

Psychometric evaluation of the System Usability Scale in the context of a childrearing app co-designed for low- and middle-income countries

Qaisar Khan , Ian B Hickie, Victoria Loblay , Mahalakshmi Ekambareshwar ,
Iqthyer Uddin Md Zahed, Aila Naderbagi , Yun JC Song and Haley M LaMonica 

Abstract

Objective: The System Usability Scale (SUS) demonstrates good psychometric properties for a range of technologies; however, its reliability and factor structure in the context of a childrearing application (app) and variation across cultures remains largely unexplored. This study investigates the reliability and factor structure of the SUS in the context of a childrearing app that was co-designed for and implemented in diverse low- and middle-income countries (LMICs).

Methods: Parents and caregivers of children aged 0–5 years in five LMICs completed the SUS after having access to the app for a minimum of 2 and maximum of 24 weeks. Survey data from participants ($n = 668$) was analysed using exploratory and confirmatory factor analysis methods.

Results: The bi-factor model shows the best fit to data (comparative fit index = 0.998; Tucker Lewis Index = 0.996; standardised root mean square residual = 0.033). Further analysis suggests that usability and learnability subscales provide additional information not contained in the total SUS score. A two-sample *t*-test shows that younger caregivers, employed full- or part-time, and with fewer children reported significantly better app usability.

Conclusion: The SUS has good psychometric properties, and it is a valid and reliable tool for assessing the usability of mobile apps when used by parents and other caregivers for children's socioemotional and cognitive development. However, it is not essentially unidimensional and appears to have a multidimensional structure that could be specific to our context owing to variations in users' experience, culture, and language. The findings have implications for other mobile health interventions implemented in contexts with cultural and linguistic differences.

Keywords

Digital technology, system usability, parenting program, early childhood development, global health, mobile app, impact evaluation

Received: 14 November 2024; accepted: 1 April 2025

Introduction

The value of mobile health

The mobile health (mHealth) market has experienced remarkable growth reaching 325,000 applications (apps) worldwide¹ and is now a popular resource for caregivers to support children's feeding and nutrition,² prenatal care,³ and mental health.^{3,4} With the rise in digital technologies, more parents seek access to information to support their child's developmental progress in a meaningful way. However, finding relevant information and translating that

into parenting practices has often been challenging owing to variations in socioeconomic conditions, cultures, and parental preferences for the type of information sought

Youth Mental Health and Technology Team, Brain and Mind Centre, The University of Sydney, Sydney, NSW, Australia

Corresponding author:

Haley M LaMonica, Brain and Mind Centre, The University of Sydney, 1 King Street, Newton NSW 2042, Australia.
Email: haley.lamonica@sydney.edu.au



about child development. Indeed, parenting practices are dynamically influenced by cultural beliefs about caregiving and child development, such as the desired skills and behaviours for young children, expectations for when children should meet specific developmental milestones, and how, when, and by whom care should be provided to children.^{5,6}

Most parenting apps provide generic information and advice such as screening and tracking of developmental milestones and tracking of health promotion behaviours for children (e.g. nutrition, sleep).⁷ There are very few apps that support socioemotional development. The literature also highlights that the absence of cultural considerations in most mobile apps adversely affects the adoption and retention of technology and contributes to health disparities.⁸ Recently, some researchers have endeavoured to use cultural frameworks and translations of content in the design of culturally responsive mobile apps.⁹ However, in relation to the latter, the findings suggest that translations, if not culturally relevant, do not support technology uptake and adoption.

In this article, we empirically evaluate the usability of a new childrearing app – Thrive by Five – that had been co-designed and implemented in nine countries in Africa (Cameroon, the Democratic Republic of the Congo, Kenya, Namibia), Central Asia (Afghanistan, Kyrgyzstan, Uzbekistan), and Southeast Asia (Indonesia, Malaysia) to support caregivers and families with evidence-based and culturally appropriate information about early childhood development (ECD). The objective of Minderoo Foundation's Thrive by Five International Program was to empower parents and other caregivers globally with the knowledge they need to support healthy development of children during the first 5 years of life. The Youth Mental Health and Technology Team from the University of Sydney's Brain and Mind Centre (including this paper's authors) led the co-design and development of the content for Thrive by Five.^{10,11} The content is underpinned by a scientific framework that highlights key neurobiological systems that can be targeted behaviourally to promote healthy ECD.¹⁰ Local parents, other caregivers (e.g. grandparents), and subject matter experts (e.g. clinical psychologists, early childhood educators, medical specialists, anthropologists, linguists) were key stakeholders throughout the research and development processes, providing invaluable insights to inform the program's design and iterative refinement across diverse contexts. Specifically, all content for each country was developed, refined, and ultimately finalised in consultation with child caregivers and subject matter experts through an extensive and iterative co-design process conducted in partnership with in-country partners and local champions. As shown in Figure 1, the content, referred to as 'Collective Actions', is comprised of scientific information about ECD (i.e. "The Why") coupled with suggested activities that parents, extended family, and trusted members of the community can engage in with the child to promote various aspects of their socioemotional and cognitive development. The Thrive by Five app is

the flagship product of this program; however, the content is also disseminated via other digital (e.g. Whatsapp chatbot) and non-digital (e.g. print media) methods based on the needs of the users in each country where the program is implemented.

System Usability Scale

The System Usability Scale (SUS) is a widely used, freely available, standardised tool to assess the perceived usability of a wide range of products and systems. Usability is the perceived ease of use of a product, system, or interface to achieve a defined goal efficiently and satisfactorily.¹² Efficiency is how effortlessly a user can use a product and satisfaction is the level of comfort of using a product.^{13,14} Developed by Brook,¹⁵ the SUS tool was initially presented as 'quick and dirty' but over time it has proven itself to be a popular and effective measure of perceived usability.¹⁶ Indeed, the SUS has been instrumental in assessing thousands of computer- and web-based applications and is considered an industry standard.^{17,18} Indeed, a recent rapid review found the SUS was a commonly used method for evaluating the quality of mobile health apps in high-, low- and middle-income countries (LMICs).¹⁹ This includes mobile health apps specific to early childhood as exemplified by NeoTree, a clinical management and education mobile app designed to improve newborn care in resource-poor settings in Malawi.²⁰ Several factors make the SUS relatively more popular compared to other measures of perceived usability, such as its short length (10 items), reliability and validity, ease of administration, and low cost as well as the fact that it is technology agnostic.^{21,22}

The SUS is comprised of 10 statements relating to different aspects of usability (e.g. ease of use, need for support, confidence, consistency of features and functions) on a 5-point Likert scale ranging from strong disagreement to strong agreement (i.e. scale of 1–5) in relation to a product, application or system (Table 1). In addition, it has a mix of positively and negatively framed items, with the odd-numbered items having a positive tone and the even-numbered items having a negative tone. There is another structure of Learnability, consisting of statements 4 and 10, that was recently explored in the factor structure of SUS.¹⁷ The SUS score is computed by converting the scale from 0 (poorest rating) to 4 (best rating) with adjustment for odd-numbered (subtract 1 from the raw score) and even-numbered (subtract the raw score from 5) questions. The adjusted score is summed and multiplied by 2.5 to get the standard SUS score. The final SUS score ranges from 0 to 100, with 0 being extremely poor usability and 100 being excellent usability.^{16,21}

While the SUS is the most widely used tool for measuring app usability, various other scales have been documented in the literature concerning user experience, satisfaction, and the suitability of mobile and web-based applications. One such scale is the User Experience Questionnaire (UEQ)

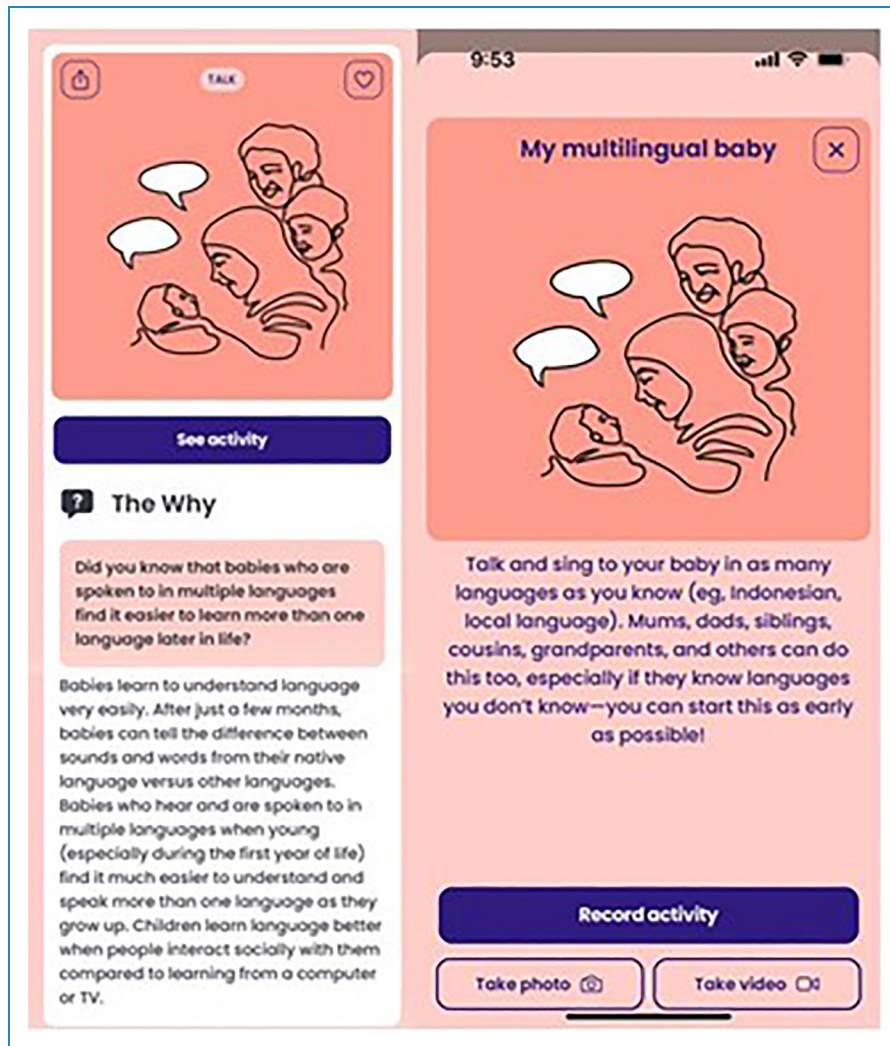


Figure 1. Example of a collective action: ‘the why’ and associated childrearing activities.

which provides a comprehensive assessment across six dimensions: attractiveness, perspicuity, efficiency, dependability, functionality, and information quality.²³ However, the UEQ was not relevant to our study due to the length of its questions and our focus on conducting a rapid and simple evaluation of the app’s basic usability.

The Usability Metric for User Experience (UMUX) serves as an alternative to SUS, concentrating on just two questions regarding usability and satisfaction.²⁴ The user experience behind UMUX aligns closely with the Technology Acceptance Model (TAM), which assumes that a user’s experience with new technology is based on its perceived ease of use and usefulness.^{25,26} However, TAM is mainly used to evaluate a product’s acceptability rather than its overall usability, which is what we aimed to assess using SUS.

Furthermore, the Technology Fit Model (TFM) evaluates the fit between technology and user needs.²⁷ TFM is often applied in the early stages of technology development

and selection, while SUS is typically utilised after users have gained experience with a product, as was the case in our study. Recent research emphasises the importance of integrating various frameworks to devise the most reliable and effective solutions for designing and implementing mobile and web-based applications.²⁸ However, integrating these frameworks was not appropriate in our study context given the specific scope of the program and the goal of assessing usability after the app’s implementation.

Psychometric properties of SUS

A range of research has been conducted on the psychometric properties of the SUS. The various features of the SUS that have been explored include its acceptable level of reliability,^{4,13} validity and sensitivity to different types of interfaces and changes to a product.²¹ From its original version in English, the SUS has been translated into and validated in many other languages such as Persian,¹³ Arabic,²⁹ Bahasa Indonesian,³⁰

Table 1. Description of individual statements in the System Usability Scale (SUS).

#	Statement ^a	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	I think that I would like to use Thrive by Five frequently	-	-	-	-	-
2	I found Thrive by Five unnecessarily complex	-	-	-	-	-
3	I thought Thrive by Five was easy to use	-	-	-	-	-
4	I think I would likely need the support of a technical person to be able to use Thrive by Five	-	-	-	-	-
5	I felt that various features and functions in Thrive by Five were well integrated	-	-	-	-	-
6	I thought there was too much inconsistency in Thrive by Five	-	-	-	-	-
7	I would imagine that most people would learn to use Thrive by Five very quickly	-	-	-	-	-
8	I found Thrive by Five very cumbersome to use	-	-	-	-	-
9	I felt very confident using Thrive by Five	-	-	-	-	-
10	I needed to learn a lot of things before I could get going with Thrive by Five	-	-	-	-	-

^aAdapted from Brooke.¹⁵

Malay,³¹ Italian,²² Polish,³² Slovene,³³ Portuguese,³⁴ and Danish³⁵ with reported reliabilities between 0.79 and 0.87.

Studies have shown both the average score and percentile ranking of the SUS score as acceptable standards to rate the usability of a product or application. Acceptable usability corresponds to a SUS score roughly above 68, with unacceptable usability indicated by scores below 50.^{18,21} Using a 7-point adjective rating scale (from worst imaginable to best imaginable) as an added eleventh question in the SUS questionnaire, Bangor et al.³⁶ provided grades (A to F) and incorporated acceptability ranges to the SUS score. SUS score ≥ 70 is considered ‘acceptable’ and rated ‘good’. Sauro and Lewis¹⁸ suggest reporting the SUS score as a percentile rank with the average score of 68 as the 50th percentile. That means a raw score above 68 is above average and below 68 is below average.

The recent trend in research shows a high inclination towards the dimensionality of the SUS – whether it provides a single score of usability and is unidimensional (one-factor) or is multidimensional. As the SUS has been used for various products and in various contexts, the findings on its dimensionality and factor structure are mixed.^{17,37,38} However, a large body of literature argues that the SUS assesses the single construct of usability, and findings suggest it is unidimensional. Most of this literature comes from straightforward contexts such as assessing the usability of hardware platforms and computer interfaces.³⁶

While the SUS has been widely applied to measure usability in mobile and web-based applications, its reliability and factor structure in the context of a childrearing app remains largely unexplored. This study aims to investigate the reliability and dimensionality of the SUS in the context of the Thrive by Five app in culturally diverse LMICs. It also explores potential factors to explain ‘usability’ and ‘learnability’ to inform future work in similar contexts.

Methods

Study design

In addition to co-designing and developing the app, the Brain and Mind Centre’s Youth Mental Health and Technology team was also responsible for conducting impact and process evaluations of the Thrive by Five International Program. As such, the evaluation study of the usability of the Thrive by Five app using the SUS reported in this paper was embedded within a larger multi-site mixed methods evaluation study of the program detailed in LaMonica et al.³⁹ The Thrive by Five Program had been implemented in nine LMICs at the time of writing; however, the evaluation had not been completed in Kenya or the Democratic Republic of the Congo. Additionally, limited app uptake and inconsistencies in the data collected from participants in Namibia and Cameroon raised notable

Table 2. Participating countries and research sites.

S.No	Organisation	Country
1	The Bayat Foundation	Afghanistan
2	The Indonesian Child Welfare Foundation	Indonesia
3	Roza Otunbayeva Initiative	Kyrgyzstan
4	Malaysian Association of Professional Early Childhood Educators	Malaysia
5	The Innovation Centre	Uzbekistan

concerns about the reliability and validity of responses to the SUS. As such, only data collected in Afghanistan, Indonesia, Kyrgyzstan, Malaysia, and Uzbekistan were included in this study.

Importantly, for each participating site (Table 2), the original SUS questionnaire was translated using international guidelines for cross-cultural adaptation to ensure consistency of meaning to the original version.⁴⁰ Specifically, the survey was translated from English into up to three local languages to enable participants to provide their responses in their preferred language. The survey translation process involved multiple steps: 1) translation by a professional translator; 2) translation reviewed and edited by a second translator; 3) any discrepancies or concerns discussed with original translator. For all languages that have NAATI accreditation, the translations were NAATI accredited. NAATI is the national standards and certifying authority for translators and interpreters in Australia (<https://www.naati.com.au/>).

Participants

Participants were parents and caregivers (e.g. grandparent, aunt, uncle, nanny) who were 18 years of age or older, self-identified as a caregiver for at least one child aged 5 years or younger, and had used the Thrive by Five app in a manner of their own choosing, either on their own smartphone or on a device shared with other family members or close friends. There were no specifications set as to the frequency of app use, time spent on the app, level of app engagement, or the number of Collective Actions completed. Participants were recruited through established networks and advertising mechanisms of the respective research sites in each country in which Thrive by Five was implemented.¹¹ The sites used the recruitment methods best suited to their community and context (e.g. emails, poster displays, paper-based and online internal news articles, handouts, digital advertisements on social media). All participants provided written informed consent for this study. Incentives to compensate participants

for their time were offered based on each site's national paid participation rates and methods of reimbursement as recommended by the local ethics practices and committee's advice.

Data collection

Participants were invited to complete the SUS questionnaire as part of a larger evaluation survey between 2 and 24 weeks after they began using the app in each respective country. The site provided participants with a web link providing them access to the impact evaluation survey via REDCap,^{41,42} a secure electronic data collection and management tool hosted at the University of Sydney. Paper-based surveys were also available and distributed by the Site Principal Investigator via post or in-person on an as needed basis for those participants who did not have reliable access to the internet or a personal smartphone. Basic demographic information (e.g. relationship to the child, sex, country of birth, language spoken at home, age, marital status) was collected for descriptive purposes. Participants were asked to complete the survey within one week of receiving it, with those completing it on paper returning the survey directly to the site. The data collection for this research occurred over 2 years (2022–2023). Notably, identifying information is not collected in the Thrive by Five app; therefore, it is not possible to link evaluation participants with their app usage data.

Statistical analysis

A pooled sample of five participating countries was used for analysis. Basic descriptive statistics (e.g. frequencies, percentages) were used to analyse demographic data. The sample was randomly divided into two halves using the split-half sample method for the robustness and generalisability of our findings.^{43,44} The two subsets of samples included 334 observations, which is consistent with the recommended criteria for exploratory factor analysis (EFA).^{45,46} EFA was conducted on the first half of the data while confirmatory factor analysis (CFA) was performed on the second half to collectively validate the robustness of our findings. Participants per item ratio for CFA calculated was 33 which aligns with the 5–10 participants per item suggested by previous studies.⁴⁷ Additional indices such as Chi-square, comparative fit index (CFI), Tucker Lewis index (TLI), and standardised root mean square residual (SRMR) were evaluated for model goodness of fit.

Several statistical analyses were performed to detect the factor structure of the SUS. EFA was performed to examine the underlying relationship among the items. As our context differs in geographical and demographic characteristics and when users have their first-time interaction with the app, EFA is desirable to understand the relationship of the

underlying variables. Bartlett's test of sphericity was performed to test whether our observed correlations differ significantly from the identity matrix. An identity matrix means that our variables were not correlated with each other. A significant Bartlett's test was an indication to carry out the factor analysis. While this test could have been affected by sample size, we also reported Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy.⁴⁸ A KMO value ≥ 0.70 is considered acceptable for factor analysis of the correlation matrix. A scree plot was used to determine the number of factors to retain in the EFA. All statistical analyses were conducted using RStudio (version 4.3.1) using *lavaan*⁴⁹ and *Psych*⁵⁰ packages. P-values less than 0.05 were considered as significant.

CFA was conducted to further assess and verify the factor structure of the variables and evaluate the construct validity of the SUS. In addition, CFA was used to check whether the SUS could be used as a single score, or if it is multidimensional. In relation to this, several statistical tests were used to determine the adequacy of various models' fit to the data. We evaluated four models including one-factor, two-factor (usability and learnability), tone structure (positive and negative statements), and bi-factor models to further explore the SUS factor structure. The rationale behind the evaluation of a two-factor and bi-factor model was to provide further insights on the dimensionality of SUS investigated in other contexts.^{4,17} A two-factor structure involves two distinct latent factors, and each observed variable is influenced by one of these factors.⁵¹ This structure explains variance in observed data through two underlying factors which limits its applicability to data where variance can be attributed to a general factor as well as specific factors. Bi-factor, on the other hand, allows for the examination of both general factor and specific factors simultaneously and offers an advantage when dealing with constructs that have both a broad overarching dimension (e.g. usability) and other secondary dimensions (e.g. usability and learnability).⁵² Because the data is collected on a 5-point Likert scale and is considered ordinal, we adjusted parameter estimates using weighted least squares mean and variance which is a recommended approach for the analysis of ordinal data.⁵¹ Accordingly, goodness-of-fit was assessed using CFI and TLI (each with a cutoff score ≥ 0.95 for an acceptable fit), and SRMR (cutoff score ≤ 0.08).⁵³

Following Dueber,⁵⁴ further advanced statistics were calculated to carefully decide on the dimensionality of the found structure. Omega was calculated to report a model-based estimate of internal reliability. While there is no pre-determined cutoff for Omega, studies recommend a value between 0.50 and 0.75 as an acceptable level of reliability. Further complimentary statistical indices calculated were Omega hierarchical (OmegaH), explained common variance (ECV), and percent of uncontaminated correlations (PUCs). Omega hierarchical or OmegaH reports the percentage of systematic variance in unit-weighted total scores

that can be attributed to the individual differences in the general factor.⁵⁵ The higher the OmegaH the more the general factor is the dominant source of systematic variance and the higher the likelihood of unidimensionality. ECV is the proportion of common variance attributable to the general dimension while PUC is the percentage of covariance terms which only reflect variance from the general dimension. PUC along with ECV influences the parameter bias of the unidimensional solution. When ECV and PUC > 0.70 , the relative bias will be slight, and the common variance can be regarded as essentially unidimensional.⁵⁵ Similarly, when PUC values are lower than 0.80, general ECV values greater than 0.60 and OmegaH > 0.70 suggest the presence of some multidimensionality is not severe enough to disqualify the interpretation of the instrument as primarily unidimensional.⁵⁶

Results

Demographic characteristics of respondents

A total of 668 participants were recruited through the research sites in the five participating countries including Afghanistan ($n=111$), Indonesia ($n=157$), Kyrgyzstan ($n=118$), Malaysia ($n=129$), and Uzbekistan ($n=153$).

Table 3 reports the pooled demographic characteristics of the participants across the five countries. All variables, except participant age and number of children, are categorical variables. The frequency (n) and percentages are reported for the categorical variables, while means and standard deviations are reported for continuous variables, such as age and number of children. Approximately 50% of the respondents had completed a degree or post-graduate qualification and 68% were employed full- or part-time. The average age of respondents was about 32 years with 79% female and 86% married. The last four rows report users' experience with the app. Daily and more than once daily use was reported by 48%, weekly use by 31%, and fortnightly and monthly use by 11% and 10%, respectively. It should be noted that this was the users' first interaction with Thrive by Five, and their experience is reported only in this context. For example, we did not collect data on user's prior experience with the app as it was implemented for the first time in the targeted countries. In this relation, we do not deal with the longitudinal aspects of usability evaluation –testing over time to take into consideration user's prior experience with the app.⁵⁷

Summary statistics of SUS statements

Table 4 reports the means of individual statements and the mean total SUS score, with the standard deviations presented in column 3 and range in column 4. The mean total SUS score is 65.07 (this is approximately the 50th percentile in our sample) which falls below the acceptable standard.

Table 3. Demographic characteristics of respondents (N = 668).

Variable	Label	n (%)
Age ^a	Age in years	32.55 (7.33)
Children ^a	Number of Children	2.55 (1.53)
Education level	Primary	16 (2.40)
	Secondary	295 (44.16)
	Degree	261 (39.07)
	Post-graduate	68 (10.18)
	Other	28 (4.19)
Employment status	Employed ^b	385 (57.63)
	Unemployed	72 (10.78)
	Home duties	145 (21.71)
	Other ^c	66 (9.88)
Marital status	Married	577 (86.38)
	Other ^d	91 (13.62)
Gender	Female	531 (79.49)
	Male	135 (20.21)
	Prefer not to answer	2 (0.30)
User's app experience	More than once daily	136 (20.36)
	Daily	187 (27.99)
	Weekly	205 (30.69)
	Fortnightly	71 (10.63)
	Monthly	69 (10.33)

^aMean (SD).^bEmployed full or part-time.^cfull-time carer, student, or unable to work due to personal or health reasons.^dDefacto relationship, separated, divorced, never married, other.

App: application.

Overall, respondents, except for statements 4 and 10 (measuring learnability), tended to disagree relatively more with the even-numbered statements (negatively worded) compared to the odd-numbered statements (positively worded). This could mean that Thrive by Five was relatively less quickly learnable for caregivers in our sample overall. However, examination of the results in relation to demographic characteristics (refer to Table 3) using univariate regression analysis found statistically significant differences in the total SUS score based on

age ($p < 0.001$), employment status ($p < 0.05$), number of children ($p < 0.01$), with young caregivers who were employed and had fewer children reporting higher SUS scores compared to their counterparts.

Psychometric outcomes

Consistent with the previous literature, the reliability coefficients show that the SUS has acceptable reliability as evidenced both by $\omega = 0.89$ and Cronbach's alpha $\alpha = 0.86$. The KMO value of 0.80 suggests sufficient evidence for factor analysis and Barlett's test statistics ($p < 0.001$) show that the variables are related and ideal for factor analysis. The Eigenvalues from the factor analysis are presented in the scree plot in Figure 2. The analysis presents a two-factor structure (two significant factors of the SUS statements). These findings show that the SUS in our case reflects participants' experience on two-factors rather than a single score of usability.

As a next step, the model fitness was investigated using CFA to see whether the SUS could be used as a single score or if there is a departure from essential unidimensionality. Four models were evaluated: one-factor, two-factor, tone structure, and bi-factor model. Table 5 shows the results of the four models. The goodness-of-fit indices suggest that one-factor does not have an acceptable fit with CFI and $TLI < 0.95$ and $SRMR > 0.08$. However, two-factor, tone structure, and bi-factor models show an acceptable fit according to the goodness-of-fit indices ($CFI \geq 0.95$, $TLI \geq 0.95$, $SRMR \leq 0.08$). Based on the fit indices, the bi-factor model shows the best fit compared to other models. Figure 3 shows a graphic representation of the fitted bi-factor model with a positive correlation between usability and learnability.

As reported previously, the tone structure has no practical and theoretical meaning in the context of the SUS and hence further analysis is not of interest. However, the two-factor structure concerning usability and learnability has practical implications and has been explored in the previous literature. Hence to model multidimensionality, the bi-factor model is used with a general factor directly influencing all manifest variables along with orthogonal specific factors (usability and learnability) additionally influencing distinct subsets of variables.

Following Reise et al.,⁵⁶ next we calculated advanced bi-factor indices such as PUC and ECV to decompose variance and determine the relative strength of the general factor versus specific factors and quantify the degree of unidimensionality. The calculated ECV is 0.60 meaning that 60% of the variation comes from the general dimension. The PUC is 0.36 and the Omega H value is 0.65. Given that the ECV and PUC values are less than 0.70, the common variance cannot be considered as essentially unidimensional. Collectively, these findings suggest that the presence of some multidimensionality cannot be eliminated and SUS cannot be completely regarded as essentially unidimensional.⁵⁶

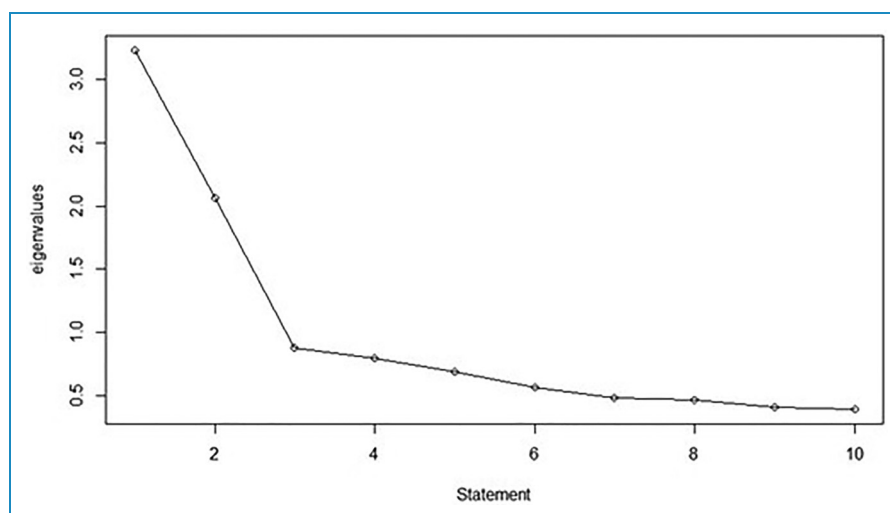


Figure 2. Scree plot from the factor analysis of SUS statements.
SUS: System Usability Scale.

Table 4. Summary statistics of individual statements and total SUS score.

Variable	Mean	Std.dev	Range ^a
SUS 1	4.02	0.75	1-5
SUS 2	2.51	0.97	1-5
SUS 3	3.95	0.64	1-5
SUS 4	3.98	1.03	1-5
SUS 5	3.91	0.80	1-5
SUS 6	2.53	0.98	1-5
SUS 7	3.89	0.76	1-5
SUS 8	2.47	0.98	1-5
SUS 9	3.91	0.76	1-5
SUS 10	3.99	1.03	1-5
Total Score	65.07	12.54	27.5–100

^aEach SUS statement ranges from strongly disagree (=1) to strongly agree (=5).

SUS: System Usability Scale.

Discussion

Mobile apps have made it possible to deliver evidence-based health and well-being information to a wider audience (e.g. parents, caregivers); however, to generate impact at scale, it is important for developers and researchers to continuously test, validate, and evaluate apps to provide optimised and culturally informed experiences for end

Table 5. Goodness of fit indices for four evaluated models.

Model	Chi-square	Df ^a	CFI ^b	TLI ^c	SRMR ^d
One-factor	390.5	35	0.716	0.635	0.137
Two-factors	139.8	34	0.949	0.950	0.069
Tone model	142.1	34	0.952	0.956	0.064
Bi-factor	27.1	24	0.998	0.996	0.033

^aDegree of freedom.

^bcomparative fit index.

^cTucker Lewis index.

^dstandardised root mean square residual.

users. In cross-cultural contexts, it is important that the reliability of evaluation instruments such as SUS is not assumed, but rather tested and validated. We explored the factor structure of the SUS in the context of a childrearing app for a diverse group of caregivers in five LMICs. Our findings indicate that the SUS has acceptable reliability and is a valid measure for mobile-based parenting apps aimed at supporting children's social, emotional, and cognitive development in diverse contexts. Notably, our findings show that the SUS is not essentially unidimensional, and that the subscales (learnability/usability) cannot be disregarded in our context. Various reasons could help explain this finding. Specifically, recent research shows the amount of user experience with a product as one of the possible determinants of the factor structure of the SUS.²² Their findings suggest the SUS is a one-factor scale when completed by users with less experience and a two-factor scale for users with more experience with the product.

The two-factor structure may also emerge when users spend more time with a product to acquire more learning

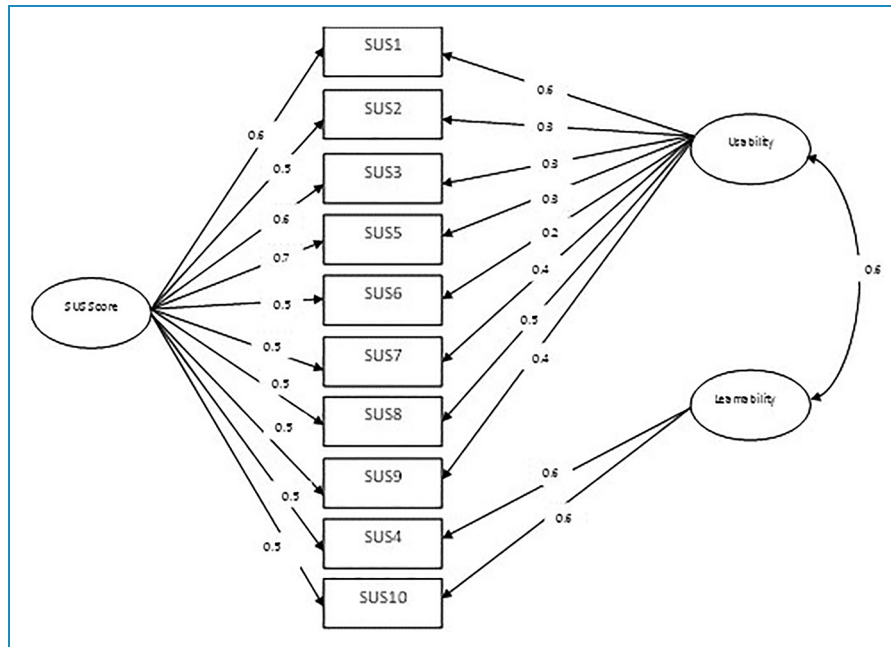


Figure 3. Factor structure of the bi-factor model.

(or what the previous research mentioned as learnability).⁴ As ECD and childrearing practices are influenced strongly by local cultures, there is potential for users to have difficulty aligning app contents with existing perspectives on childrearing, hence requiring more learning to close this gap. In other words, more time may be required for users to personally adapt the content to their own culture and context. Indeed, beyond new learnings from digital health interventions, such as Thrive by Five, parenting practices are ongoingly and dynamically influenced by a myriad of factors including cultural beliefs, personal experiences, economic and social status, personal goals and priorities, family dynamics, and social media.⁵ In such circumstances where a product or application serves as an educational platform, the learnability dimension may have more weight than usability particularly depending on how the new learnings align with other information sources or beliefs with regard to parenting.

Our heterogeneous sample contains large variations in users' experience with the app, education levels, and employment status. Almost half the caregivers reported use of the app either daily or more than once daily. This high frequency of use may be explained by users having to adapt to a new platform. Thus, it can be inferred that more learning was required to translate knowledge into positive caregiving practices. This pattern appears in the statement-wise summary statistics reported in Table 4. The table shows that a relatively higher proportion of the respondents reported learnability issues by agreeing with the statements that they needed the support of a technical person and that they needed to learn a lot of things before getting going with the app. There is also a relatively higher variation in respondents' responses to these two statements.

The total SUS score in our pooled sample falls below acceptable standards. As is explained above, a relatively higher tendency of participants toward learnability challenges could be one of the reasons for a low total usability score in our data. Further, examination of differences based on demographic variables allows us to draw additional inferences. Specifically, the statistically significant negative association between the age of the participant and usability in our sample may indicate that young users find the app relatively more usable compared to older users, a finding consistent with the previous research.²¹ We also found that parents and caregivers with relatively more children who were not employed full- or part-time reported poorer usability scores compared to their counterparts. For example, parents (particularly mothers) with more children are less likely to engage in paid employment,⁵⁸ may not have time to 'learn' the app due to finite levels of resources (time, energy, etc.) and therefore report poor usability. Collectively, these factors provide suggestive evidence of the underlying mechanisms behind the below acceptable usability score in our data.

Based on our analyses, the SUS appears to have a multi-dimensional structure and the subscales usability and learnability provide additional information that could be specific to our context – typical caregivers in LMICs with no or very limited exposure to the product and substantial variability in relation to education levels, employment status, age, and number of children. This has significant implications for interpreting SUS scores for other mHealth apps scaling in global settings where issues of learnability may need to be addressed to achieve acceptable usability. However,

there are some limitations that we would like to mention. The sample comes from culturally and linguistically diverse contexts that may have influenced variability (standard errors) in our reported estimates. As we mentioned earlier, the survey was translated from English into up to three local languages in participating countries to enable participants to provide their responses in their preferred language. The reliability (ω) coefficients range from 0.79 to 0.89 in five countries suggesting that SUS is a reliable tool measuring users' experience with Thrive by Five. However, we were unable to produce convincing findings on the factor structure of SUS by each participating country owing to small sample sizes that influence goodness-of-fit indices and may contaminate results.⁵⁹

Conclusion

With the universal endorsement of ECD in the 2030 sustainable development goals, there is a growing trend in the use of digital learning platforms to help parents and caregivers in LMICs address the risks of suboptimal development in young children. While technology has the potential to support the scaling up of ECD programs in diverse contexts, it entails challenges regarding culture, language, digital literacy, and training. It is therefore important to understand the usability of digital applications targeted toward ECD in specific contexts. In pursuit of this goal, it is imperative that evaluation tools for assessing these impact measures are scrutinised for validity and reliability, particularly when they are used in diverse contexts. In this article, we explore the reliability and factor structure of the SUS in the context of a childrearing app in five LMICs. We demonstrated that the SUS had good psychometric properties and is a valid and reliable tool for assessing the usability of mobile apps when used by parents and other caregivers for children's socioemotional and cognitive development. However, it appears to have a multidimensional structure that could be specific to contexts where there users' experiences, exposures, cultures, and languages are variable. The subscales of usability and learnability provide additional information that is not already contained in the total SUS score, highlighting the need to examine demographic, knowledge-based, cultural, or contextual factors that may make it more difficult for a user to 'learn' an app. Thrive by Five and other parenting applications interacting with cultures and requiring cultural adaptation need further inquiry ideally at the levels of individual countries or cultures to produce targeted findings for the designers and consumers in this discourse.

Acknowledgments

The authors would like to thank all in-country partners including the Bayat Foundation (Afghanistan), the Indonesian Child Welfare Foundation (Indonesia), Roza Otunbayeva Initiative (Kyrgyzstan), the Malaysian Association of Professional Early


Childhood Educators (Malaysia), and the Innovation Centre (Uzbekistan) for their assistance in facilitating local ethics and governance approvals as necessary and for recruiting parents and caregivers to participate in this research study. Additionally, we would like to thank all study participants who contributed their valuable, knowledge, and feedback. We also would like to thank the technology team at BBE for their efforts in developing, piloting, and testing the Thrive by Five app. Finally, we are very appreciative of our partner, Minderoo Foundation, for their support of this research.


Guarantor


IBH


ORCID iDs

Qaisar Khan  <https://orcid.org/0000-0002-9689-1666>

Victoria Loblay  <https://orcid.org/0000-0003-4094-9619>

Mahalakshmi Ekambareshwar  <https://orcid.org/0000-0003-1936-7120>

Aila Naderbagi  <https://orcid.org/0009-0003-9558-1142>

Haley M LaMonica  <https://orcid.org/0000-0002-6563-5467>

Statements and declarations

Ethical considerations

This study has been approved by the University of Sydney Human Research Ethics Committee (HREC) (Project 2021/956). Where a country-specific HREC exists, a site-specific protocol and supporting documents were submitted for local ethics approval, prior to initiating the research at the identified site. The Site Principal Investigator assisted in identifying the appropriate country-specific HREC as well as with the preparation and submission of ethics application as required. In instances where the country did not have a governing ethics body (e.g. Afghanistan), the approval from the University of Sydney applied as advised by the University of Sydney's HREC Office.

Consent to participate

All participants provided informed consent electronically prior to participating in this study.

Author contributions/CRedit

Authors, HL, YS, and IH were integral in securing funding to support the study. The impact evaluation study was designed by HL, YS, VL, and ME, with additional insights and support from QK, AN, and IZ. Scientific oversight and guidance were provided by IH to ensure all activities were conducted responsibly and in a culturally appropriate manner. QK was responsible for all data analyses and drafted the original manuscript. All authors contributed to and have approved the final manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was conducted by the University of Sydney's

Brain and Mind Centre pursuant to an agreement between the University and Minderoo Foundation Limited (Minderoo). Minderoo's Thrive by Five International Program targets parents and caregivers of children 0–5 years to support the cognitive, socioemotional development and well-being of young children across diverse cultures. IBH is supported by a NHMRC L3 Investigator Grant (GNT2016346).

Conflicts of interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: IBH is the Co-Director, Health and Policy at the Brain and Mind Centre (BMC) University of Sydney, Australia. The BMC operates an early-intervention youth services at Camperdown under contract to headspace. He is the Chief Scientific Advisor to, and a 3.2% equity shareholder in, InnoWell Pty Ltd which aims to transform mental health services through the use of innovative technologies.

Data availability

The data analysed during the current study are not publicly available to protect participant's privacy and confidentiality. However, R codes used for the analysis are available from the corresponding author upon reasonable request.

References

- Research 2 Guidance. mHealth App Economics 2017/2018. 2017.
- Taki S, Campbell KJ, Russell CG, et al. Infant feeding websites and apps: a systematic assessment of quality and content. *Interact J Med Res* 2015; 4: e18.
- Kraschnewski JL, Chuang CH, Poole ES, et al. Paging "Dr. Google": does technology fill the gap created by the prenatal care visit structure? Qualitative focus group study with pregnant women. *J Med Internet Res* 2014; 16: e147.
- Mol M, van Schaik A, Dozeman E, et al. Dimensionality of the System Usability Scale among professionals using internet-based interventions for depression: a confirmatory factor analysis. *BMC Psychiatry* 2020; 20: 218.
- Bornstein MH. Cultural approaches to parenting. *Parent Sci Pract* 2012; 12: 212–221.
- Chen X, Fu R and Yiu WYV. Culture and parenting. In: *Handbook of parenting*. United Kingdom: Routledge, 2019, pp.448–473.
- DeWitt A, Kientz J and Liljenquist K. Quality of mobile apps for child development support: search in app stores and content analysis. *JMIR Pediatr Parent* 2022; 5: e38793.
- Siek K, Veinot T and Mynatt B. Research opportunities in sociotechnical interventions for health disparity reduction. *arXiv preprint arXiv:190801035*. 2019.
- Armenta V, Warrell L, Nazneen N, et al. Actearly: a bi-national evaluation study of a mobile application for tracking developmental milestones. In: *Proceedings of the IX Latin American conference on human computer interaction*, Panama City, Panama, 2020, p.Article 1: Association for Computing Machinery.
- Crouse JJ, LaMonica HM, Song YJC, et al. Designing an app for parents and caregivers to promote cognitive and socioemotional development and well-being among children aged 0 to 5 years in diverse cultural settings: scientific framework. *JMIR Pediatr Parent* 2023; 6: e38921.
- LaMonica HM, Crouse JJ, Song YJ, et al. Developing a parenting app to support young children's socioemotional and cognitive development in culturally diverse low-and middle-income countries: protocol for a co-design study. *JMIR Res Protoc* 2022; 11: e39225.
- Interaction ITSEoH-S. Ergonomic requirements for office work with Visual Display Terminals (VDTs): Guidance on usability: International Organization for Standardization. 1998.
- Dianat I, Ghanbari Z and AsghariJafarabadi M. Psychometric properties of the Persian language version of the System Usability Scale. *Health Promot Perspect* 2014; 4: 82.
- Jordan PW. *An introduction to usability*. London: CRC Press, 2020.
- Brooke J. Sus: a "quick and dirty" usability". *Usability Evaluat Ind* 1996; 189: 189–194.
- Lewis JR. The System Usability Scale: past, present, and future. *Int J Hum Comput Interact* 2018; 34: 577–590.
- Lewis JR and Sauro J (eds). The factor structure of the System Usability Scale. In: *Human centered design: first international conference, HCD 2009, held as part of HCI international 2009, San Diego, CA, USA, 19–24 July 2009, Proceedings 1*: Springer.
- Sauro J and Lewis JR. *Quantifying the user experience: practical statistics for user research*. Massachusetts: Morgan Kaufmann, 2016.
- Woulfe F, Fadahunsi KP, Smith S, et al. Identification and evaluation of methodologies to assess the quality of mobile health apps in high-, low-, and middle-income countries: rapid review. *JMIR Mhealth Uhealth* 2021; 9: e28384.
- Crehan C, Kesler E, Nambiar B, et al. *G286 (P) The acceptability, feasibility and usability of the neotree application in Malawi: an integrated data collection, clinical management and education mHealth solution to improve quality of newborn care and thus newborn survival in health facilities in resource-poor settings*. London: BMJ Publishing Group Ltd, 2018.
- Bangor A, Kortum PT and Miller JT. An empirical evaluation of the System Usability Scale. *Int J Hum Comput Interact* 2008; 24: 574–594.
- Borsci S, Federici S, Bacci S, et al. Assessing user satisfaction in the era of user experience: comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *Int J Hum Comput Interact* 2015; 31: 484–495.
- Laugwitz B, Held T and Schrepp M (eds). Construction and evaluation of a user experience questionnaire. In: *HCI and usability for education and work: 4th symposium of the workgroup human-computer interaction and usability engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, 20–21 November 2008, Proceedings 4*: Springer.

24. Finstad K. The usability metric for user experience. *Interact Comput* 2010; 22: 323–327.
25. Davis FD. User acceptance of information systems: the technology acceptance model (TAM). 1987.
26. Davis FD. Technology acceptance model: TAM. In: Al-Suqri MN and Al-Aufi AS (eds) *Information seeking behavior and technology adoption*. Pennsylvania: IGI Global, 1989, pp.205–219.
27. Goodhue DL and Thompson RL. Task-technology fit and individual performance. *MIS Q* 1995; 19: 213–236.
28. Chuenyindee T, Montenegro LD, Ong AKS, et al. The perceived usability of the learning management system during the COVID-19 pandemic: integrating System Usability Scale, technology acceptance model, and task-technology fit. *Work* 2022; 73: 41–58.
29. AlGhannam BA, Albustan SA, Al-Hassan AA, et al. Towards a standard Arabic System Usability Scale: psychometric evaluation using communication disorder app. *Int J Hum Comput Interact* 2018; 34: 799–804.
30. Sharfina Z and Santoso HB (eds). An Indonesian adaptation of the System Usability Scale (SUS). In: 2016 international conference on advanced computer science and information systems (ICACSIS), 15–16 October 2016.
31. Marzuki MFM, Yaacob NA and Yaacob NM. Translation, cross-cultural adaptation, and validation of the Malay version of the System Usability Scale questionnaire for the assessment of mobile apps. *JMIR Human Factors* 2018; 5: e10308.
32. Borkowska A and Jach K (eds). Pre-testing of Polish translation of System Usability Scale (SUS). In: Information systems architecture and technology: proceedings of 37th international conference on information systems architecture and technology—ISAT 2016—Part I, 2017: Springer.
33. Blažica B and Lewis JR. A Slovene translation of the System Usability Scale: the SUS-SI. *Int J Hum Comput Interact* 2015; 31: 112–117.
34. Martins AI, Rosa AF, Queirós A, et al. European Portuguese validation of the System Usability Scale (SUS). *Procedia Comput Sci* 2015; 67: 293–300.
35. Hvidt JCS, Christensen LF, Sibbersen C, et al. Translation and validation of the System Usability Scale in a Danish mental health setting using digital technologies in treatment interventions. *Int J Hum Comput Interact* 2020; 36: 709–716.
36. Bangor A, Kortum P and Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. *J Usability Stud* 2009; 4: 114–123.
37. Borsci S, Federici S and Lauriola M. On the dimensionality of the System Usability Scale: a test of alternative measurement models. *Cogn Process* 2009; 10: 193–197.
38. Kortum P and Sorber M. Measuring the usability of mobile applications for phones and tablets. *Int J Hum Comput Interact* 2015; 31: 518–529.
39. LaMonica HM, Song YJ, Loblay V, et al. Promoting social, emotional, and cognitive development in early childhood: a protocol for early valuation of a culturally adapted digital tool for supporting optimal childrearing practices. *Digit Health* 2024; 10: 20552076241242559.
40. Sousa VD and Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract* 2011; 17: 268–274.
41. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009; 42: 377–381.
42. PH RT, Minor B, Elliott V, et al. The REDCap consortium: building an international community of software partners. *J Biomed Inform* 2019; 95: 103208.
43. Lorenzo-Seva U. SOLOMON: a method for splitting a sample into equivalent subsamples in factor analysis. *Behav Res Methods* 2022; 54: 2665–2677.
44. Fokkema M and Greiff S. *How performing PCA and CFA on the same data equals trouble*. Massachusetts: Hogrefe Publishing, 2017.
45. Tabachnick B and Fidell L. *Using multivariate statistics*. Boston: Pearson Education, Inc, 2007.
46. Field A. *Discovering statistics using IBM SPSS statistics: (and sex and drugs and rock'n'roll)*. Andy Field: Sage, 2013.
47. Kline RB. *Principles and practice of structural equation modeling*. 2nd ed. New York, NY, US: Guilford Press, 2005, pp.xviii, 366.
48. Kaiser HF. An index of factorial simplicity. *Psychometrika* 1974; 39: 31–36.
49. Rosseel Y. Lavaan: an R package for structural equation modeling. *J Stat Softw* 2012; 48: 1–36.
50. Revelle WR. psych: procedures for personality and psychological research. 2017.
51. Brown TA. *Confirmatory factor analysis for applied research*. New York: Guilford Publications, 2015.
52. Reise SP. The rediscovery of bifactor measurement models. *Multivariate Behav Res* 2012; 47: 667–696.
53. Hu L and Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model* 1999; 6: 1–55.
54. Dueber DM. Bifactor indices calculator: a Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. 2017.
55. Rodriguez A, Reise SP and Haviland MG. Applying bifactor statistical indices in the evaluation of psychological measures. *J Pers Assess* 2016; 98: 223–237.
56. Reise SP, Scheines R, Widaman KF, et al. Multidimensionality and structural coefficient bias in structural equation modeling: a bifactor perspective. *Educ Psychol Meas* 2013; 73: 5–26.
57. McLellan S, Muddimer A and Peres SC. The effect of experience on System Usability Scale ratings. *J Usability Stud* 2012; 7: 56–67.
58. Cools S, Markussen S and Strøm M. Children and careers: how family size affects parents' labor market outcomes in the long run. *Demography* 2017; 54: 1773–1793.
59. Bader M, Jobst LJ and Moshagen M. Sample size requirements for bifactor models. *Struct Equ Model* 2022; 29: 772–783.