A Multi-Sample Comparison and Rasch Analysis of the Evaluation of Children's Listening and Processing Skills Questionnaire

Sam Denys,[1,2] Johanna Barry,[3] David R. Moore,[4,5] Nicolas Verhaert,[1,2] and Astrid van Wieringen[1,6]

Objectives: Assessing listening difficulties and associated complaints can be challenging. Often, measures of peripheral auditory functions are within normal ranges, making clinicians feel unsure about proper management strategies. The range and nature of observed or experienced difficulties might be better captured using a qualitative measure. The Evaluation of Children's Listening and Processing Skills (ECLiPS) questionnaire was designed to broadly profile the auditory and cognitive problems often present in children with listening difficulties. This 38-item questionnaire was initially standardized in British children aged 6 to 11 years, was subsequently modified for use with North-American children, and was recently translated into Flemish–Dutch. This study aimed to compare typical scores of the Flemish version with the UK and US versions, and to evaluate and compare its psychometric quality based on Rasch analysis.

Design: We selected 112 Flemish children aged 6 to 11 years with verified normal hearing and typical development, and asked two caregivers of every child to fill out the ECLiPS. Data from two comparator samples were analyzed, including responses for 71 North-American children and 650 British children. Typical values for ECLiPS factors and aggregates were determined as a function of age and gender, and meaningful differences across samples were analyzed. Rasch analyses were performed to evaluate whether ECLiPS response categories work as intended, and whether item scores fit a linear equal interval measurement scale that works the same way for everyone. Item and person metrics were derived, including separation and reliability indices. We investigated whether items function similarly across linguistically and culturally different samples.

Results: ECLiPS scores were relatively invariant to age. Girls obtained higher scores compared with boys, mainly for items related to memory and attention, and pragmatic and social skills. Across ECLiPS versions, the most pronounced differences were found for items probing social skills. With respect to its psychometric quality, ECLiPS response categories work as intended, and ECLiPS items were found to fit the Rasch measurement scale. Cultural differences in responses were noted for some items, belonging to different factors. Item separation and reliability indices generally pointed toward sufficient variation in item difficulty. In general, person separation (and reliability) metrics, quantifying the instrument's ability to distinguish between poor and strong performers (in a reproducible manner), were low. This is expected from samples of typically developing children with homogeneous and high levels of listening ability.

Conclusions: Across the languages assessed here, the ECLiPS caregiver questionnaire was verified to be a psychometrically valid qualitative measure to assess listening and processing skills, which can be used to support the assessment and management of elementary school children referred with LiD.

Key words: Auditory processing, Children, Listening, Questionnaires.

Abbreviations: ANOVA = analysis of variance; CELF-IV-NL = Clinical Evaluation of Language Fundamentals, fourth version, Dutch edition; CPC = category probability curve; DIF = differential item functioning; DTT=digittriplettest;EAS=environmentalandauditorysensitivity;ECLiPS= Evaluation of Children's Listening and Processing Skills; ECLiPS-FL = Evaluation of Children's Listening and Processing Skills, Flemish version; ECLiPS-US = Evaluation of Children's Listening and Processing Skills, American version; ECLiPS-UK = Evaluation of Children's Listening and Processing Skills, British version; EMT = één-minuut-test; ICC = item characteristic curve; LAN = language; LiD = listening difficulties; LIS = listening; LLL = language, literacy, and laterality; MA = memory and attention; MANOVA = multivariate analysis of variance; MNSQ = mean of the squared residuals; N = number; PPVT-III-NL = Peabody Picture Vocabulary Test, third version, Dutch edition; PSS = pragmatic and social skills; SAP = speech and auditory processing; SNR = signal to noise ratio; SOC = social; SRT = speech-reception threshold; Tea-Ch = Test of Everyday Attention for Children; WISC-III-NL = Wechsler Intelligence Scale for Children, third version, Dutch edition.

(Ear & Hearing 2024;45;1202–1215)

INTRODUCTION

For some children, listening is challenging. Caregivers may perceive these children to struggle with hearing and understanding speech. The experienced or reported listening difficulties (LiD) are often judged by ENT-specialists or audiologists to be unrelated or disproportionate to hearing thresholds, or any other peripheral measures of hearing (Hind et al. 2011; Hunter et al. 2021). Adequate bottom-up auditory processing and speech decoding are undoubtedly essential for successful listening but, more importantly, active listening is mediated by top-down cognitive processing, including memory, attention, and language (Moore 2012; Tomlin et al. 2015; de Wit et al. 2016; Roebuck & Barry 2018; Stavrinos et al. 2018; Dillon & Cameron 2021). Recent neuroimaging studies support this view (Farah et al. 2014; Stewart et al. 2022; Alvand et al. 2023), and weaker or impaired language, reading, and

[1]University of Leuven, Department of Neurosciences, Research Group Experimental Otorhinolaryngology (ExpORL), Leuven, Belgium; [2]University Hospitals of Leuven, Department of Otorhinolaryngology - Head and Neck Surgery, Multidisciplinary University Center for Speech-Language Pathology and Audiology, Leuven, Belgium; [3]Otorhinolaryngology - Head and Neck Surgery, Nottingham University Hospitals National Health Service Trust, Nottingham, United Kingdom; [4]Communication Sciences Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA; [5]Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, United Kingdom; and [6]Department of Special Needs Education, University of Oslo, Oslo, Norway.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

0196/0202/2024/455-1202/0 • Ear & Hearing • Copyright © 2024 The Authors. Ear & Hearing is published on behalf of the American Auditory Society, by Wolters Kluwer Health, Inc. • Printed in the U.S.A.

<zdoi; 10.1097/AUD.0000000000001509>

social-emotional abilities have been observed in children with LiD (Sharma et al. 2009; Ferguson et al. 2011; Kreisman et al. 2012; Ahmmed et al. 2014; Moore et al. 2018; Ahmmed 2020; Petley et al. 2021).

LiD and associated complaints are an expression of many neurodevelopmental disorders, and not only a symptom of a pure (central) auditory processing deficit (Moore et al. 2010; DeBonis 2015; de Wit et al. 2018; Seeto et al. 2021). When there are concerns about the listening abilities of a child, a holistic approach is required to ascertain qualitative differences among children with overlapping symptoms, and broadly profile their cognitive strengths and weaknesses. This approach might give rise to more targeted referrals, diagnostic evaluation, and management (Moore et al. 2018).

However, the assessment of LiD is not trivial. Formal auditory and cognitive behavioral test batteries might be insensitive to the listening and processing difficulties experienced by children in their daily lives (Ahmmed & Ahmmed 2016; Magimairaj et al. 2020). In the absence of appropriately targeted tests, the range and nature of these difficulties are presumably better captured using a psychometrically sound (reliable and valid) qualitative measure (Moore et al. 2013). The Evaluation of Children's Listening and Processing Skills (ECLiPS; Barry et al. 2015; Barry & Moore 2021; Petley et al. 2021) was developed to address the need for a questionnaire that measured the broad range of auditory and cognitive problems displayed by children referred with LiD. The scale consists of 38 simply formulated statements of behaviors (items) that caregivers of the child can observe. These items belong to five distinct subscales (factors)—speech and auditory processing (SAP), memory and attention (MA), language, literacy, and laterality (LLL), pragmatic and social skills (PSS), and environmental and auditory sensitivity (EAS). Scores are then averaged to provide a profile of a child's listening (LIS = [SAP + MA + PSS]/3), language (LAN = [MA + LLL]/2), social (SOC = [PSS + EAS]/2), and total (TOT = average of all factor scores) abilities. Example statements are: "when there is a sudden noise, is confused about where to look" (SAP, item 5), "needs strategies for remembering" (MA, item 4), "needs help from others when writing homework down" (LLL, item 20), "says some sentences or words over and over" (PSS, item 23), and "finds it tiring being in groups of people" (EAS, item 3). Caregivers need to indicate to what extent they agree or disagree with these statements on a five-point Likert scale, ranging from "totally disagree" to "totally agree," with a neutral category in the middle of the response scale.

The items comprising a questionnaire aim to measure specific latent/hidden traits meaningfully and rigorously. In the context of the ECLiPS questionnaire, listening ability is viewed as a multidimensional construct. Items are summed and averaged to result in factor, aggregated, and total scores to quantify ability levels along different dimensions related to listening.

Most developers have applied classical test theory when assessing and improving the psychometric properties of their measuring instruments. Classical test theory is based on the idea that a person's obtained score reflects the person's true ability and some degree of measurement error, and makes no assumptions about responses for individual items (Tesio 2012). Two different people with the same score are assumed to have the same ability level, even when their answers vary across different items assessing the same latent trait (Stemler & Naples 2021).

Item response theory, on the other hand, states that two different people with the same score are said to have the same ability level, only if they provide similar responses to the same items.

Items vary in their level of difficulty. Easy items are more likely to be observed or endorsed. Difficult items are less likely to be observed or endorsed. The probability that a person obtains a certain score on a specific item is governed by the person's ability and the item's difficulty.

Next to item difficulty, item response theory also models item discrimination, which represents how well items discriminate between ability levels, with scores for more discriminating items weighing more heavily in the estimate of the ability level. It is about building a predictive statistical model that explains as much variance in the data as possible (Stemler & Naples 2021). Latent traits are measured through inferences from the obtained scores: raw scores are transformed into measures based on response probabilities. The ECLiPS questionnaire was designed in accordance with the principles of good scale development to obtain a valid and reliable measurement scale that is easy to understand for respondents (Barry & Moore 2021).

Its original development started with a large pool of 168 items, which was iteratively refined by rejecting redundant or uninformative items to the current set of 38 items forming the five subscales of the questionnaire (SAP, MA, LLL, EAS, and PSS). Item response theory modeling supported this process. Item information curves, showing the range of ability levels which an item is most sensitive to, were used to create a scale that was sensitive to a broad range of underlying abilities in each of the five subscales assessed by the ECLiPS.

In this study, we applied a combination of classical test and Rasch modeling to evaluate the psychometric properties of the ECLiPS. Item response theory and Rasch modeling differ in their assumptions and perspectives, but not their mathematical formulation. Rasch modeling assumes that ability levels are normally distributed. As a consequence, whereas people obtaining the same score are said to have the same ability level (classical test theory), differences between scores have a different meaning at different locations on the scale (Tesio et al. 2023c). A key property of Rasch modeling is that person ability estimates are independent of which items are used to probe the level of ability (separability theorem). Like item response theory, Rasch modeling makes assumptions about item difficulty. More precisely, Rasch modeling states that there is a ranking in item difficulty, and this ranking is invariant of ability level. More challenging items are more challenging for all people, and people with higher ability are able to respond correctly to more difficult items. Rasch fit statistics evaluate whether a linear structure is observed in the data or whether there are discrepancies between expected performance on the items (based on ability level) and observed performance on these items. Item difficulty is also assumed to follow a normal distribution, and is placed on the same linear scale as person ability. Contrary to item response theory, Rasch modeling requires all items to be equally discriminating. The goal is not to build a model that best predicts observed scores (item response theory), but to evaluate whether observed scores fit a linear equal interval measurement scale that works the same way for everyone (Embretson & Reise 2013; De Ayala 2022).

The present study aimed to evaluate the measurement properties of the ECLiPS using Rasch analysis. Rasch analysis facilitates the detection of item bias or local dependence which may be easily overlooked using more traditional validation methods (Müller 2020). Other advantages of Rasch analysis are the ability to decrease the number of items on a scale without affecting its psychometric properties, and to pool data drawn from different samples (Smith et al. 2008). These advantages have obvious clinical benefits, such as a more time efficient and valid questionnaire.

The ECLiPS questionnaire was initially standardized in a relatively large sample of British children aged 6 to 11 years. It has subsequently been modified with respect to the order of items for use with North-American populations (Petley et al. 2021), and was recently translated to Flemish/Dutch with the ultimate goal of incorporating the instrument into routine clinical practice to support the assessment of children referred with LiD. As part of achieving this goal, this article describes our work to standardize the Flemish ECLiPS (ECLiPS-FL), and to compare typical scores and psychometric properties, based on Rasch analysis, with the original (ECLiPS-UK) and US (ECLiPS-US) versions.

MATERIALS AND METHODS

Participants

Between May 2015 and October 2021, we recruited 135 typically developing Flemish children with normal hearing thresholds aged between 6; 00 and 11; 11 years old. We asked two caregivers of every child to independently fill out the ECLiPS-FL questionnaire (on paper or digitally), and invited every child for up to three testing sessions (at home or at school). Typical development (*z*-score >−2 on all (sub)tests of an elaborated behavioral neurocognitive assessment) and normal hearing sensitivity (bilateral pure-tone average$_{250\text{-}8000\text{Hz}}$ <20 dBHL) were checked. Data from children with missing questionnaires (N = 5), caregiver-reported developmental concerns or cognitive diagnoses (such as language-, learning- or attention disorders; N = 10), and/or very poor performance on any of the cognitive tests (N = 8) were excluded. The total Flemish sample consisted of 112 children (Table 1).

ECLiPS data from two comparator samples were analyzed (Table 1): a convenience sample of N = 71 typically developing normal hearing American-English children aged 6 to 11 years (ECLiPS-US; Petley et al. 2021) and a population sample of N = 650 British children within the same age range, that is, the original ECLiPS standardization sample (ECLiPS-UK; Barry et al. 2015; Barry & Moore 2021). This study was approved by the Ethics Committee of the University Hospitals of Leuven.

TABLE 1. Sample sizes as a function of gender and age for the different ECLiPS samples

	ECLiPS-FL		ECLiPS-US		ECLiPS-UK	
Age, yrs; mos (yrs)	Boys	Girls	Boys	Girls	Boys	Girls
6; 00–6; 11 (6)	5	9	3	9	48	52
7; 00–7; 11 (7)	16	12	6	8	65	64
8; 00–8; 11 (8)	7	7	3	9	66	64
9; 00–9; 11 (9)	11	12	4	4	53	49
10; 00–10; 11 (10)	10	5	7	7	44	71
11; 00–11; 11 (11)	8	10	6	5	43	31
Total	57	55	29	42	319	331

ECLiPS, Evaluation of Children's Listening and Processing Skills.

Materials and Procedures

Caregiver-Report Measures • Two caregivers for each child filled out the ECLiPS-FL separately. Either one or both parents completed the ECLiPS-FL for ~72% of the children in the study sample. We do not have information on the relationship of the caregivers to the child for ~28% of the sample. The caregivers also filled out a "background questionnaire" providing information on birthday, age, and gender of the child, relationship of the respondent to the child, history of ear and hearing problems, and concerns or diagnosis of hearing, language, learning, attentional, and/or social-emotional problems.

Behavioral Test Battery • A behavioral test battery was completed, including pure-tone audiometry, a digit triplet speech-in-noise screening test (DTT), and various cognitive tests (Table 2).

Pure-tone audiometry was performed using a portable Natus Madsen Midimate 622 audiometer connected to either a calibrated TDH-39 or a Sennheiser HDA-200 headphone. Hearing thresholds for octave frequencies between 250 and 8000 Hz were measured for both ears.

In a subset of our sample, bottom-up speech perception in noise ability was measured using an operator-controlled Flemish DTT (Jansen et al. 2013; Denys et al. 2021) with binaural antiphasic presentation of digit triplets in phasic speech-shaped noise (De Sousa et al. 2020) using a 7-inch Samsung Galaxy tablet connected to a calibrated DD65 headphone. The noise level was fixed at 70 dB SPL. Triplet levels were adaptively varied with fixed steps of 2 dB for the first six trials according to triplet scoring, and with variable step sizes from trial seven onward according to digit scoring (Denys et al. 2019). A test consisted of 17 trials, and started at a signal to noise ratio (SNR) of −8 dB SNR. Speech-reception thresholds (SRT) were calculated as the mean SNR$_{\text{trial7-i18}}$. A training test with 12 trials was administered to familiarize the children with the test procedure.

Cognitive abilities were measured using standardized tests for which Flemish normative values are available. Tasks were administered and scored according to instructions provided in the instruments' manuals. The test battery included a (nonverbal) intelligence test (block patterns; Wechsler Intelligence Scale for Children, third version, Dutch edition [WISC-III-NL, Kort et al. 2005]), a short-term auditory and working memory test (digit span forward and backward; Clinical Evaluation of Language Fundamentals, fourth version, Dutch edition [CELF-IV-NL, Kort et al. 2008]), reading tests (including real- [één-minuut-test {Brus & Voeten 1973}] and pseudowords KLEPEL {Van den Bos et al. 1999}), subtests from the Test of Everyday Attention for Children (TEA-Ch, Manly et al. 2004) measuring attentional abilities, and (receptive and expressive) language tests (Peabody Picture Vocabulary Test, third version, Dutch edition [PPVT-III-NL, Schlichting 2005], and subtests from the CELF-IV-NL).

The number of testing sessions varied from one to three, largely depending on the age of the child (more sessions for younger children to avoid fatigue). Also, due to updates with respect to the protocol over time, sample sizes varied across tests (Table 2).

Data Analysis

ECLiPS Typical Scores and Multi-Sample Comparisons • ECLiPS questionnaires were scored, providing raw factor (SAP, MA, LLL, PSS, and EAS) and aggregate (LIS, LAN, SOC, and

TABLE 2. Overview of cognitive tests and obtained scores (expressed as mean and median *z*-scores) for the Flemish sample

Cognitive Tests	N	Mean *z*-Score (SD)	Median *z*-Score (IQR)
Language tests			
PPVT-III-NL (receptive vocabulary)	54	0.18 (0.57)	0.20 (0.65)
CELF-IV-NL (general language ability)			
Concepts and following directions (receptive semantics and memory)	112	0.24 (0.71)	0.33 (1.33)*
Word structure (expressive morphology)	22	−0.09 (0.75)	0.00 (0.92)
Recalling sentences (receptive syntax and memory)	54	0.25 (0.84)	0.00 (1.00)
Formulated sentences (expressive syntax and semantics)	54	0.02 (0.69)	0.00 (1.33)
Word classes (semantics)	32	0.15 (0.73)	0.17 (1.00)
Expressive	90	0.22 (0.91)	0.33 (1.33)
Receptive			
Reading tests†			
EMT (real words)	49	0.03 (0.87)	0.00 (1.33)
KLEPEL (pseudowords)	49	0.37 (0.85)	0.33 (0.67)
Nonverbal intelligence test			
WISC-III-NL			
Block design	54	0.64 (0.84)	0.33 (1.33)*
Memory tests			
CELF-IV-NL			
Number repetition	112	0.28 (0.92)	0.33 (1.33)*
Forward (short-term auditory memory)	112	0.35 (0.87)	0.33 (1.33)*
Backward (working memory)			
Attention tests			
TEA-Ch			
Sky search (selective attention)	54	0.14 (0.67)	0.17 (0.67)
Score (sustained attention)	54	0.02 (0.73)	0.00 (1.00)
Walk, don't walk (attention switching)	49	0.61 (0.88)	0.67 (1.33)*

**Significant differences from 0 according to the nonparametric Wilcoxon rank test.*

†From 7 yrs onward.

CELF-IV-NL, Clinical Evaluation of Language Fundamentals, fourth version, Dutch edition; EMT, één-minuut-test; IQR, interquartile range; PPVT-III-NL, Peabody Picture Vocabulary Test, third version, Dutch edition; Tea-Ch, Test of Everyday Attention for Children; WISC-III-NL, Wechsler Intelligence Scale for Children, third version, Dutch edition.

TOT) scores, ranging between −2 and +2. Distributions of raw scores are shown in Supplementary Figure S1 in Supplemental Digital Content, http://links.lww.com/EANDH/B384. Scores were log-transformed (after adding 3 to all scores to bring them above 0) before statistical analyses to remediate skewness in the data. First, repeated-measures analyses of variance (ANOVA) were performed to investigate inter-rater agreement on ECLiPS-FL scores (separate analyses for factor and aggregate scores to avoid multicollinearity). Differences among raters for TOT scores were investigated with a paired-samples *t* test. These analyses were backed-up with one-way intraclass correlational analyses of absolute agreement. Next, for every child, the log-transformed scores were averaged across respondents (for ECLiPS-FL), and subjected to ANOVA (or MANOVA in case of multiple dependent variables, e.g., factor and aggregate scores) analyses, to investigate the effects of age and gender. Similar analyses were performed with sample (FL, US, or UK) added as a fixed factor to evaluate multi-sample comparability. These analyses were conducted using Jamovi (version 2.2.3.0) open-source statistical software.

Rasch-Based Psychometric Evaluation, and Multi-Samples Comparison • The Rasch model states that the probability of a person (the caregiver in case of the ECLiPS) agreeing with an item is a logistic function of the difference between a person's ability (the child's level of ability judged by the caregiver in case of the ECLiPS) and the item's difficulty (the level of ability measured by the item). Both person ability and item difficulty are estimated using log-odds or logits, and are placed on a linear scale with equal intervals (Boone 2016). The higher a child's estimated ability level relative to an item's difficulty level, the higher the probability that a respondent will agree. It is important to note that item difficulty is independent of the sample, and person ability is independent of the items (parameter separation), according to the Rasch measurement model (Bond & Fox 2001).

We applied Rasch analysis to evaluate the internal structure of the ECLiPS, investigating the extent to which items, rating scale categories, and people coalesce to form a measure that adheres to the requirements of the Rasch measurement model, and the extent to which Rasch metrics are reproducible and invariant across samples. In describing our findings, we adhered to the recently published RULER guideline (Mallinson et al. 2022; Van de Winckel et al. 2022).

Rating scale model, software, and data selection and preparation. For all ECLiPS versions, per factor items were evaluated against Andrich's Rating Scale Model, suitable for scales with multiple response categories per item, and an equal number of response categories across items (Andrich 1978), such as the ECLiPS, using WINSTEPS (version 5.2.0) software.

ECLiPS-FL data from one respondent per child were chosen. Response sets were pseudo-randomly selected to achieve a maximally balanced distribution of responses from mothers, fathers, and other (or unknown) caregivers as a function of the children's age and gender. For US and UK samples, only one response set was available (with the relationship of the respondent to the child being unknown). Before Rasch analyses, ECLiPS item scores were converted to scores ranging from 1 to 5 (with higher scores indicating higher ability).

Verification of Rasch assumptions. First, core Rasch assumptions of unidimensionality (i.e., items should measure one latent trait) and local independence (i.e., correlations between items are captured by the latent trait) were verified.

Rating scale category structure. Second, we investigated whether the rating scale categories work as intended: higher scores should represent higher listening ability or require higher listening ability levels. Mean category difficulty (i.e., the average model-estimated ability level of people selecting a given category across all items, in logits) should be ordered. The boundaries (or thresholds, in logits) between categories should rise as well. However, disordered thresholds and submerged (i.e., snowed under) categories can coexist with ordered categories, especially when certain categories are seldomly chosen or when respondents have difficulty discriminating between response options (Tesio et al. 2023a, b). Submerged categories were inspected using category probability curves (CPC; i.e., the probability of observing a response category as a function of the difference between latent trait ability and item difficulty).

Quantitative and qualitative assessment of data-model fit. Third, item fits were evaluated. Items with mean square (MNSQ) fit statistics >1 can be interpreted as demonstrating more variation between the Rasch measurement model-estimated and observed scores, and are said to underfit the model. Conversely, an item with a fit statistic <1 would indicate less variation than predicted, and is said to overfit the model (Bond & Fox 2001). Both inlier-sensitive (infit-MNSQ) and outlier-sensitive (outfit-MNSQ) fit statistics were analyzed. Infit is more sensitive to the pattern of responses to items with difficulties more targeted to person ability. Outfit is more sensitive to the pattern of responses to items with difficulties less targeted to person ability. High infit values reflect a more structural misfit of items. High outfits are less of a threat to measurement.

Because fit statistics are affected by response dependency (multidimensionality), rating scale categories (e.g., disordered thresholds), the number of items and sample size, using rules of thumb can be misleading. Therefore, alignment of observed and predicted scores was verified by visually inspecting the item characteristic curves (ICC; i.e., the probability of responding to an item as a function of ability level) (Bond & Fox 2001). Global model fit statistics, that is, Pearson global Chi-square indices, were also evaluated (Tesio et al. 2023b).

Measurement accuracy and score-to-measure conversion. Fourth, measurement accuracy was evaluated. Items used should collectively form a hierarchy of difficulty levels. Item separation indices indicate the number of distinct levels of difficulty that can be distinguished. Analogously, person separation indices indicate the number of ability levels that can be distinguished. Reliability indices quantify the reproducibility of item and person ordering. Reliability estimates are indices of how much differences between measures reflect real differences rather than measurement errors; they depend on the variance of item difficulty and sample ability levels, and the size of the item pool and sample (Tesio et al. 2023a, b).

A well-targeted instrument must have a distribution of item difficulty that closely matches the distribution of person ability, to ensure the items measure the full range of people. Targeting is the difference between the mean person ability and the mean item difficulty or the relative item and person distributions when placed upon the same scale. In Rasch analysis, item difficulty is centered around 0 logits, with 0 representing the item of average difficulty; person ability is expressed relative to this value. The closer the mean person ability is to the mean item difficulty, the better the targeting, with a difference of 0 logits between both indicating perfect targeting. Extreme scores are typically discarded when evaluating targeting (Tesio et al. 2023a,b).

Score-to-measure conversion graphs (test characteristic curves) were constructed to convert raw ordinal scores into linear equal interval ability measures.

Measurement invariance across samples. Last, measurement invariance was evaluated across samples. Item functioning is intended to be invariant with respect to sample characteristics. In order for composite measures to be unidimensional, this measure to be linear, and to make meaningful comparisons between people, the items of a scale have to function invariantly across sample groups. Lack of invariance among sample groups or items functioning differently between groups who otherwise share the same ability level, is called differential item functioning (DIF, Humphry & Montuoro 2021). Across samples, scores may be similar, but the item difficulty may not. DIF metrics were inspected to check whether the items measured the latent trait similarly across samples or, stated differently, whether item difficulty changed as a function of the sample. Both the size and significance of the DIF contrast are important. Accurate DIF analysis relies on sample sizes that are similar with N ≥ 100. Therefore, to compare DIF for ECLiPS-UK, -US, and -FL, data (n = 120) from ECLiPS-UK (N = 650) were pseudo-randomly selected to ensure a balanced distribution across age and gender.

RESULTS

Demographic, Auditory, and Cognitive Characteristics of the Flemish Sample

The median age of the Flemish sample was 8.5 years (95% confidence interval of median [8 years; 9 years]) and was similar for boys and girls (U = 1502, p = 0.70). With respect to hearing sensitivity of the whole sample (N = 112), 72% had hearing thresholds ≤15 dB HL in both ears for all frequencies measured, and ~92% had thresholds ≤20 dB HL. Those with a history of ventilation tubes (N = 20 children, 18%) had slightly poorer (2.5 dB HL) mean hearing sensitivity [F(1,110) = 5.01, p = 0.03]. In a subset of our sample (N = 51), a mean SRT of −15.1 dB (SD = 1.4 dB) was found for the anti-phasic DTT. Age (p < 0.05), but not gender (p = 0.79) nor worse-ear mean hearing sensitivity (p = 0.86) were significantly associated with the SRT, with children aged 6 years (−12.7 dB SNR) performing significantly worse compared with older children (−15.3 dB SNR). With respect to cognitive performance, our sample generally performed better compared with population normative values, especially concerning nonverbal intelligence, auditory and working memory, and attention switching (Table 2).

Inter-Rater Reliability and Typical Scores for ECLiPS-FL

ECLiPS-FL: Inter-Rater Reliability • An RM-ANOVA with ECLiPS factor and rater as within-subject factors did not show a significant main effect of rater [F(1,111) = 0.00, p = 0.99] and no Significant Rater × Factor Interaction [F(4,444) = 1.05, p = 0.38]. With respect to aggregate scores, no effect of rater [F(1,111) = 0.03, p = 0.85] was observed, but the analysis did reveal a Significant Rater × Aggregate Interaction [F(1.4,155.79) = 4.67, p = 0.02]. TOT scores did not differ significantly among raters (t[111] = 0.20, p = 0.85). Intraclass correlation values of 0.52 (SAP), 0.73 (MA), 0.59 (LLL), 0.59 (PSS), and 0.70 (EAS) were found for factor scores. Aggregate score intraclass correlation values were all 0.68. A value of 0.67 was found for the TOT score, indicating moderate agreement (Koo & Li 2016).

ECLiPS-FL: Typical Scores as a Function of Age and Gender • A MANOVA analysis with ECLiPS-FL factor scores as dependent variables showed a significant effect of gender [*F*(5,96) = 3.04, *p* = 0.01] but not age [*F*(25,500) = 1.00, *p* = 0.47]. Post hoc univariate ANOVA tests showed differences in the scores between boys and girls for MA [*F*(1,100) = 5.65, *p* = 0.02] and PSS [*F*(1,100) = 5.99, *p* = 0.02], with boys obtaining higher (poorer) scores. Similarly, with respect to aggregate scores, the gender factor [*F*(3,98) = 2.72, *p* = 0.05], but not age [*F*(15,300) = 0.56, *p* = 0.89] was significant, with girls obtaining better scores than boys for LIS [*F*(1,100) = 5.10, *p* = 0.03]. TOT ECLiPS score were not affected by gender [*F*(1,100) = 3.70, *p* = 0.06] or age [*F*(5,100) = 0.38, *p* = 0.86]. Age × Gender Interactions were never significant (see top panels in Figs. 1 and 2 for gender and age effects, respectively).

Typical Scores for Comparator Samples, and Multi-Samples Comparison

ECLiPS-US and ECLiPS-UK: Typical Scores as a Function of Age and Gender • Comparable effects were found in the US-sample, with a significant effect of gender [*F*(5,55) = 2.61,

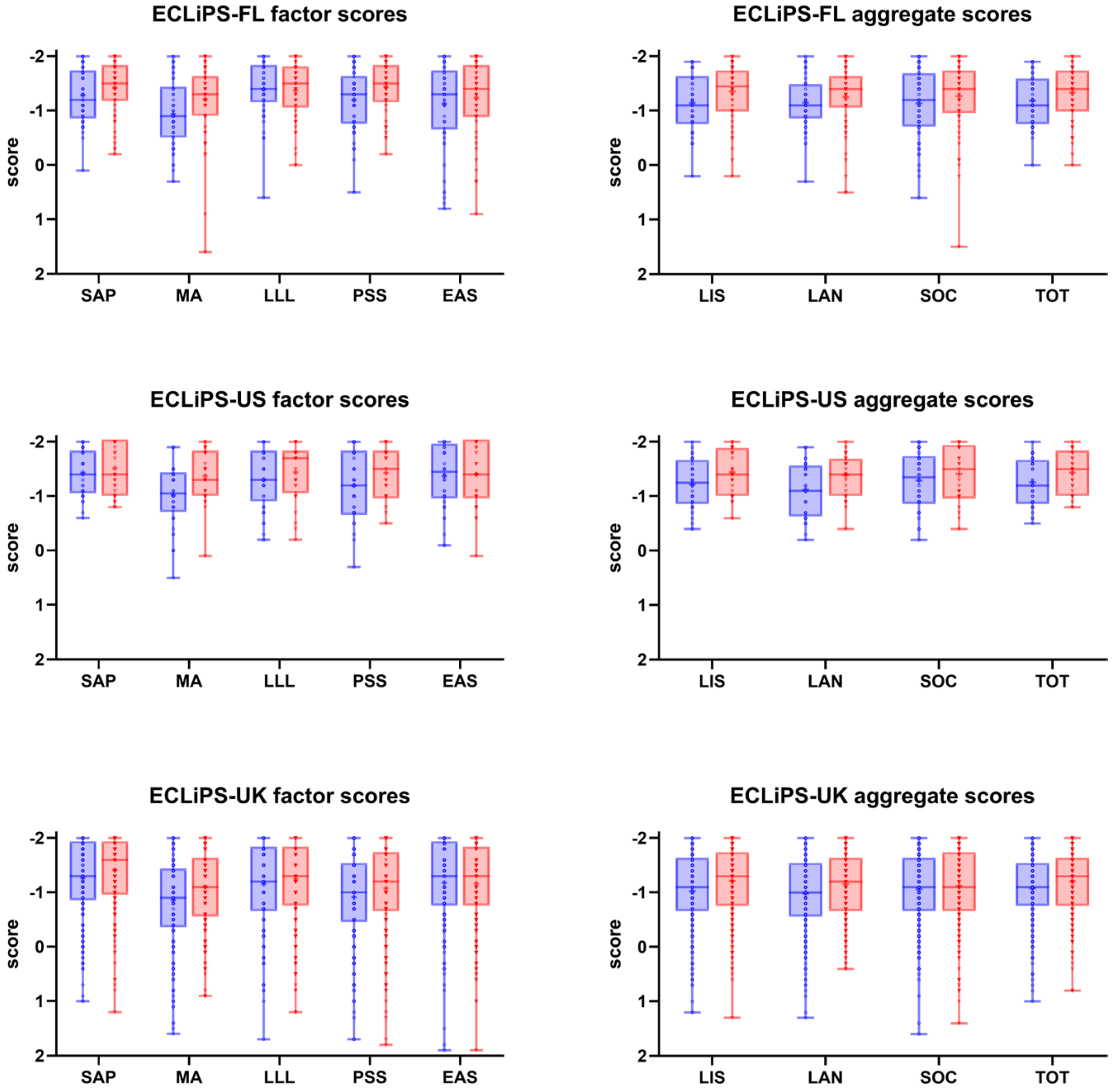

Fig. 1. ECLiPS factor (left-sided panels) and aggregate raw scores (right-sided panels) for the different ECLiPS versions as a function of gender. Blue data points and boxplots represent data obtained from boys. Red data points and boxplots represent data obtained from girls. More negative scores indicate higher ability levels. ECLiPS indicates Evaluation of Children's Listening and Processing Skills.

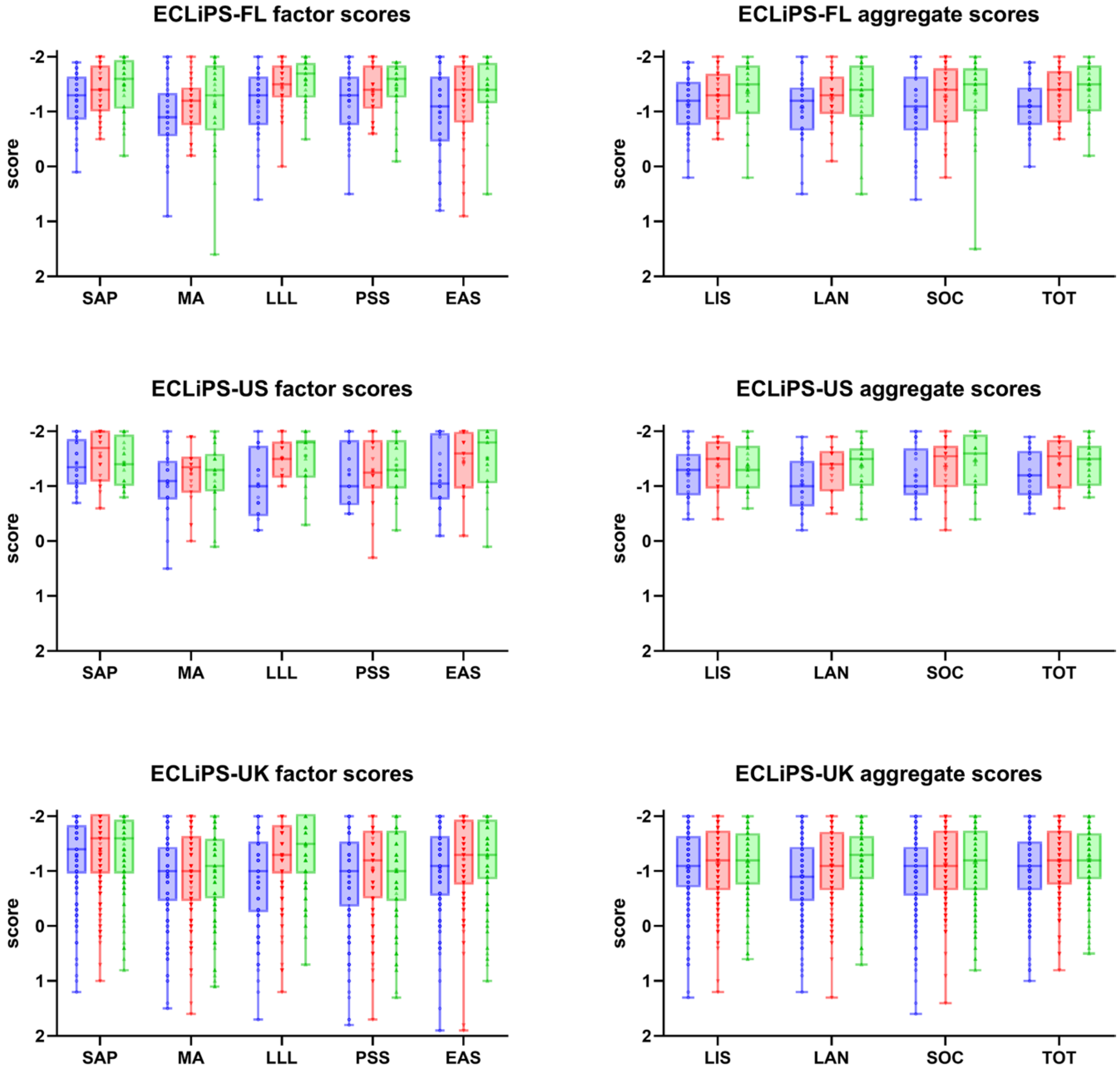

Fig. 2. ECLiPS factor (left-sided panels) and aggregate (right-sided panels) raw scores for the different ECLiPS versions as a function of age. For visualization purposes, the data were grouped into three age categories: data from children aged 6–7 yrs are represented in blue, data from children aged 8–9 yrs are represented in red, and data from children aged 10–11 yrs are represented in green. More negative scores indicate higher ability levels. ECLiPS indicates Evaluation of Children's Listening and Processing Skills.

$p = 0.03$] for MA [$F(1,59) = 8.73$, $p < 0.01$] and absence of age effects [$F(25,295) = 1.45$, $p = 0.08$]. No effect of age and gender was observed for aggregate [age: $F(15,177) = 1.67$, $p = 0.06$; gender: $F(3,57) = 2.61$, $p = 0.06$] or TOT [age: $F(5,59) = 0.01$, $p = 0.55$; gender: $F(1,59) = 3.49$, $p = 0.07$] scores (see middle panels in Figs. 1 and 2 for gender and age effects, respectively). In the larger UK-sample, a significant effect of age was found [$F(25,3190) = 5.57$, $p < 0.01$] for LLL [$F(5,638) = 19.53$, $p < 0.01$] and EAS [$F(5,638) = 3.04$, $p = 0.01$] factor scores, as well as for LAN [$F(5,638) = 5.64$, $p < 0.01$], SOC [$F(5,638) = 3.05$, $p = 0.01$], and TOT [$F(5,638) = 3.91$, $p < 0.01$] scores. However, except for LLL ($r = -0.34$), Spearman correlational analyses indicate very small correlation coefficients ($r < -0.18$). A significant gender effect [$F(5,634) = 5.79$, $p < 0.01$] was found for SAP [$F(1,638) = 8.36$, $p < 0.01$], MA [$F(1,638) = 17.27$, $p < 0.01$], and PSS [$F(1,638) = 7.46$, $p < 0.01$] factor scores, and for LIS [$F(1,638) = 14.11$, $p < 0.01$], LAN [$F(1,638) = 11.58$, $p < 0.01$], and TOT [$F(1,638) = 5.88$, $p = 0.02$] scores. Age × Gender Interactions never reached significance (see bottom panels in Figs. 1 and 2 for gender and age effects, respectively).

Across-Samples Comparison of Typical Scores as a Function of Age and Gender • Significant differences across samples were found for ECLiPS factor scores [$F(10,1588) = 4.50$, $p < 0.01$]. These never interacted with the age [$F(50,3985) = 0.73$, $p = 0.93$] or gender [$F(10,1588) = 0.61$, $p = 0.81$] of the children. Differences were observed for LLL [$F(2,797) = 3.33$, $p = 0.04$], PSS [$F(2,797) = 10.37$, $p < 0.01$], and EAS [$F(2,797) = 3.19$,

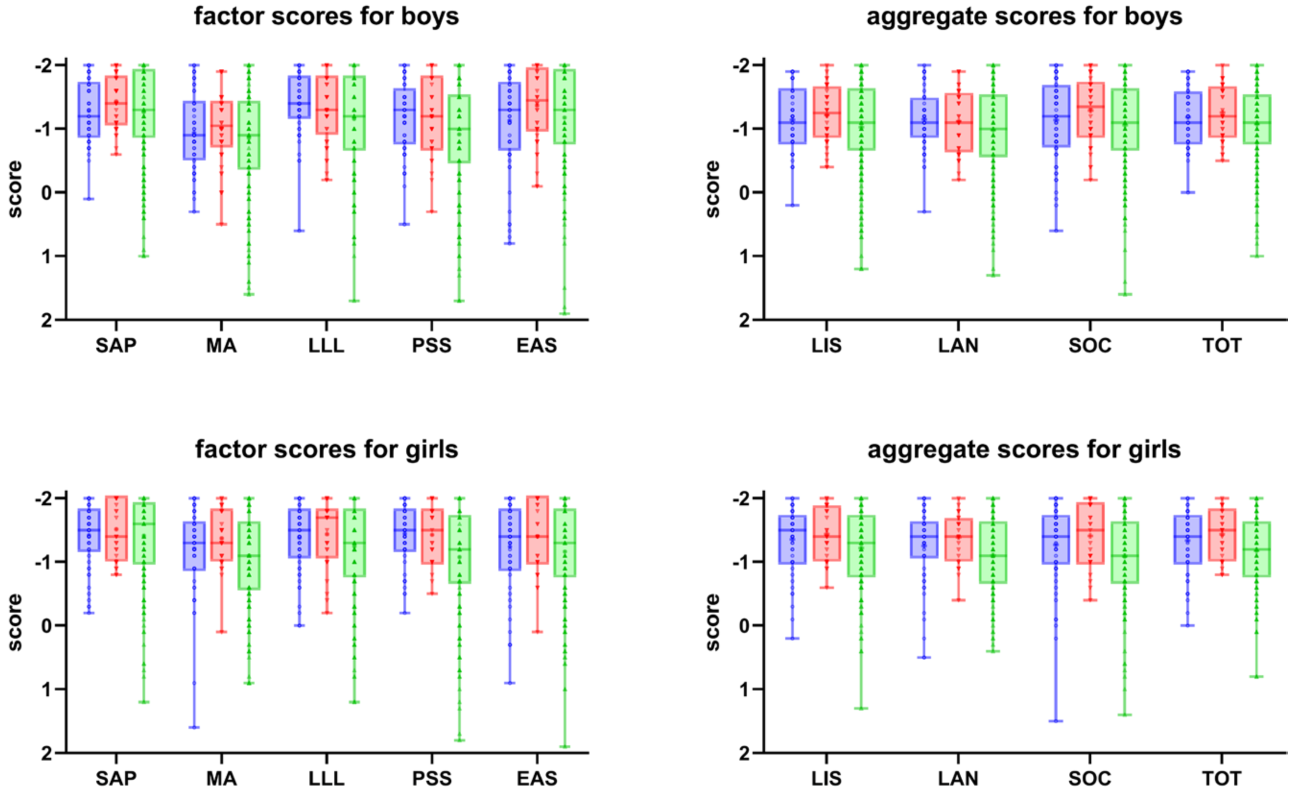

Fig. 3. ECLiPS factor (left-sided panels) and aggregate (right-sided panels) raw scores as a function of sample. Given the absence of clear age effects, the data were pooled across age and plotted as a function of gender. Data from Flemish children are shown in blue, data from US children are shown in red, and data from UK children are green. More negative scores indicate higher ability levels. ECLiPS indicates Evaluation of Children's Listening and Processing Skills.

p = 0.04]. Similarly, borderline significant differences were found across samples for aggregate scores [F(6,1592) = 2.12, p = 0.05]. Both LIS [F(2,797) = 3.57, p = 0.03], SOC [F(2,797) = 5.78, p < 0.01], and TOT [F(2,797) = 3.75, p = 0.02] scores showed differences.

Post hoc univariate ANOVAs and t tests (with Bonferroni correction) showed differences between UK and both FL (t[830] = 3.77, p < 0.01) and US (t[830] = 2.93, p = 0.01) samples for PSS, and between UK and US samples for EAS (t[830] = 2.51, p = 0.04), SOC (t[830] = 3.10, p < 0.01), and TOT (t[830] = 2.47, p = 0.04) scores. No significant differences between samples for LLL [F(2,830) = 2.92, p = 0.06] and LIS [F(2,830) = 3.50, p = 0.03 − post hoc t tests did not reveal differences, with p > 0.10] were observed after post hoc univariate analyses. Figure 3 shows sample effects as a function of gender.

Interim Summary of Results

ECLiPS scores seem to be relatively invariant with age. With respect to gender, separate normative values for boys and girls are needed, especially for the MA and PSS factors. Across-sample differences are most pronounced for PSS and EAS (the factors constituting the SOC aggregate).

Psychometric Evaluation Based on Rasch Analysis, and Multi-Samples Comparison

Verification of Rasch Assumptions • Except for EAS in the ECLiPS-US, core Rasch assumptions of unidimensionality and local independence were reasonably well met (Supplementary Table S1 in Supplemental Digital Content, http://links.lww.com/EANDH/B391). Item 6 ("complains about loud sounds") and item 10 ("complains about sounds being unpleasant") were found to be linked beyond what is described by the underlying trait. Because this response dependency corrupts unidimensionality, the Rasch analysis for this factor was rerun after removal of item 6 (the most misfitting one of the two). An Eigenvalue of 1.89 was subsequently found, and there was no longer any significant local item dependence.

Rating Scale Category Structure • At least 10 responses per rating scale are recommended (Mallinson et al. 2022). This was only achieved for ECLiPS-UK (Supplementary Table S2 in Supplemental Digital Content, http://links.lww.com/EANDH/B392). However, a consistent finding across factors and ECLiPS versions was the presence of disordered thresholds with ordered categories (Table 3). CPC curves clearly show submergence of category 3 (the neutral one) for most ECLiPS factors (Supplementary Figure S2 in Supplemental Digital Content, http://links.lww.com/EANDH/B385).

Quantitative and Qualitative Assessment of Data-Model Fit • ICC curves are shown in Figure S3 in Supplemental Digital Content, http://links.lww.com/EANDH/B386. Global data-model fit statistics are summarized in Supplementary Table S4 in Supplemental Digital Content, http://links.lww.com/EANDH/B394. To evaluate item fit, sample sizes of >200 are recommended. This was only achieved for ECLiPS-UK. On the other hand, for item calibrations and person measures stable within ±1 logit, a sample of N = 50 is deemed appropriate for polytomous scales (Mallinson et al. 2022), which was achieved in the present study.

TABLE 3. Rating scale category difficulty (in logits) and thresholds (between brackets, in logits) for the different ECLiPS factors as a function of sample

		FL		US		UK	
Factor	Category	Difficulty Measure	Threshold	Difficulty Measure	Threshold	Difficulty Measure	Threshold
SAP	1	−2.98	—	−2.42	—	−3.73	—
	2 (1–2)	−1.21	−1.75	−1.11	−1.08	−1.51	−2.57
	3 (2–3)	−0.12	**−0.06**	−0.33	**0.12**	−0.07	**−0.03**
	4 (3–4)	1.15	−0.33	0.86	**−1.31**	1.49	−0.15
	5 (4–5)	3.29	2.13	3.39	2.27	3.89	2.74
MA	1	−3.88	—	−2.94	—	−3.46	—
	2 (1–2)	−1.42	−2.74	−1.08	−1.75	−1.28	−2.30
	3 (2–3)	0.12	**0.30**	−0.05	**0.41**	0.04	**0.26**
	4 (3–4)	1.47	0.00	1.05	−0.57	1.30	−0.16
	5 (4–5)	3.61	2.44	3.07	1.91	3.36	2.19
LLL	1	−4.04	—	−2.72	—	−3.04	—
	2 (1–2)	−1.67	−2.89	−1.16	−1.40	−1.11	−1.85
	3 (2–3)	0.01	−0.18	−0.13	−0.37	0.01	**0.28**
	4 (3–4)	1.67	0.21	1.10	−0.08	1.12	−0.25
	5 (4–5)	4.00	2.85	3.05	1.85	3.02	1.82
PSS	1	−3.21	—	−2.69	—	−3.00	—
	2 (1–2)	−1.35	−1.98	−1.16	−1.39	−1.09	−1.81
	3 (2–3)	−0.11	−0.29	−0.21	**−0.18**	0.01	**0.30**
	4 (3–4)	1.31	−0.03	1.04	−0.54	1.10	−0.27
	5 (4–5)	3.47	2.30	3.25	2.10	2.98	1.78
EAS	1	−3.41	—	−3.82	—	−3.34	—
	2 (1–2)	−1.35	−2.24	−1.40	−2.70	−1.35	−2.16
	3 (2–3)	−0.10	**0.11**	−0.05	**0.64**	−0.09	**−0.03**
	4 (3–4)	1.31	−0.38	1.37	−0.77	1.31	−0.21
	5 (4–5)	3.65	2.51	3.95	2.83	3.55	2.40

Values in bold indicate disordered, that is, not monotonically rising, thresholds.
Category 1: totally agree; category 2: agree; category 3: neutral; category 4: disagree; category 5: totally disagree.
EAS, environmental and auditory sensitivity; LLL, language, literacy, and laterality; MA, memory and attention; PSS, pragmatic and social skills; SAP, speech and auditory processing.

TABLE 4. Rasch item, person, and targeting metrics as a function of sample, and measurement invariance (DIF) across samples

			Item Metrics			Person Metrics			
Factor	Sample	Items With DIF Across Samples	Misfitting Items	Separation Index	Reliability Index	Separation Index	Reliability Index	% of Extremes	Targeting (Logits)
SAP	FL	1, 11, 16, 38	5	**2.07**	0.81	**1.66**	**0.73**	17%	**1.81**
	US		1	3.34	0.92	**1.18**	**0.58**	21%	**2.19**
	UK		38	6.14	0.97	**1.91**	**0.79**	20%	**2.32**
MA	FL	8, 25, 32	—	4.14	0.94	**1.61**	**0.72**	9%	**1.73**
	US		4	**1.31**	**0.63**	**1.31**	**0.63**	3%	**1.59**
	UK		—	4.12	0.94	**1.77**	**0.76**	7%	**1.44**
LLL	FL	20	—	3.84	0.94	**1.42**	**0.67**	20%	**2.45**
	US		—	**1.86**	0.78	**0.98**	**0.49**	13%	**1.96**
	UK		—	5.55	0.97	**1.19**	**0.59**	20%	**1.44**
PSS	FL	7, 14	—	3.05	0.90	**1.25**	**0.61**	20%	**1.82**
	US		—	**1.76**	0.76	**1.24**	**0.61**	15%	**1.75**
	UK		—	7.53	0.98	**1.39**	**0.66**	12%	**1.21**
EAS	FL	29	—	**1.85**	0.77	2.18	0.83	17%	**1.66**
	US		3, 15	**0.00**	**0.00**	**1.29**	**0.62**	27%	**2.12**
	UK		—	7.15	0.98	**1.78**	**0.76**	15%	**1.74**

Values in bold fall outside the range of pre-set critical values for separation indices (≥3 for item separation; ≥2 for person separation), reliability coefficients (≥0.70 and ≥0.80 for item and person reliability, respectively), and proper targeting (>0.5 logits).
Item 1: "takes time to realize that someone has said something to him/her"; item 3: "finds it tiring being in groups of people"; item 4: "needs strategies for remembering"; item 5: "when there is a sudden noise, is confused about where to look"; item 7: "has obsessive interests"; item 8: "finds it difficult to do more than one thing at a time"; item 11: "seems to struggle to hear at times"; item 14: "gets frustrated because others misunderstand what he/she means to say"; item 15: "becomes upset if daily routine is changed"; item 16: "seems deaf when lots of people are talking"; item 20: "needs help from others when writing homework down"; item 25: "remembers series of instructions"; item 29: "becomes upset in crowded spaces"; item 32: "forgets about things that happen on a schedule"; item 38: "has problems understanding what people say."
DIF, differential item functioning; EAS, environmental and auditory sensitivity; LLL, language, literacy and laterality; MA, memory and attention; PSS, pragmatic and social skills; SAP, speech and auditory processing.

Overall, misfitting (underfit) was noted for six items (items 1, 3, 4, 5, 15, and 38), with most items belonging to SAP (Table 4, Supplementary Table S3 in Supplemental Digital Content, http://links.lww.com/EANDH/B393, and Supplementary Figure S4 in Supplemental Digital Content, http://links.lww.com/EANDH/B387). Misfitting items were not reproduced across samples.

Measurement Accuracy and Score-to-Measure Conversion • Item difficulty estimates (in logits) are shown in Supplementary Table S3 in Supplemental Digital Content, http://links.lww.com/EANDH/B393. SEs for item difficulty estimates were generally very small, indicating high overall precision of the ECLiPS items.

Separation and reliability metrics can be found in Table 4. In general, item separation and reliability indices were higher for the largest UK-sample, with between five (MA) and eight (EAS, PSS) distinguishable levels of difficulty and excellent reliability (>0.9).

ECLiPS-FL items' difficulty hierarchy ranged from two (EAS) to five (MA). ECLiPS-US item separation indices were somewhat lower, with two (MA, LLL, PSS) to four (SAP) distinguishable difficulty levels. For EAS-items, there was no hierarchy in item difficulty. Reliability estimates for ECLiPS-FL items ranged from acceptable (EAS) to excellent (MA, LLL, PSS). Regarding the ECLiPS-US, item reliability was unacceptable (<0.7) for MA and EAS, acceptable for LLL and PSS, and excellent for SAP (Linacre n.d.). The ECLiPS-UK demonstrated excellent item reliability. In general, person separation was low.

Targeting metrics deviate largely from 0 in the positive direction, indicating that the samples consisted of high performers or, stated differently, that items were too easy relative to person ability. Low person separation and poor targeting can be expected in a sample of typically developing children with homogeneous (and high) ability levels among people (Table 4).

Steep test characteristic curves, or test score to ability conversion graphs, also demonstrate high person ability levels. Nearly the complete range of factor scores corresponds to ability levels over a range of four logits around the average ability level of 0 logits (Supplementary Figure in Supplemental Digital Content, S4 http://links.lww.com/EANDH/B387).

Measurement Invariance Across Samples • Eleven items (items 1, 7, 8, 11, 14, 16, 20, 25, 29, 32, and 38) were functioning differentially as a function of the sample (Table 3). With 4/9, 3/8, and 2/6, DIF was most prominent for items belonging to the factors SAP, MA, and PSS, respectively (the factors that form the LIS aggregate). EAS (1/8) and LLL (1/6) showed considerably higher invariance across samples. DIF metrics can be found in Supplementary Table S5 in Supplemental Digital Content, http://links.lww.com/EANDH/B395.

DISCUSSION

This study's objectives were: (1) to collect normative data for the Flemish ECLiPS, a caregiver questionnaire to identify childhood LiD, and to compare these data with typical scores obtained by children from two comparator samples, and (2) to perform a psychometric evaluation of the questionnaire against the Rasch measurement model. To our knowledge, this is the first study doing so in the field of questionnaires for LiD or auditory processing difficulties.

Associations of ECLiPS Scores With Age and Gender

ECLiPS scores were relatively invariant with age (Fig. 2). With respect to gender, differences between boys and girls were noted (Fig. 1), with girls typically obtaining more favorable scores than boys (most markedly for MA and PSS when considering all samples). Age and gender effects were most pronounced in the UK population sample.

On the one hand, the absence of clear age effects on the ECLiPS facilitates interpretation by the clinician, who does not need to convert scores to age-appropriate normative values. On the other hand, as childhood is a period of marked neurocognitive and sensory development, the ECLiPS appears to be rather insensitive to these developmental trajectories. This might be due to differences between ability assessment by means of formal behavioral tests and caregiver observations of ability probed by questionnaire items.

More complex language functions as well as executive functions are known to develop well into adolescence (Korkman et al. 2001). In their Generation R study, Mous et al. (2017) investigated associations of gender and age with neurocognitive functioning in the domains of memory (working memory and recall), (selective and sustained) attention, and language (verbal fluency) with tasks from the NEPSY-II-NL in a large group of young (6 to 10 years old) typically developing children from the Netherlands. In agreement with our data, they found girls outperforming boys for behavioral tasks measuring MA abilities. They also found a clear association of neurocognitive task performance with age.

To be successful communicators, children not only need to acquire morpho-syntactic and semantic language skills, but also require a specific set of so-called pragmatic abilities for the proper use and interpretation of language in different situations, for example, to understand communicative intentions. Rothermich et al. (2020) demonstrated that children's ability to comprehend speaker intentions (using acoustic and nonverbal cues) increases with age between 8 and 12 years. Furthermore, they found girls outperforming boys in classifying lies and sarcasm as insincere and point to advanced social perspective-taking abilities, backing up our finding of poorer PSS scores in boys compared with girls to some extent.

Across-Samples Comparability of ECLiPS Scores

Across samples, the most pronounced differences were found for the factors that make up the social ECLiPS scale (PSS and EAS), with UK children systematically obtaining poorer scores. No significant differences were found between ECLiPS scores obtained from US and FL children. Because of unequal sample sizes between the US and FL samples on the one hand, and the UK sample on the other hand, we verified these findings using nonparametric Kruskal–Wallis tests and post hoc Dwass–Steel–Critchlow–Fligner pairwise comparisons (separate analyses for boys and girls, with p values adjusted for multiple comparisons). Significant differences across samples with respect to the PSS factor remained, especially among FL and UK children.*

Overall, except for SAP, ECLiPS-UK population scores tended to be somewhat poorer, compared with scores obtained from the other samples (especially for girls, see Fig. 3). This is not surprising, given the population-based sampling technique

*SAP: $\chi^2(2) = 0.93$, $p = 0.63$ (boys), $\chi^2(2) = 0.73$, $p = 0.69$ (girls); MA: $\chi^2(2) = 1.03$, $p = 0.60$ (boys), $\chi^2(2) = 6.26$, **$p = 0.04$** (girls); LLL: $\chi^2(2) = 4.13$, $p = 0.13$ (boys), $\chi^2(2) = 1.22$, $p = 0.54$ (girls); PSS: $\chi^2(2) = 6.71$, **$p = 0.04$** (boys), $\chi^2(2) = 11.74$, **$p < 0.01$** (girls); EAS: $\chi^2(2) = 2.66$, $p = 0.26$ (boys), $\chi^2(2) = 4.69$, $p = 0.10$ (girls). PSS scores for girls are significantly poorer in the UK sample compared to the FL sample ($p < 0.01$). Pairwise comparisons did not reveal other sample-related differences for ECLiPS factor scores.

to collect typical data and, as such, minimize bias toward including respondents from more educated backgrounds only. Whereas equal definitions for "normal hearing" and "typical development" were used, both the Flemish and US samples were convenience samples, with children recruited from unchallenging socio-economic areas. Demographic factors, such as socio-economic status and parental education have an influence on children's cognitive skills (Bradley et al. 2001). Furthermore, both children from the US (Petley et al. 2021) as well as the FL sample were found to have neurocognitive scores that were generally higher than those of the normative samples, suggesting stronger neurocognitive ability (Table 2). Social communication is also influenced by culture (Hwa-Froelich & Vigil 2004), which may also contribute to observed differences among samples, specifically for the PSS factor. In sum, the comparability of normative ECLiPS scores across versions seems to be influenced by the sampling technique. ECLiPS scores from strictly selected children with verified normal hearing and age-appropriate neurocognitive ability, whose caregivers are typically highly educated, seem to agree quite well. Population data for ECLiPS-FL and ECLiPS-US are not available at this time, but are expected to be more in line with ECLiPS-UK population data.

The Structural Psychometric Validity of the ECLiPS

Rating Scale Categories • ECLiPS items are mostly formulated in such a way that (totally) not agreeing with them indicates higher ability levels, and (completely) agreeing with them indicates lower levels of ability. A neutral response category lies in the middle of the response scale. The validity of the ECLiPS rating scale was evaluated across samples. A consistent finding was the appearance of disordered category thresholds, while, on average, categories work as intended: higher scores require higher ability levels to be observed. Disordered boundaries or submerged categories are often the results of respondents seldomly choosing a certain response category (Tesio et al. 2023a, b). This was certainly observed in our data, and could be expected from respondents answering items about their typically developing children. However, the neutral response category was mostly submerged, and not a category at the lower ability end of the scale, which might also be a consequence of poorer conceptualization. A remedy could be to collapse this middle category with an adjacent one, and rescore responses. We did not evaluate whether this would result in ordered category boundaries. There is controversy about the need for ordered thresholds to conclude whether a rating scale works as intended (Adams et al. 2012; Andrich 2013).

Data-Model Fit • Underfitting items degrade the measuring instrument and may lead to an under-detection of difficulties. Overfitting items, on the other hand, will tend to overestimate differences in raw scores, interfering in comparisons within and between people (Bond & Fox 2001). With acceptable item fit statistics, we can be confident that scores accurately and predictably fit the Rasch measurement model (Stemler & Naples 2021). They can be considered a first pass, after which ICC curves can be reviewed in detail.

Across samples, fit statistics indicated a clear misfit for four items (~10%). Overfitting items were never reproduced across samples (different items were found to be misfitting in different samples). Two items of the ECLiPS-US demonstrated high outfit statistics only, which is less problematic.

There is considerable debate about the most appropriate fit statistics to use, the range of fit statistics to be used when evaluating fit, and the interpretation of fit statistics. Some researchers have proposed adjusting the critical range used for fitting items depending on the sample size. However, Smith et al. (2008) showed that in- and outfit MNSQs are relatively insensitive to sample size for polytomous data. Qualitative evaluation of data-model fit, by inspecting ICCs, showed reasonable fit for some items considered "misfitting" based on statistical criteria. Other items (e.g., items 14 and 22) were more "misfitting" graphically while demonstrating fit statistics within the acceptable range.

Measurement Accuracy • Reasonable item separation and reliability were observed for most ECLiPS factors, with separation indices being generally lower for the ECLiPS-US.

Item separation is used to verify the item hierarchy, and its associated reliability quantifies the reproducibility of item hierarchy. Item reliability is higher when the items have a wider difficulty range, and/or when the sample size is larger. This was confirmed in our data.

Person separation quantifies the sensitivity of an instrument to distinguish between low and high performers. Person reliability is higher when the sample demonstrates a wide range in ability level and when there is better sample-item targeting. It is independent of sample size.

Generally, our analyses demonstrate poor person metrics (and poor targeting). This was expected, given that the samples considered comprised typically developing normal-hearing children without LiD.

Measurement Invariance • DIF inspects the items in a questionnaire for signs of interactions with sample characteristics. Groups are compared, stratified based on matching ability levels, and their relative performance on each item is quantified. The ability levels are usually based on total scores. In this way, item-specific DIF analysis is relatively independent of the DIF analyses of the other items. On the other hand, the overall impact of item DIF, accumulated across items, is unclear. While resolving an item showing evidence of DIF may improve fit statistics, one should be careful not to degrade the content validity of the scale (Hagquist & Andrich 2017). Interestingly, two items with misfit also had DIF (items 1 and 38).

DIF analysis can be useful to check translations among different versions of a questionnaire (Petersen et al. 2003). Whereas adapting a questionnaire for use in other languages, merely seems to be a matter of translating the items it is composed of, it may well be that items have a different meaning in a specific cultural context. As such, translations may induce bias, affect item difficulty and equivalence.

The samples compared in this study provide an interesting opportunity to investigate linguistic and cultural influences on the ECLiPS items, based on DIF. If linguistic effects were the main driver for differential functioning, we would expect differences between the FL sample and both the US and UK samples, in the absence of differences between US and UK samples (where the same ECLiPS wordings were used). This was only the case for item 32 ("forgets about things that happen on a schedule"). Four items with DIF for sample showed differences between the US and UK samples, mostly in combination with differences between the FL sample and either the US (three items) or UK (one item) sample. The remaining six items showed differential functioning between the FL and US (two items) or FL and UK (four items). This leads to the conclusion that primarily cultural effects must be at play, but linguistic effects cannot be ruled out. It would be interesting to investigate whether removing items with DIF and recalculating factor scores improves the across-samples comparison. As mentioned earlier, the most distinct differences between samples were found for the PSS factor. It was definitely not the case that items belonging to this factor show more DIF or poorer fit. Rather, DIF was most pronounced for items belonging to the SAP factor.

Contributions of ECLiPS to Clinical Care

LiD in children can be an expression of (interacting) deficits in auditory processing, cognitive processing and/or language processing. The ECLiPS questionnaire is a caregiver-report measure that was developed with the aim of capturing the underlying sources that contribute to the experienced processing difficulties related to listening in daily life. This could support clinical practice, for instance in deciding which specialists to involve in the assessment and management of children who present with difficulties. Such a questionnaire would be useful in audiology clinics in Flanders, Belgium, where clinicians often have difficulties knowing how best to manage children referred for auditory processing difficulties and with normal hearing according to the pure-tone audiogram. This is particularly important because many children are referred for auditory assessment, following routine school-age hearing screening. This program refers children based on a failed self-test based on digit triplet identification in noise (Denys et al. 2018; Guérin et al. 2018) who often turn out to have normal hearing sensitivity. The ECLiPS may help to identify LiD that have a different management pathway compared with a traditional hearing loss.

CONCLUSION

Typical values for the ECLiPS, a caregiver-report outcome measure to profile auditory and cognitive real-world abilities important for successful listening and (auditory) processing, demonstrate gender-based differences for some domains. Age-based differences were less pronounced. Across samples, factor and aggregate scores were quite comparable, with a notable exception for the social domain scores, which were considerably poorer for the less strictly selected UK population sample. For future translated versions of the scale, it is advisable to collect typical values and evaluate gender-based differences. Regarding its psychometric quality, moderate interrater agreement was demonstrated for the ECLiPS-FL. ECLiPS items were found to fit the Rasch measurement model within reasonable margins of error, and could generally distinguish between two or more difficulty levels. Some items were found to function differently across samples, which seemed to be a cultural rather than a linguistic effect. Rasch analyses confirmed the children considered in this comparative psychometric validation study to be quite homogeneous with respect to (listening) ability levels. Future studies could investigate whether removing the biased items further improves comparability across samples without affecting the questionnaire's content validity and other psychometric properties. Preferably, these studies should include children from more diverse populations with respect to socio-economic and linguistic backgrounds, and also consider children with self-reported or observed LiD. In its current form, the ECLiPS caregiver questionnaire appears to be a psychometrically valid qualitative measure to capture LiD in elementary school children.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the following master students of the Speech-Language Pathology and Audiology educational program (KU Leuven, Faculty of Medicine) for their help with data collection in the framework of their master thesis: Céline Mondelaers, Karlien Smolders, Laura Aerts, Sofie Princen, Liesbeth Vander Velde, Tine Weerens, Catho Nulmans, and Esra Cetinkaya.

S.D. and A.v.W. were funded by the Horizon 2020 research and innovation program under the Marie Sklowdoska-Curie Grant (agreement number 860755), and by a TBM grant from the Research Foundation Flanders (FWO, grant number: T000823N). J.B. was supported by the UK Nottingham University Hospitals NHS trust and the UK Medical Research Council. D.R.M. received funding from the US National Institute of Deafness and other Communicative Disorders Grant R01DC014078, the UK Medical Research Council, and the NIHR Manchester Biomedical Research Center. N.V. received a Research Foundation Flanders (FWO) senior clinical investigator fund (1804816N).

S.D. and A.v.W. designed the study. D.R.M. and J.B. provided the data from comparator samples. S.D. analyzed the data and wrote the paper. All authors discussed the results and implications and commented on the manuscript at all stages.

The authors have no conflicts of interest to disclose.

Address for correspondence: Sam Denys, Department of Neurosciences, University of Leuven, Research Group Experimental Otorhinolaryngology, Leuven, Belgium. E-mail: sam.denys@kuleuven.be

Received April 24, 2023; accepted March 15, 2024; published online ahead of print June 3, 2024

REFERENCES

Adams, R. J., Wu, M. L., Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educ Psychol Meas*, *72*, 547–573.

Ahmmed, A. U. (2020). Auditory processing, co-morbidities, and parental report of sleep disturbance in children with auditory processing disorder (APD). *Int J Pediatr Otorhinolaryngol*, *135*, 110117.

Ahmmed, A. U., & Ahmmed, A. A. (2016). Setting appropriate pass or fail cut-off criteria for tests to reflect real life listening difficulties in children with suspected auditory processing disorder. *Int J Pediatr Otorhinolaryngol*, *84*, 166–173.

Ahmmed, A. U., Ahmmed, A. A., Bath, J. R., Ferguson, M. A., Plack, C. J., Moore, D. R. (2014). Assessment of children with suspected auditory processing disorder: A factor analysis study. *Ear Hear*, *35*, 295–305.

Alvand, A., Kuruvilla-Mathew, A., Roberts, R. P., Pedersen, M., Kirk, I. J., Purdy, S. C. (2023). Altered structural connectome of children with auditory processing disorder: A diffusion MRI study. *Cereb Cortex*, *33*, 7727–7740.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "Threshold Disorder Controversy." *Educ Psychol Meas*, *73*, 78–124.

Barry, J. G., & Moore, D. R. (2021). *ECLIPS: Evaluation of Children's Listening and Processing Skills* (2nd ed.). Cincinnati Children's Hospital Medical Center.

Barry, J. G., Tomlin, D., Moore, D. R., Dillon, H. (2015). Use of questionnaire-based measures in the assessment of listening difficulties in school-aged children. *Ear Hear*, *36*, e300–e313.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Baum Associates.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sci Educ*, *15*, rm4.

Bradley, R. H., Convyn, R. F., Burchinal, M., McAdoo, H. P., Coll, C. G. (2001). The home environments of children in the United States part II: Relations with behavioral development through age thirteen. *Child Dev*, *72*, 1868–1886.

Brus, B. T., & Voeten, M. J. M. (1973). *Een-Minuut-Test*. Pearson Clinical.

De Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory*. Guilford Press.

De Sousa, K. C., Swanepoel, D. W., Moore, D. R., Myburgh, H. C., Smits, C. (2020). Improving sensitivity of the digits-in-noise test using antiphasic stimuli. *Ear Hear*, *41*, 442–450.

de Wit, E., van Dijk, P., Hanekamp, S., Visser-Bochane, M. I., Steenbergen, B., van der Schans, C. P., Luinge, M. R. (2018). Same or different: The overlap between children with auditory processing disorders and children with other developmental disorders: A systematic review. *Ear Hear*, *39*, 1–19.

de Wit, E., Visser-Bochane, M. I., Steenbergen, B., van Dijk, P., van der Schans, C. P., Luinge, M. R. (2016). Characteristics of auditory processing disorders: A systematic review. *J Speech Lang Hear Res*, *59*, 384–413.

DeBonis, D. A. (2015). It is time to rethink central auditory processing disorder protocols for school-aged children. *Am J Audiol*, *24*, 124–136.

Denys, S., Hofmann, M., Luts, H., Guérin, C., Keymeulen, A., Van Hoeck, K., Wouters, J. (2018). School-age hearing screening based on speech-in-noise perception using the digit triplet test. *Ear Hear*, *39*, 1104–1115.https://doi.org/10.1097/AUD.0000000000000563.

Denys, S., Hofmann, M., van Wieringen, A., Wouters, J. (2019). Improving the efficiency of the digit triplet test using digit scoring with variable adaptive step sizes. *Int J Audiol*, *58*, 670–677.

Denys, S., Wouters, J., van Wieringen, A. (2021). The digit triplet test as a self-test for hearing screening at the age of school-entry. *Int J Audiol*, *61*, 408–415.

Dillon, H., & Cameron, S. (2021). Separating the causes of listening difficulties in children. *Ear Hear*, *42*, 1097–1108.

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory*. Psychology Press.

Farah, R., Schmithorst, V. J., Keith, R. W., Holland, S. K. (2014). Altered white matter microstructure underlies listening difficulties in children suspected of auditory processing disorders: A DTI study. *Brain Behav*, *4*, 531–543.

Ferguson, M. A., Hall, R. L., Riley, A., Moore, D. R. (2011). Communication, listening, cognitive and speech perception skills in children with auditory processing disorder (APD) or Specific Language Impairment (SLI). *J Speech Lang Hear Res*, *54*, 211–227.

Guérin, C., Van Hoeck, K., Denys, S., van Wieringen, A., Wouters, J., Hoppenbrouwers, K. (2018). Systematische opsporing van lawaaischade bij jongeren. *JGZ Tijdschrift Voor Jeugdgezondheidszorg*, *50*, 132–137.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health Qual Life Outcomes*, *15*, 181.

Hind, S. E., Haines-Bazrafshan, R., Benton, C. L., Brassington, W., Towle, B., Moore, D. R. (2011). Prevalence of clinical referrals having hearing thresholds within normal limits. *Int J Audiol*, *50*, 708–716.

Humphry, S., & Montuoro, P. (2021). The Rasch model cannot reveal systematic differential item functioning in single tests: Subset DIF analysis as an alternative methodology. *Front Educ*, *6*, 1–8.

Hunter, L. L., Blankenship, C. M., Lin, L., Sloat, N. T., Perdew, A., Stewart, H., Moore, D. R. (2021). Peripheral auditory involvement in childhood listening difficulty. *Ear Hear*, *42*, 29–41.

Hwa-Froelich, D. A., & Vigil, D. C. (2004). Three aspects of cultural influence on communication: A literature review. *Commun Disord Q*, *25*, 107–118.

Jansen, S., Luts, H., Dejonckere, P., van Wieringen, A., Wouters, J. (2013). Efficient hearing screening in noise-exposed listeners using the digit triplet test. *Ear Hear*, *34*, 773–778.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*, *15*, 155–163.

Kreisman, N., John, A., Kreisman, B., Hall, J., & Crandell, C. (2012). Psychosocial status of children with auditory processing disorder. *J Am Acad Audiol*, *23*, 222–233.

Korkman, M., Kemp, S. L., Kirk, U. (2001). Effects of age on neurocognitive measures of children ages 5 to 12: A cross-sectional study on 800 children from the United States. *Dev Neuropsychol*, *20*, 331–354.

Kort, W., Compaan, E., Schittekatte, M., Dekker, P. (2008). *Clinical Evaluation of Language Fundamentals 4 Nederlandse Versie (CELF-4-NL)*. Pearson.

Kort, W., Schittekatte, M., Dekker, P. H., Verhaeghe, P., Compaan, E., Bosmans, M., Vermeir, G. (2005). WISC-III NL Wechsler Intelligence Scale for Children. In N. L. Derde Editie (Ed.), *Handleiding en Verantwoording*. Harcourt Test Publishers/Nederlands Instituut voor Psychologen.

Linacre, J. M. (n.d.). *Winsteps Rasch measurement computer program User's guide*. Beaverton.

Magimairaj, B. M., Nagaraj, N. K., Sergeev, A. V., Benafield, N. J. (2020). Comparison of auditory, language, memory, and attention abilities in children with and without listening difficulties. *Am J Audiol*, *29*, 710–727.

Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Van de Winckel, A. (2022). Rasch Reporting Guideline for Rehabilitation Research (RULER): The RULER Statement. *Arch Phys Med Rehabil*, *103*, 1477–1486.

Manly, T., Robertson, I. H., Anderson, V., Nimmo-Smith, I. (2004). *Tea-Ch: Test of Everyday Attention for Children*. Pearson.

Moore, D. R. (2012). Listening difficulties in children: Bottom-up and top-down contributions. *J Commun Disord*, *45*, 411–418.

Moore, D. R., Ferguson, M. A., Edmondson-Jones, A. M., Ratib, S., Riley, A. (2010). Nature of auditory processing disorder in children. *Pediatrics*, *126*, e382–e390.

Moore, D. R., Rosen, S., Bamiou, D.-E., Campbell, N. G., Sirimanna, T. (2013). Evolving concepts of developmental auditory processing disorder (APD): A British Society of Audiology APD special interest group "white paper.". *Int J Audiol*, *52*, 3–13.

Moore, D. R., Sieswerda, S. L., Grainger, M. M., Bowling, A., Smith, N., Perdew, A., Hunter, L. L. (2018). Referral and diagnosis of developmental auditory processing disorder in a large, United States hospital-based audiology service. *J Am Acad Audiol*, *29*, 364–377.

Mous, S. E., Schoemaker, N. K., Blanken, L. M. E., Thijssen, S., van der Ende, J., Polderman, T. J. C., White, T. (2017). The association of gender, age, and intelligence with neuropsychological functioning in young typically developing children: The Generation R study. *Appl Neuropsychol Child*, *6*, 22–40.

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *J Stat Distrib Appl*, *7*, 1–12.

Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., Sullivan, M. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Qual Life Res*, *12*, 373–385.

Petley, L., Hunter, L. L., Motlagh Zadeh, L., Stewart, H. J., Sloat, N. T., Perdew, A., Moore, D. R. (2021). Listening difficulties in children with normal audiograms: Relation to hearing and cognition. *Ear Hear*, *42*, 1640–1655.

Roebuck, H., & Barry, J. G. (2018). Parental perception of listening difficulties: An interaction between weaknesses in language processing and ability to sustain attention. *Sci Rep*, *8*, 6985.

Rothermich, K., Caivano, O., Knoll, L. J., Talwar, V. (2020). Do they really mean it? Children's inference of speaker intentions and the role of age and gender. *Lang Speech*, *63*, 689–712.

Schlichting, L. (2005). *Peabody Picture Vocabulary Test III-NL* (3 rd ed.). Harcourt, Test Publishers.

Seeto, M., Tomlin, D., Dillon, H. (2021). The relations between auditory processing scores and cognitive, listening and reading abilities. *Ear Hear*, *42*, 803–813.

Sharma, M., Purdy, S. C., Kelly, A. S. (2009). Comorbidity of auditory processing, language, and reading disorders. *J Speech Lang Hear Res*, *52*, 706–722.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol*, *8*, 33.

Stavrinos, G., Iliadou, V.-M., Edwards, L., Sirimanna, T., Bamiou, D.-E. (2018). The relationship between types of attention and auditory processing skills: Reconsidering auditory processing disorder diagnosis. *Front Psychol*, *9*, 34.

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. item response theory: Knowing when to cross the line. *Pract Assess Res Eval*, *26*, 1–16.

Stewart, H. J., Cash, E. K., Hunter, L. L., Maloney, T., Vannest, J., Moore, D. R. (2022). Speech cortical activation and connectivity in typically developing children and those with listening difficulties. *Neuroimage Clin*, *36*, 103172.

Tesio, L. (2012). Outcome measurement in behavioural sciences: A view on how to shift attention from means to individuals and why. *Int J Rehabil Res*, *35*, 1–12.

Tesio, L., Caronni, A., Kumbhare, D., Scarano, S. (2023a). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disabil Rehabil*, *46*, 591–603.

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., Scarano, S. (2023b). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disabil Rehabil*, *46*, 604–617.

Tesio, L., Scarano, S., Hassan, S., Kumbhare, D., Caronni, A. (2023c). Why questionnaire scores are not measures: A question-raising article. *Am J Phys Med Rehabil*, *102*, 75–82.

Tomlin, D., Dillon, H., Sharma, M., Rance, G. (2015). The impact of auditory processing and cognitive abilities in children. *Ear Hear*, *36*, 527–542.

Van de Winckel, A., Kozlowski, A. J., Johnston, M. V., Weaver, J., Grampurohit, N., Terhorst, L., Mallinson, T. (2022). Reporting Guideline for RULER: Rasch Reporting Guideline for Rehabilitation Research: Explanation and Elaboration. *Arch Phys Med Rehabil*, *103*, 1487–1498.

Van den Bos, K., Spelberg, H., Scheepstra, A., de Vries, J. (1999). *De Klepel Pseudowoordentest Vorm A en B*. Swets Test Publishers.

Erratum

Electrode Montage Induced Changes in Air-Conducted Ocular Vestibular-Evoked Myogenic Potential?: Erratum

In the article that published in the in the Jan-Feb 2024, volume 45 issue of *Ear and Hearing,* "Electrode Montage Induced Changes in Air-Conducted Ocular Vestibular-Evoked Myogenic Potential?" by Raveendran, R.K., and Singh, N.K., errors were discovered in the table titles by the authors:

The published paper titles reads:

TABLE 1. Mean, SD, median and inter-quartile range of latency and peak-to-peak amplitude of ocular vestibular-evoked myogenic potentials for various electrode montages

TABLE 2. Outcomes of the Wilcoxon signed-rank test for pairwise comparison of absolute latencies, peak-to-peak amplitude, and SNR between the electrode montages

TABLE 3. Mean, SD, median, and interquartile range of latency, peak-to-peak amplitude, and SNR of oVEMP recorded using various electrode montages across the test sessions, and the outcomes of Friedman's test for comparison of these parameters among the test sessions

TABLE 4. The intraclass correlation coefficient and Cronbach's alpha of latency, amplitude, and SNR of ocular vestibular-evoked myogenic potentials for various electrode montages

The correct titles are noted below:

TABLE 1. Mean, SD, median, and inter-quartile range of latency, peak-to-peak amplitude, inter-aural measures, signal-to-noise ratio, and morphology rating scores of ocular vestibular-evoked myogenic potentials for various electrode montages

TABLE 2. Outcomes of the Wilcoxon signed-rank test for pairwise comparison of absolute latency, peak-to-peak amplitude, inter-aural measures, signal-to-noise ratio, and morphology rating scores between the electrode montages

TABLE 3. Median, interquartile range of latency, peak-to-peak amplitude, inter-aural measures, and signal-to-noise ratio of oVEMP recorded using various electrode montages across the test sessions, and the outcomes of Friedman's test for comparison of these parameters among the test sessions

TABLE 4. The intraclass correlation coefficient and Cronbach's alpha of latency, peak-to-peak amplitude, inter-aural measures, and signal-to-noise ratio of ocular vestibular-evoked myogenic potentials for various electrode montages

The authors apologize for these errors.

Reference

Raveendran, R. K., & Singh, N. K. (2024). Electrode montage induced changes in air-conducted ocular vestibular-evoked myogenic potential. *Ear Hear*, *45*(1), 227–238.