurement error from the total score in addition to or instead of more traditional analyses (e.g. Cronbach's alpha, Pearson's *r*).

Recently, Seo, Little, Shogren, and Lang (2016) described a way of using structural equation modeling (SEM) to develop norms. They described the application of their technique to the Supports Intensity Scale – Children's Version (SIS-C; Thompson et al., 2016) in a US sample of 4,015 children and adolescents aged 5–16 years with intellectual disability. To norm the SIS-C, the large US standardization sample was stratified by two-year age-bands reflecting the hypothesized development changes in the targeted age range as well as by disability-related characteristics. Effects-coded scaling constraints, which provide a non-arbitrary metric for testing differences among groups (Little, Slegers, & Card, 2006) in the multi-group SEM framework, allowed for testing of age related-differences by age-bands and the creation of norms for each age-band. The latent means and variances estimated in this process were used to map raw scores to a percentile rank and then a standard score as McDonald (2011) defines it, where the score has been rescaled to a mean of 10 with standard deviation of 3 for the subscales and a mean of 100 with standard deviation of 15 for overall support needs.

[1] Kongju National University, Gongju, Republic of Korea
[2] University of Kansas, Lawrence, KS, USA
[3] Texas Tech University, Lubbock, TX, USA

Corresponding author:
Karrie A. Shogren, University of Kansas, 1200 Sunnyside Ave RM 3136, Lawrence, KS 66045, USA.
Email: shogren@ku.edu

After using multi-group SEM to norm the SIS-C with a sample of children and youth from the United States (see Seo et al., 2016), the need for separate norms for translated versions of the scale remained. Specifically, research teams from multiple countries or regions, including Spain, Catalonia, Italy, and Iceland, received permission from the publisher of the SIS-C to translate the assessment. Because of the differing cultural contexts and the fact that the norming sample was US-based, a method to establish international norms was critical. Given that many of the partner countries would struggle to generate a sufficiently large standardization sample, the SIS-C Development Team proposed to leverage the power of the US norming sample and the use of multiple-group mean and covariance structures (MACS; Little, 1997) models to provide robust norms for each international sample.

The purpose of this article is to describe how the approach introduced by Seo et al. (2016) can be extended to norm translated versions of previously validated scales when the context precludes collecting a large standardization sample. We use the Spanish translation of the SIS-C as a motivating example. Specifically, we elucidate our strategy by describing how the small standardization sample of SIS-C Spanish Translation ($n = 450$) can be linked to the original US standardization sample ($n = 4{,}015$) to generate norms. We demonstrate that the larger sample, in combination with the smaller sample from the translated version, can be leveraged to provide sufficient power and stability for the norming process for the translated version. In the following sections, we briefly describe the SIS-C and international SIS-C norming projects. We then describe the proposed extension of the Seo et al. (2016) method using the SIS-C Spanish Translation norming analysis as a contextualizing example. Finally, we provide considerations for when to use or not use the proposed approach.

Case study: Supports Intensity Scale – Children's Version

The SIS-C (Thompson et al., 2016) was normed on a sample of 4,015 children and youth with intellectual disability aged 5–16 years using a latent variable norming approach. The four-step norming process described in Seo et al. (2016) that utilized data from a large US standardization sample served as a foundation to generate unique norms for the translated versions of the scale.

Standardization samples

The US standardization sample consisted of 4,015 children and adolescents with intellectual disability between the ages of 5 and 16 years. The US norms and relevant standard scores were generated by stratification based on age bands, as support needs were assumed and confirmed to be correlated with age. The stratification bands were: 5–6, 7–8, 9–10, 11–12, 13–14, and 15–16 years. Within each age band, the sample was further stratified based on three levels of intellectual functioning (i.e., mild, moderate, severe/profound). Thus, in the original US standardization sample, there were 18 cells, and the target was 215 participants in each cell based on power and representativeness considerations. Further demographic information on US standardization sample is provided in the SIS-C Manual (Thompson et al., 2016).

The Spanish standardization sample contained 450 children and youth with intellectual disability aged between 5 and 16 years. Like the US sample, participants were stratified into the same six age bands and then three levels of intellectual functioning (18 cells total). In both the US and Spain, the same definition of intellectual disability (Schalock et al., 2010) is used to diagnose and classify children and youth with intellectual disability based on IQ and adaptive behavior deficits. To achieve the overall sample size goal of 450, we targeted 25 completed SIS-C protocols per cell. The goal of 450 total was based on both power and representativeness considerations. The percentage of males in the Spanish sample was lower (63.8%) as compared to the US sample (67.5%). With respect to the stratification variables of age and intelligence level, the Spanish sample was more evenly distributed than the US sample. Comparisons between the US and Spanish norming samples are provided in Table 1.

Table 1. Demographic characteristics of US and Spanish norming data.

	US ($N = 4{,}015$)		Spanish ($N = 450$)	
Variable	*n*	%	*n*	%
Gender				
Male	2,710	67.5	287	63.8
Female	1,202	29.9	163	36.2
Missing	103	2.6	–	–
Age cohort				
5–6	513 (513)	12.8	76	16.9
7–8	562 (562)	14.0	75	16.7
9–10	762 (787)	19.0	71	15.8
11–12	804 (844)	20.0	77	17.1
13–14	818 (822)	20.4	76	16.9
15–16	487 (487)	12.1	75	16.7
Missing	69 (0)	1.7	–	–
Intelligence level				
55–70 or Mild	1,157	28.8	150	33.3
40–55 or Moderate	1,321	32.9	150	33.3
< 39 or Severe/Profound	1,321	32.9	150	33.4
Missing	216	5.4	–	–
Adaptive behavior level				
Mild	948	23.6	124	27.6
Moderate	1,335	33.3	173	38.4
Severe/Profound	1,615	40.2	158	33.7
Missing	117	2.9	1	0.2

Note. Sample sizes in parentheses are estimates after imputing missing data. Only age was imputed because that variable was used in the norming process.

Measure

The SIS-C is completed by a trained interviewer with at least two respondents (e.g. teacher, family) who know the child with an intellectual disability well and can report reliably on support needs. When different respondents bring unique perspectives during the SIS-C interview, the qualified interviewers make the final decision on the best rating for the item based on their clinical judgement. The assessment consists of two sections: (a) Section 1—Exceptional Medical and Behavioral Needs and (b) Section 2—Supports Needs Index Scale. Section 1 measures 19 medical conditions (e.g. respiratory care, feeding assistance) and 13 challenging behaviors (e.g. externally directed behavior, self-directed behavior) that would impact support needs of children with intellectual disability. Scores from Section 1 are not included in the standardization process but instead provide descriptive information to guide supports planning. Section 2 consists of 61 items organized in seven life-

activity domains: Home Life, Community and Neighborhood, School Participation, School Learning, Health and Safety, Social, and Advocacy. There is one subscale score for each domain. All 61 items in Section 2 are rated on a 0–4 point-scale across three dimensions: frequency, daily support time, and type of support. These items are used to generate average scores for each subscale and an overall support needs score that is an average of the seven subscale scores.

The translation of the SIS-C to Spanish was done using a systematic process developed to guide translation activities developed by Tassé and Thompson (2010). The process involved four steps. First, two teams that included a professional translator and bilingual content expert worked independently to translate each item on the scale, and then met with one another to compare their translations. Disparities were resolved and the *Preliminary Translation* emerged. Second, the *Preliminary Translation* was given to a second small group of bilingual content experts and translators who verified the translation. Third, the translation was piloted with a group of potential test users who were asked to provide feedback. Fourth, any additional edits were made and the translation finalized.

Missing data

Keeping in mind that the process followed for norming the SIS-C Spanish Translation would be replicated for norming the measure with data from other countries, the imputation process started with an item level imputation of the US sample. The amount of item-level missingness ranged from 0.07% to 1.97%, so a single data set was imputed (Little, Jorgensen, Lang, & Moore, 2014) with predictive mean matching in the mice package (van Buuren & Groothuis-Oudshoorn, 2011) in R 3.2.0 (R Core Team, 2015). This imputed data set was compared to the US norming imputed data, a data set that had been imputed on parceled indicators, to ensure that the results were unchanged. Comparison of invariance testing results indicated that the biggest difference, which was still negligible, from the two imputation approaches was found on the mean of the social activities construct (difference = .003). Given this, a copy of the imputed US data set was exported for use with all future international SIS-C norming projects. Inspection of missing data in the Spanish data set revealed that only one question was missing two responses out of 450. The imputed US data were merged with the Spanish data set for a single, item-level imputation using the same imputation software.

Norming procedure

We employed the same general process used to norm scores for the US version of the SIS-C to generate the Spanish norms. In norming US SIS-C scores, Seo et al. (2016) established measurement invariance and then tested the equality constraints imposed on the latent means and variances to detect group differences on each support needs subscale construct. It is crucial to emphasize that the effects-coding method of identification (Little et al., 2006) was used throughout the norming process to estimate latent parameters on non-arbitrary scales that reflect the metric of the manifest variables. Once latent means and variances were estimated and tested for equality across groups for each support needs subscale construct, the process was repeated for the overall support needs model. Lastly, standard scores and percentile ranks were generated for each subscale and overall support needs across the age groups. The procedures employed to create the US SIS-C norms have been fully described by Seo et al. (2016) and Shogren et al. (2015).

To generate Spanish norms, we mirrored the analytic procedure used in the US norming to test measurement invariance and equality constraints imposed on latent means or variances; however, this testing occurred in three steps. First, we tested measurement invariance in a 2-group model, Spanish data and US data. Second, we evaluated measurement invariance across the 6 age groups with only the Spanish data. Models that passed measurement invariance in the 6-group model were then merged with the US data for models with 12 groups (6 US age groups and 6 Spanish age groups). Spanish-only models that failed invariance testing were modified to pass partial invariance and those same changes were made in the 12-group models prior to evaluating measurement invariance again. The inclusion of the US sample added power and parameter stability during the norming process. To be more specific, the method we propose reduces noise in the measurement parameters (i.e., factor loadings and item intercepts) of the small sample (i.e., the Spanish sample, in our example). By enforcing strong invariance constraints, the factor loadings and item intercepts for the small group are equated to those of the large group (i.e., the US sample, in our example) thereby reducing estimation uncertainty for these parameters. This effect is confirmed by the relatively smaller standard errors of the 12-group model's measurement parameters found in the sensitivity analysis we describe in what follows. In MACS CFA, the measurement parameters dictate how the observed information is translated into the latent parameter estimates used for norm generation. Therefore, more precise estimates of these measurement parameters lead to a more certain mapping of observed scale characteristics to latent parameters. The inevitable consequence of our proposed process is a set of norms for the small group that are less noisy (i.e., more stable over hypothetical repeated norming analyses) than norms constructed from only the small sample would have been.

Parceling. The parceling scheme used for the norming process was described by Seo et al. (2016). Parcels were used to represent more parsimonious renditions of the constructs and minimize problems with model estimation (Little, Rhemtulla, Gibson, & Schoemann, 2013). The procedures we describe herein can be also be applied using item-level models, but doing so is not necessary to achieve veridical norms with our technique. In what follows, we describe the necessity of enforcing strong measurement invariance before creating the norms, but it is only necessary to do so for whatever measured items are used to fit the MACS CFA models (be they raw indicators or parcels). If the parceled solution supports strong invariance, then the latent parameters used to construct the norms are not contaminated by the effects of differential item functioning (DIF), even if item-level DIF may exist when not using parcels. Our technique relies only on optimal estimates of the latent means and standard deviations, not on any item or parcel level parameters. As a consequence, researchers using our proposed technique need not concern themselves with what DIF may exist in a model that was not used to estimate the latent parameters from which the norms were generated. We used the same parceling scheme created for the US norming process; the 61 items in Section 2 of the SIS-C were averaged into 21 parcels representing three indicators per construct. To create the final standard scores and percentile ranks for the overall score, these 21 indicators were further averaged into seven indicators loading onto an overall support needs factor, with one indicator for each life activity.

Table 2. Fit indices for the nested sequence in the multiple-group CFA.

Model	χ^2	*df*	*p*	RMSEA	RMSEA 90% CI	CFI	TLI	SRMR	Change in CFI	Constraint tenable
Country										
Configural	3503.4	336	.00	.065	.063–.067	.975	0.969	.019	–	–
Weak	3570.6	350	.00	.064	.062–.066	.975	0.970	.021	.000	Yes
Strong	2852.9	364	.00	.064	.062–.066	.974	0.970	.022	.001	Yes
Subscales Spain										
Configural	2125.5	1008	.00	.122	.114–.129	.936	0.920	.027	–	–
Weak	2226.0	1078	.00	.119	.112–.126	.934	0.923	.040	.002	Yes
Strong	2365.2	1148	.00	.119	.112–.126	.931	0.924	.045	.003	Yes
Subscales Spain + US										
Configural	6676.8	2016	.00	.079	.077–.081	.964	0.955	.023	–	–
Weak	7042.4	2170	.00	.078	.076–.080	.962	0.956	.033	.002	Yes
Strong	7640.5	2324	.00	.078	.076–.080	.959	0.955	.035	.003	Yes
Overall support needs index Spain										
Configural	358.2	84	.00	.209	.187–.231	.942	0.913	.026	–	–
Weak	404.6	114	.00	.184	.165–.204	.938	0.932	.079	.004	Yes
Strong	520.8	144	.00	.187	.170–.204	.920	0.930	.096	.018	No
Partial Strong	403.8	136	.00	.162	.144–.180	.943	0.947	.088	.005	Yes
Overall support needs index Spain + US										
Configural	2509.6	168	.00	.194	.187–.200	.934	0.900	.031	–	–
Weak	2638.6	233	.00	.167	.161–.172	.932	0.926	.061	.005	Yes
Strong	2976.4	292	.00	.157	.152–.162	.924	0.934	.070	.008	Yes

Note. Each nested model contains its constraints, plus the constraints of all previous, tenable models. When the constraint is tenable, it means that the equality constraints placed on measurement parameters of interest can be retained.

Multiple-group confirmatory factor analysis. For both subscale and overall scores, we conducted a multiple-group mean and covariance structure (MACS; Little, 1997) CFA for the 12-group model to establish measurement invariance and evaluate invariance of the latent constructs. Mplus 7.2 (Muthén & Muthén, 2012) was used for all analyses. Please see the supplementary materials for Mplus syntax examples for the MACS CFA to norm the SIS-C Spanish translation (see Supplemental Appendix 1: Mplus Syntax Examples).

Measurement invariance test. Prior to the norming process, we examined measurement invariance within the Spanish sample and then within the pooled US and Spanish samples. Measurement invariance was evaluated for both the life activity constructs and the overall support needs model across age groups. Tests of measurement invariance mainly consisted of three distinct levels: configural, weak, and strong invariance (Brown, 2015) for equivalence of structural form, factor loadings, and intercepts, respectively. If change in CFI is less than .01 when moving from one level of invariance constraint to the next, then the more restrictive level of invariance was supported (Cheung & Rensvold, 2002), and we proceeded to test the next level. In this case, the invariance constraints imposed on the measurement parameters are regarded as tenable (see the last column of Table 2). When either weak or strong invariance is not established, partial measurement invariance testing can be conducted. Findings from measurement invariance testing suggested that invariance could be established for each of the life activity constructs, suggesting that the latent variables reflect equivalent concepts across age bands, and most importantly, across countries. With measurement invariance, the constructs are equivalently defined in all groups (Brown, 2015), and the stabilizing power of the US norming sample can be leveraged to reduce the sources of sampling error that could arise from utilizing only the Spanish sample. The overall support needs model was invariant at the weak invariance stage indicating equal factor loadings but not fully invariant at the strong invariance stage indicating intercept differences requiring further testing to establish a partially invariant model for overall support needs. Results from the five sets of measurement invariance tests are listed in Table 2.

If either weak or strong measurement invariance is not supported, nested model testing should be used to determine which groups' factor loadings and/or intercepts are statistically different from the others. After re-estimating the model using fixed-factor coding, instead of effects coding, to identify and set the scale of the latent parameters, indicators should be freed one at a time and statistical significance compared with nested model testing and using an adjusted α-level to protect against Type I errors. Once differentially functioning indicators are identified, changes should be introduced in the model with effects coding. Only the parameters that have been identified as non-invariant should be freed across groups so that the groups remain linked by the parameters that can be equated. Relaxing the invariance constraints in this way produces a partially invariant model. Using Mplus syntax (Muthén & Muthén, 2012) for an example, the effects coding for intercepts would require the following model constraint:

$$t1 = 0 - t2 - t3 - t4 - t5 - t6 - t7 \quad (1)$$

where t1 – t7 represent parameter labels, one for each of seven indicators; the effects code averages to 0. The full strong invariance model would apply these same seven labels to the intercepts for all groups. If one group was found to differ on the first (t1) and last (t7) intercepts, partial invariance could be achieved by introducing a second effects code for that group with invariant items t2 – t6 still being equated across all groups as follows:

$$t11 = 0 - t2 - t3 - t4 - t5 - t6 - t17 \quad (2)$$

Table 3. Mean comparisons across age groups for the home life activity model.

Model	Model name	χ^2	*df*	Model comparison	$\Delta\chi^2$	Δ df	*p*	Constraint tenable
Strong invariance model (Subscale scores) (Bonferroni correction = .01/5 = .002)	M1	7640.53	2324	–	–	–	–	–
5–6 = 7–8	A1	7642.13	2325	M1 vs. A1	1.60	1	.206	Yes
5–6 = 7–8 = 9–10	A2	7642.42	2326	A1 vs. A2	0.29	1	.590	Yes
5–6 = 7–8 = 9–10 = 11–12	A3	7652.33	2327	A2 vs. A3	9.90	1	.002	No
11–12 = 13–14	A4	7640.94	2325	M1 vs. A4	0.41	1	.523	Yes
11–12 = 13–14 = 15–16	A5	7641.25	2326	A4 vs. A5	0.30	1	.581	Yes
[5–6 = 7–8 = 9–10] ≠ [11–12 = 13–14 = 15–16]	A6	7643.13	2328	M1 vs. A6	2.60	4	.627	Yes

Note. Values are calculated to three decimal places but reported with two decimal places. There are rounding errors in some cases. Highlighted models are the final latent mean models. Values for RMSEA, CFI, TLI, and SRMR were unchanged from Strong model results from the Subscales Spain + US model. When the constraint is tenable, it means that the equality constraints placed on measurement parameters of interest can be retained.

In the special case where only one factor loading or one intercept has been identified as different, an additional parameter of the same type will also need to be freed in order for effects coding to work. For example, if the intercept corresponding to t7 was freed but all other intercepts were still equated, t7 would be unchanged and model fit would not differ from the model with all intercepts were equated across groups (Little et al., 2006).

Estimation of latent means. The second set of MACS CFA analyses evaluated mean differences in the Spanish age groups across the 12-group norm-generating models. Equality constraints were placed only on Spanish latent means (not on US means). To parallel the US norming process, nested model testing was used to evaluate whether latent means could be constrained to equality across age groups, and this comparison was conducted sequentially across the age groups, one life activity construct at a time. Once a latent mean was evaluated across all Spanish age groups, the strong invariant model was used as the starting point to test the next life activity construct. The procedure was repeated with each of the other six constructs and the overall support needs model. To determine the tenability of equality constraints placed on a given factor, χ^2 difference tests between nested models (i.e., model with equality constraints vs. model without equality constraints) were performed using Bonferroni corrections. A sequence of nested model test results for one of the life activity constructs, Home Life, is shown in Table 3, and the results used to determine the tenability of equality constraints were presented in the last column of Table 3.

Estimation of latent variances. The third set of analyses for MACS CFA was to test variance differences across the six Spanish age groups in the 12-group norm-generating models. As in the latent mean comparisons, sequential tests were conducted per construct by gradually increasing the number of constraints to compare variances among the age groups. The final latent mean model for each factor was used as a baseline model for the nested model testing for the corresponding latent variances in order to determine if variances could be equated across the age groups. These steps were repeated for the overall support needs model. An example of the final latent variance model for one subscale, Home Life, is shown in Figure 1. In this final model, factor loadings and intercepts were equal across all 12 groups though residuals and latent covariances in the subscale models were not. Latent means and variances for the other subscales also varied freely, along with the latent parameters for the US age groups.

Standard scores and percentile ranks. After latent means and standard deviations of the life activity constructs were determined, *Z* scores based on latent means and standard deviations were computed for each life activity subscale and age group. The overall support needs model latent mean and standard deviation were used to generate *Z* scores for the overall score for each age group. *Z* scores were converted into standard scores with a mean of 10 and standard deviation of 3 for subscale scores and a mean of 100 and standard deviation of 15 for the overall support needs score. To further interpret raw scores against the population, the standardized percentile ranks were calculated for subscale and overall scores. Each of these standardized percentile ranks represents the proportion of students in the population who have lower scores on a given subscale or overall score than people having that given score. Standardized percentile ranks were computed by calculating the quantiles of a normal cumulative distribution function with the same means and standard deviations used for *Z* score transformations. Table 4 provides an example of the standard scores, standardized percentile ranks, raw scores, and raw score ranges for two age groups on the Home Life activities construct. The raw scores and ranges were based on the final latent mean and variance estimates obtained in the previous step.

Sensitivity analysis. Although we established measurement invariance between two groups (US 5–16- vs. Spanish 5–16-year-olds) in preliminary analyses and then for all 12 groups, we also tested each age group alone in six separate 2-group models, comparing US to Spain for the subscale and the overall support needs model. This step was taken to further inspect the function of the parceling scheme used for the SIS-C international norming process across age groups. Measurement invariance was established between countries for each age group.

Methodological checks. Because the process described in this article is new, additional model comparisons were run throughout the process to ensure that results were not unduly influenced by the large US sample. The method we describe is only meant to reduce uncertainty in the small sample's measurement parameters, not to substantively shift any of its latent parameter estimates. Summary results of these analyses are provided below, and tables documenting the results from all sensitivity analyses are provided in the supplemental materials (see Supplemental Appendix 2: Results from sensitivity analyses).

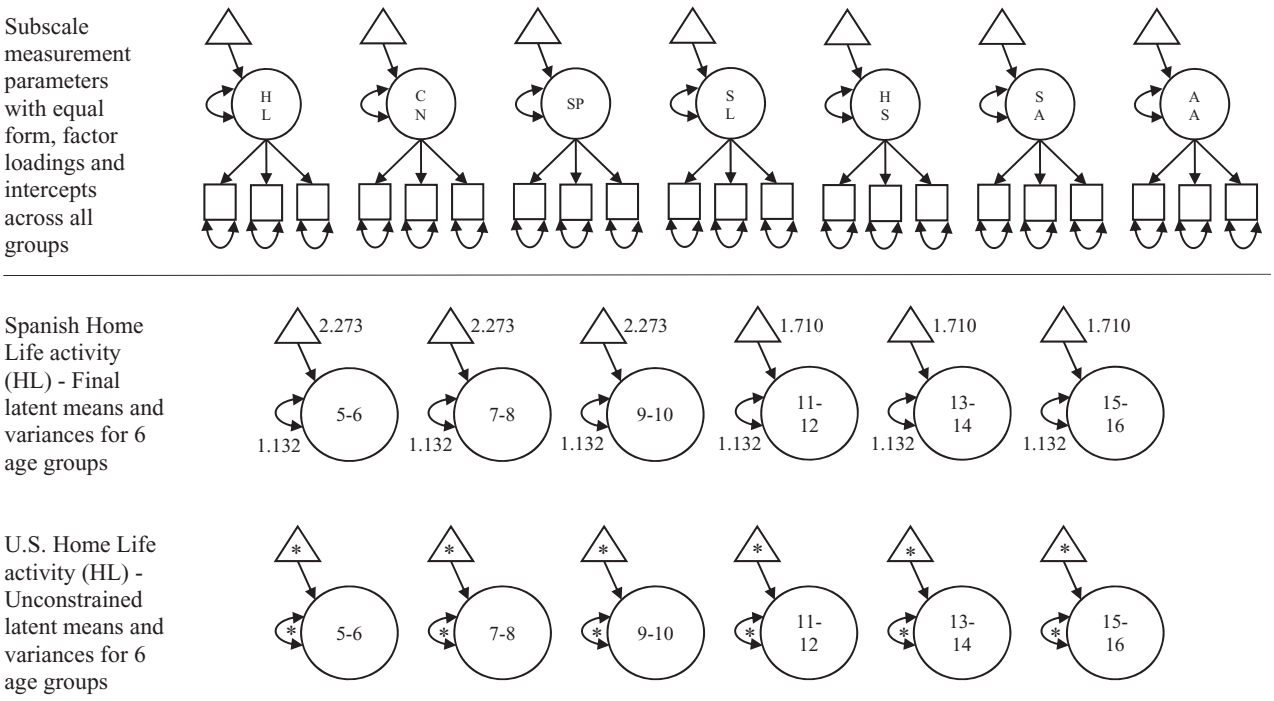

Figure 1. A single subscale measurement model with constrained Home Life latent means and variances for the Spanish age groups and unconstrained latent parameters in the US age groups. The subscale model freely estimated residuals, latent covariances, and all other latent means and variances across all twelve groups. HL = Home Life, CN = Community and Neighborhood, SP = School Participation, SL = School Learning, HS = Health and Safety, SA = Social Activities, and AA = Advocacy Activities.

Table 4. Standard score and standardized percentile ranks for home life activities construct.

		Home life 5–10-year-olds		Home life 11–16-year-olds	
Standard score	Standardized percentile rank	Raw score	Raw-score range	Raw score	Raw-score range
16	97.7			3.84	3.66–4.00
15	95.2		3.87–4.00	3.48	3.31–3.65
14	90.9	3.69	3.51–3.86	3.13	2.95–3.30
13	84.1	3.34	3.16–3.50	2.77	2.60–2.94
12	74.8	2.98	2.81–3.15	2.42	2.24–2.59
11	63.1	2.63	2.45–2.80	2.06	1.89–2.23
10	50.0	2.27	2.10–2.44	1.71	1.53–1.88
9	36.9	1.92	1.74–2.09	1.36	1.18–1.52
8	25.2	1.56	1.39–1.73	1.00	0.82–1.17
7	15.9	1.21	1.03–1.38	0.65	0.47–0.81
6	9.1	0.85	0.68–1.02	0.29	0.11–0.46
5	4.8	0.50	0.32–0.67		< 0.11
4	2.3	0.15	< 0.32		

Note. The highlighted row indicates the means of Home Life activities constructs in 5–10- and 11–16-year-olds and their corresponding standard scores, standardized percentile ranks, and raw-score ranges.

To examine whether the US data impacted latent means and variances of the Spanish standardization sample, we compared Spanish sample estimates between models with (12-group) and without (6-group) the US standardization sample. Life activity models were evaluated first by examining the factor loadings and intercepts from strong invariance models. The factor loading and intercept estimates differed slightly between the combined and Spanish-only models, with 28 of 42 (66.7%) factor loadings and intercepts from the 12-group model falling within the 95% confidence intervals for the 6-group model. Differences between the measurement parameters of the combined and Spanish-only models were expected because of the instability of the Spanish-only measurement model given the small sample size.

Additionally, the standard errors for the measurement parameters were systematically larger for the Spanish-only models. This result is unsurprising because standard errors are a function of sample size, so it is expected that the standard errors in the 12-group model (n = 4,465) would be smaller than those in the Spanish 6-group model (n = 450). For example, the standard errors for the three indicators of Home Life were 0.013, 0.015, and 0.012 in the 6-group model, and 0.006, 0.006, and 0.005 in the 12-group model. A simulation was designed to determine what sample size for Spain would have been needed to obtain standard errors of equal size. A population model was specified with the parameter estimates from the 6-group model and data were simulated and analysed in Mplus 7.31 (Muthén & Muthén, 2012). Results indicated that a sample size of 400 participants per group, for a total of 2,400, would be needed to obtain standard errors equal to those obtained from the 12-group model. As discussed above, collecting a Spanish sample of 2,400 participants was unfeasible, so these findings clearly demonstrate that leveraging the US norming sample was crucial for the outcome of the analyses reported in this article.

Interestingly, the simulation results indicated a smaller hypothetical sample for Spain than was modeled with the 12 groups (2,400 versus 4,465). The simulation was then repeated using the 12-group estimates for the population instead of the 6-group estimates. The results for this model indicated that a sample size of 425 per group (n = 2,550) was needed to reproduce standard errors equal to those

in the 12-group model. There could be a few reasons for these findings. One, simulated data are rarely as noisy as real data. Two, the US portion of the 12-group model may reflect a more heterogeneous sample than the Spanish sample, given the diversity of the US population. The results from the second simulation support that the Spanish sample is more homogenous (which may also relate to the finding that the age bands could be collapsed in norming the Spanish Transition, while not in the US norming sample), otherwise a sample less than that of the 12-group model would have been required to obtain standard errors of the same size.

The final subscale models used to generate norm tables were evaluated next (Spain only vs. Spain + US). Comparisons of latent means and variances across the models showed minor differences (< .02) between estimates in the 6- and 12-group models. When the two latent means and one variance in the 6-group models were fixed to the estimate obtained from the 12-group model, nested model testing indicated that none of the subscale models were different from the models where the latent constructs were freely estimated; χ^2 (3) ranged from 0.015 to 0.157 with p values that ranged from .98 to 1.00. These results indicate that the latent estimates from the 6- and 12-group models are not significantly different from each other.

These subscale findings were replicated in the overall support needs models with the exception that the 6-group model did not pass strong invariance testing because the change in comparative fit index (CFI) exceeded .01. By adding a correlation between Home Life and School Participation in the 15–16 age group, the change in CFI between the weak and strong models became acceptable (i.e., partial strong invariance). The correlated residual was not needed in the 12-group model. The partial strong invariance model was then estimated with fixed latent means and variance, using latent parameter estimates from the 12-group overall support needs model, nested model testing indicated that the fixed model was not significantly different from the freely estimated model, $\Delta\chi^2$ (3) = 4.804, p = .187. Although the measurement model differed by one parameter in one group, the differences in measurement model did not impact the latent constructs.

These sensitivity analyses clearly suggest that leveraging the US norming sample to help construct the Spanish norms has had the desired effect. Namely, the large amount of information in the US sample has stabilized the parameter estimates in the combined sample measurement model, but including the US data has not had any significant impact on the latent parameters that were actually used to construct the Spanish norms. This confirms that the process described in this article US data stabilized the overall model parameters and the inclusion of the US sample did not influence on the norms of the SIS-C Spanish Translation, and thus, can be used to norm translated versions of previously validated scales when the context precludes collecting a large standardization sample.

Discussion

The current article describes the development and application of a process to use an extended data set to generate unique norms for a translated version of a standardized scale. We demonstrate the suggested approach by describing the application of the norming process to the SIS-C Spanish Translation, where we leveraged a large US standardization sample to generate unique norms with a smaller Spanish standardization sample. The study findings suggest that the larger sample can be leveraged to generate norms, given the measurement level equivalence in the measurement of support needs of children with intellectual disability between the US and Spain. Further work on translated versions of the SIS-C will provide additional information on the performance of our proposed technique across a range of cultural contexts. We are currently repeating the international norming process in several other countries or regions (i.e., Catalonia, Iceland, Italy) where small standardization samples have also been generated following the same procedures described for the Spanish sample.

Considerations for use

When implementing this international norming process with other data sets or with other assessments, the following three issues must be considered in evaluating its appropriateness for the data collection procedures and analytic goals.

Samples from larger population with subgroups. For the SIS-C norming process described here, the sample was children aged 5 to 16 years with intellectual disability. The sample used in this study was subdivided based on country of origin (US vs. Spain), but the overall population was children with intellectual disability. In structuring groups based on the country, one group must have a large enough sample to conduct a stand-alone norm generation, and be representative of the population of interest (in this case the US population of children with intellectual disability). For example, the US sample was drawn from all geographic areas of the country, but did not extend outside national borders (Seo et al., 2016). Likewise, the Spanish sample, while smaller and dependent on the US sample for norm generation, was collected to be representative of the population of children with intellectual disability in Spain. Because the data differed by country, we were able to separate the observations into groups first based on country and then by country and age group. If there is overlap between the samples being considered, this approach may not be appropriate.

Strong invariance. As noted above, the technique we propose is only valid when strong invariance constraints are enforced. Without measurement invariance constraints, there is no benefit to including the large group standardization sample because doing so only reduces uncertainty in the estimated small group measurement parameters when measurement invariance constraints hold. If the MACS CFA models do not support full strong invariance, the goal becomes establishment of partial measurement invariance by freeing as few parameter estimates as possible before testing latent means and variances. The primary role that the US normative sample provides is model stabilization, and a sample that induces differences on a large number of factor loadings or intercepts may not add the model stability.

Sample size. Lastly, there is a constraint on how small the sample can be for the larger of the two samples utilized in the proposed process. The size of the larger of the standardization sample is not based solely on representativeness but is also dependent on the measure being standardized. MacCullum, Widaman, Zhang, and Hong (1999) conducted simulations to provide guidance with respect to sample size for CFA, and their results highlight that how many indicators load on a construct and the strength of factor loadings play a role in sample size requirements for any model. The stronger the factor loadings, the smaller the sample needed in the larger group. In addition, optimally, the measure that is being standardized was

previously validated. Validation information paired with the size of the smallest group can be used to determine the necessary size of the larger sample. As long as both samples are representative and the multiple group model with both the large and small samples has sufficient degrees of freedom—more observations than parameters being estimated—then this process can be used.

The normalization process for the SIS-C stratified data by age (and within age, disability characteristics) so the smallest number of observations needed was based on the number of parameters in the model and the number of observations in the smallest age group. A single group in the SIS-C CFA model consists of 69 parameter estimates, and the Spanish 9–10 age group contained 71 observations. By the end of the modeling process, the 9–10 age group no longer had 69 parameters being estimated because factor loadings, indicator intercepts, and other parameters have been equated across groups; however, the configural model did require sufficient group membership to estimate the model. Without more observations than parameters, the model could suffer from convergence issues, regardless of how large the extended sample is.

Conclusions

In this article, we demonstrated additional merits of utilizing the statistical strengths of MACS (Little, 1997) modeling to norm scales. Building on the clear advantages of MACS modeling outlined in Seo et al. (2016), we have demonstrated how to leverage the power of a large, pre-existing standardization sample from one population to facilitate establishing novel norms for a separate population that does not have the capacity to generate large enough sample sizes to derive norms independently. This capability is facilitated by measurement invariance constraints that increase the stability and precision of the overall models' parameters estimates. We believe that the proposed innovative use of additional populations to increase power and stability will be beneficial for future researchers who wish to establish norms for scales with the relatively small standardization sample size across diverse social cultural contexts.

Funding

The authors declared receipt of the following financial support for the research, authorship,and/or publication of this article: This study was supported in part by grant IES R324A120407 (Carolyn Hughes, James R. Thompson, and Michael L. Wehmeyer, co-PIs), by grant NSF 1053160 (Wei Wu and Todd D. Little, co-PIs), and by the Institute for Measurement, Methodology, Analysis, and Policy (Todd D. Little, Director) at Texas Tech University. Todd D. Little owns and receives remuneration from Yhat Enterprises, LLC, which runs educational workshops such as Stats Camp (statscamp.org), and processes his royalties and his fees for consulting on statistics and methods with life-science science researchers.

Supplemental material

Supplementary material for this article is available online.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational testing service.

Bartram, D. (2008). Global norms: Towards some guidelines for aggregating personality norms across countries. *International Journal of Testing*, *8*, 315–333.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Elosua, P., & Iliescu, D. (2012). Tests in Europe: Where we are and where we should go. *International Journal of Testing*, *12*, 157–175.

Lappalainen, K. M., Savolainen, H., Kuorelahti, M., & Epstein, M. H. (2009). An international assessment of the emotional and behavioral strengths of youth. *Journal of Child & Family Studies*, *18*, 746–753.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. G. M. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*, 151–162.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the item versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285–300.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99.

McDonald, R. P. (2011). *Test theory: A unified treatment*. New York: Routledge.

Meyer, G. J., Shaffer, T. W., Erdberg, P., & Horn, S. L. (2015). Addressing issues in the development and use of the composite international reference values as Rorschach norms for adults. *Journal of Personality Assessment*, *97*, 330–347.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles: Muthén & Muthén.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Schalock, R. L., Borthwick-Duffy, S., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., . . . Yeager, M. H. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Seo, H., Little, T. D., Shogren, K. A., & Lang, K. M. (2016). On the benefits of latent variable modeling for norming scales: The case of the Supports Intensity Scale – Children's Version. *International Journal of Behavioral Development*, *40*, 373–384.

Shogren, K. A., Seo, H., Wehmeyer, M., Hughes, C., Thompson, J., Little, T., & Palmer, S. (2015). Support needs of children with intellectual and developmental disabilities: Age-related implications for assessment. *Psychology in the Schools*, *52*, 874–891.

Tassé, M. J., & Thompson, J. R. (2010, June). *Supports Intensity Scale for Children Translation Guidelines*. Paper presented at the 134th Meeting of the American Association on Intellectual and Developmental Disabilities, Providence, RI.

Thompson, J. R., Wehmeyer, M. L., Hughes, C., Shogren, K. A., Seo, H., & Little, T. D., . . . TasséM. J. (2016). *Supports Intensity Scale – Children's Version: User's Manual*. Washington, DC: American Association on Intellectual and Developmental Disabilities.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67.