ASHA American Speech-Language-Hearing Association

PERSPECTIVES

SIG 14

Research Article

First Steps Toward Implementation of the Online Test Battery LITMUS-NL: A Usability and Feasibility Study

Linda Wouda,[a] Tessel Boerma,[b] Ellen Gerrits,[b,c] and Elma Blom[a]

[a] Department of Development & Education of Youth in Diverse Societies, Faculty of Social and Behavioural Sciences, Utrecht University, the Netherlands [b] Department of Languages, Literature and Communication, Faculty of Humanities, Utrecht University, the Netherlands [c] Research Group Speech and Language Therapy: Participation Through Communication, HU University of Applied Sciences Utrecht, the Netherlands

ARTICLE INFO

Article History:
Received December 29, 2023
Revision received April 11, 2024
Accepted July 8, 2024

Editor-in-Chief: Celeste Domsch
Editor: Danai Kasambira Fannin

https://doi.org/10.1044/2024_PERSP-23-00308

ABSTRACT

Purpose: Language Impairment Testing in Multilingual Settings (LITMUS) instruments were developed to improve the identification of developmental language disorders in multilingual children. The current study investigated the usability and feasibility of the online Dutch version of several of these instruments (LITMUS-NL) according to speech-language pathologists (SLPs), thereby taking the first steps toward implementation in clinical practice.
Method: We first conducted a usability study in which 24 SLPs performed the LITMUS-NL tests while using a think-aloud protocol. They subsequently filled out a questionnaire to investigate the degree of usability and added value of LITMUS-NL. After adapting LITMUS-NL based on the results of the usability study, a feasibility study was carried out in which 25 other SLPs each used LITMUS-NL with three multilingual children. Afterward, they completed a feasibility questionnaire and questionnaires about the reactions of the children.
Results: In the first study, many usability issues emerged, mainly concerning technical problems, instructions, and test construction. Despite these issues, the SLPs evaluated the degree of usability and added value as positive. The feasibility study revealed a lower degree of usability and, despite the adaptations, feasibility issues in the same categories.
Conclusions: By involving the intended users in the process toward implementing a new product, we identified (and solved) many issues that would interfere with successful implementation in daily clinical practice. A systematic, iterative approach toward implementation helps identify what is deemed important by the intended users of a new product and bridges the gap between research and practice.
Supplemental Material: https://doi.org/10.23641/asha.26864344

Worldwide, many children grow up learning multiple languages (Grosjean & Pavlenko, 2021). In the Netherlands, where the current study was situated, a recent survey showed that a quarter of the respondents used another language at home, in addition to Dutch (Schmeets & Cornips, 2021). In the United States, more than one fifth of the population speaks a language other than English at home (Dietrich & Hernandez, 2022), and it often goes unnoticed that in the Global South, multilingualism is the rule rather than the exception (Léglise, 2017). Children possess a natural aptitude for acquiring multiple languages, yet if they learn languages sequentially, they may face greater challenges compared to those who acquire the languages from birth, especially in the earlier stages when exposure is still limited (Bedore & Peña, 2008; Peña et al., 2020). Several studies have shown that children with language delays due to insufficient

Correspondence to Elma Blom: w.b.t.blom@uu.nl. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

language exposure and children with a developmental language disorder (DLD) tend to struggle with similar aspects of the language (Bedore & Peña, 2008; Boerma, 2017; Kohnert, 2010; Orgassa & Weerman, 2008; Paradis, 2010). This overlap in language difficulties can lead to the over- and underdiagnosis of DLD in multilingual children: A language disorder can be interpreted as caused by insufficient exposure to the target language, or language delays are mistakenly identified as a DLD. In both cases, children may receive support and education that are not optimal for their development (Bedore & Peña, 2008; Boerma et al., 2015; Grimm & Schulz, 2014).

Research in the past years has been dedicated to supporting a reliable diagnosis of DLD in multilingual children, as this is neither simple nor straightforward (see, e.g., Boerma & Blom, 2017; Peña et al., 2020). To this end, new language assessment instruments, for example, the Language Impairment Testing in Multilingual Settings (LITMUS) tests, have been developed. Such tests have proven to be successful in distinguishing between language delay and language disorder (e.g., Boerma, 2017; Boerma et al., 2015, 2016; Chiat & Polišenská, 2016). However, subsequent steps toward the implementation of such new instruments in clinical practice are often forgotten, although these are crucial for actual progress in the field with respect to a better identification of DLD. Many newly developed tests do not cross research boundaries despite showing promise for clinical practice, potentially leading to research waste (e.g., Macleod et al., 2014). The challenges to identify DLD in multilingual children therefore persist in daily clinical practice. With the ultimate goal of bringing promising research findings to the professional work field, we report on a systematic investigation of the first steps that can be taken toward the implementation of new assessment instruments, setting an example for fellow researchers and stressing the importance of the involvement of professionals (i.e., end users) in this process.

A New Test Battery for Multilingual Children: LITMUS-NL

LITMUS refers to various tests that were developed as a result of a European Cooperation in Science and Technology (COST) Action.[1] The research in this COST Action aimed to study typical and atypical multilingual language development across many languages to distinguish between features of DLD and multilingualism and to improve the language assessment of multilingual children (Armon-Lotem et al., 2015). The development of these new tests was necessary, as standardized instruments used for language assessment with monolingual children often disadvantage children with less experience in the target language (Armon-Lotem et al., 2015; Bedore & Peña, 2008; Paradis, 2010). Standardized instruments with multilingual norms could do justice to these specific language experiences. However, because multilingual experiences vary between children, multilingual norms will apply only to a limited group of children (Bedore & Peña, 2008; Kohnert, 2010; Paradis, 2010; Pearson, 2013). Another option is to assess all languages of multilingual children (American Speech-Language-Hearing Association, 2004), but skilled interpreters, translators, and multilingual speech-language pathologists (SLPs) are often not available. In addition, these dual-language assessments are time consuming for the SLPs and burdensome to the families. Moreover, the results of the assessments are difficult to interpret, as validated reference scores are lacking for multilingual children (Hamann & Abed Ibrahim, 2017; Kohnert, 2010).

To overcome the challenges related to multilingual norms and dual-language testing, a comprehensive set of instruments was developed within the aforementioned COST Action: LITMUS.[2] The LITMUS cross-linguistic nonword repetition task (CL-NWR; Chiat, 2015), narrative task (Multilingual Assessment Instrument for Narratives [MAIN]; Gagarina et al., 2019), sentence repetition task (SRep; Marinis & Armon-Lotem, 2015), and cross-linguistic lexical task (CLT; Simonsen & Haman, 2017) as well as a risk index for early language development based on parental report (Tuller, 2018) were translated into Dutch and studied within the Dutch context (Boerma & Blom, 2017; Boerma et al., 2015, 2016; de Jong et al., 2021; van Wonderen et al., 2017). These five measures were combined and developed into an online instrument, which resulted in the first test battery with multiple LITMUS tests: LITMUS-NL. LITMUS-NL was not available to SLPs prior to the current study. The paper versions of the MAIN and CL-NWR were used by a few SLPs working in multidisciplinary speech and hearing centers to support the diagnosis of multilingual children. Norm data were not available at the time.

LITMUS-NL consists of Core and Additional subtests, together with a manual, norm scores, and a reports section presenting the results. The LITMUS-NL Core subtest battery includes the CL-NWR, MAIN, and risk index. The CL-NWR assesses phonological short-term memory; the MAIN evaluates macrostructural narrative abilities; and the risk index is based on a parental questionnaire, including questions on early developmental language

[1]COST Action IS0804: "Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment"

[2]For more information, see https://www.bi-sli.org.

milestones and parental concerns (Armon-Lotem et al., 2015). These three measures do not draw on abilities or knowledge specific to one language but, instead, take into account skills that children have acquired in any language. In contrast to many traditional standardized language tests, these measures are therefore not biased against multilingual children, as has been shown in previous studies (Boerma & Blom, 2017; Boerma et al., 2015, 2016; Chiat & Polišenská, 2016). Boerma and Blom (2017) evaluated the diagnostic accuracy of these Core subtests and were the first to combine different LITMUS instruments. The study showed that the combination of the CL-NWR, MAIN, and risk index resulted in excellent diagnostic accuracy and reliable identification of DLD in monolingual and multilingual children. As such, the Core subtests can play a pivotal role in the early phase of the diagnostic process, supporting the identification of multilingual children who are both likely and unlikely to have DLD. In assessing language skills that are relatively independent of properties of specific languages, the Core subtests of LITMUS-NL are unique. LITMUS-NL should, however, not be used as a stand-alone instrument; to diagnose DLD, a more elaborate insight is needed into the different aspects of a child's language skills, in addition to excluding the possible causes of language difficulties. Based on the results of Boerma and Blom, norm scores for the Core subtests were developed during the current research. As performance on these tests does not rely on language-specific experience, norms are equal for monolingual and multilingual children; hence, for the Core subtests, there is no need for separate norms.

The SRep and CLT are additional tests that are included in LITMUS-NL. They assess children's Dutch grammar and vocabulary skills, respectively, and are more dependent on experience with Dutch than the Core measures. They are included in LITMUS-NL to provide detailed information on the strengths and weaknesses of a child's language abilities in Dutch, which can, for example, be used to set therapy goals. The SRep assesses the processing of sentences and representations in long-term memory (Marinis & Armon-Lotem, 2015) and contains sentence structures that are difficult for children with DLD across languages (Marinis & Armon-Lotem, 2015). The CLT consists of tasks for receptive and expressive noun and verb vocabulary (Simonsen & Haman, 2017). The SRep and CLT are developed in multiple languages, which might enable comparing a multilingual child's abilities in different languages. Norm data for the LITMUS-NL Additional measures are not (yet) available.

Steps Toward Implementation

Even though previous research with the tests included in LITMUS-NL showed promising results (Boerma & Blom, 2017), this does not necessarily mean that implementation in clinical practice will be a success. Implementation refers to the elaborate process of several activities to introduce and maintain a product or an intervention in a particular context (van Gemert-Pijnen, 2022; World Health Organization [WHO], 2016). This process is iterative, meaning that it consists of repeated cycles of evaluation and adaptation (O'Cathain et al., 2019). Activities that are recommended during the implementation of digital instruments include the evaluation of usability and feasibility in the early stages of the iterative process and of effectiveness and affordability in later stages (WHO, 2016). In this current study, we report on the early stages of implementation of LITMUS-NL, investigating its usability and feasibility.

The LITMUS tests were developed by multidisciplinary groups of linguists, psychologists, and researchers in the field of speech-language pathology, but what these working groups deem important could differ from the opinions and practices of SLPs working in daily clinical settings. One key aspect to increasing successful implementation is engaging the intended users (Gravitt, 2023; Skivington et al., 2021). This can be done systematically through investigations of usability and feasibility, which offer insight into the intended users and user preferences (WHO, 2016). Usability and feasibility studies, conducted in a controlled and an uncontrolled setting, respectively, result in a list of improvements to the products recommended by the intended users.

Usability investigations assess whether the new product is used as intended and enable an understanding of the setting in which the product is used (O'Cathain et al., 2019; WHO, 2016). A usability study evaluates the domains of user performance (i.e., how users interact with the product), user satisfaction, and acceptability (i.e., the way intended users and those involved in implementation—in this case, multilingual children—react to the product) to describe the usability of the product as well as to find and solve usability problems (Bevan et al., 1991; Bowen et al., 2009; Sauro & Lewis, 2016).

Usability is often studied in controlled situations to refine the product (Bevan et al., 1991; Fonda et al., 2008; O'Cathain et al., 2019). However, the product will eventually be used in daily clinical practice. Therefore, a necessary step after the investigation of usability is conducting a feasibility study, during which the product is used in uncontrolled "real-life" settings. The domains of user performance, user satisfaction, and acceptability are then studied in a specific context. Feasibility investigations assess whether the product will be used as intended in this specific context—for example, in clinical practice—and help identify aspects of the product that should be redesigned (Bowen et al., 2009; Mulkey et al., 2019; WHO, 2016).

The Current Study

LITMUS tests have been developed in many languages and have, as of yet, mainly been used in research to enhance our insight into multilingual language development or to evaluate the diagnostic potential. The current study presents the first test battery that combines multiple LITMUS tests, integrated into an online environment that is designed for clinical use: LITMUS-NL. We are also one of the few to move beyond research boundaries and work toward the implementation of this test battery in clinical practice, reducing research waste (a need emphasized by Kulkarni et al., 2022). The current study can hereby function as a model for others working with LITMUS tests and for those working with newly developed language assessment instruments in general.

In this current study, we report on the early stages of implementation of LITMUS-NL in daily clinical practice, investigating its usability and feasibility according to SLPs. These two steps in the implementation process are decisive for the successful use of a digital instrument in the work field (Bowen et al., 2009; Skivington et al., 2021; WHO, 2016). SLPs are the intended users of LITMUS-NL, for whom the instrument needs to work optimally and who should be willing to use it. Their involvement in the development of LITMUS-NL is therefore crucial. In the usability study, SLPs used LITMUS-NL by themselves in a controlled setting. After improving LITMUS-NL based on the feedback of the participating SLPs, a feasibility study was conducted. In the feasibility study, SLPs used LITMUS-NL with multilingual children in daily clinical practice.

Usability Study

Method

The usability study was performed from February 2021 to April 2021. The study was approved by the Ethics Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University.

Participants

Twenty-five SLPs participated in the study. They were recruited by promoting the study on social media. All were female and worked with multilingual children. One participant was excluded because the instructions were not followed properly. The final sample consisted of 16 SLPs working in speech-language pathology practices and eight SLPs working in speech and hearing centers. In the Netherlands, language assessments are often carried out in speech-language pathology practices before starting language therapy. However, when DLD is suspected, (multilingual) children are also assessed with additional tests or dual-language assessments by SLPs in speech and hearing centers. Therefore, we included both SLPs in speech and hearing centers and SLPs in speech-language pathology practices.

Demographic data concerning gender, age, work setting, work experience, and self-reported digital literacy were collected with an online questionnaire to describe the participant characteristics. The data are displayed in Table 1. Self-reported digital literacy, which was relevant given the online nature of LITMUS-NL, was assessed with a visual analog scale (VAS) from 1 (*low*) to 10 (*high*).

Instruments and Procedure

LITMUS-NL was recently developed as part of an online testing environment, namely, the Utrecht University Developmental Assessment Battery (UU-DAB),[3] by the IT services of Utrecht University. The participating SLPs received an account to log in to the online environment of LITMUS-NL. After logging in, the SLPs performed the CL-NWR, MAIN, SRep, and CLT tests by themselves while thinking aloud. The risk index and the norm scores for the Core subtests (CL-NWR, MAIN, and risk index) were not yet available during the usability study and were therefore not included in the protocol. A synchronized concurrent think-aloud protocol (TAP) was performed to collect usability issues (Fonda et al., 2008; Stefano et al., 2010). While performing the four tests of LITMUS-NL, the SLPs verbalized their thoughts and experiences. Audio recordings were made, which allowed for a systematic analysis of these thoughts and experiences. This was used to identify unsatisfactory features and usability issues of the instrument (Fonda et al., 2008; van Velsen et al., 2011), which could cover all domains of usability, including user experience, user satisfaction, and acceptability.

After finishing the LITMUS-NL test battery, the participants completed the online usability questionnaire. We used questionnaires because they offer the opportunity to collect data from many subjects with low costs and high flexibility (Gillham, 2008). The outcome parameters of the usability questionnaire were categorized into three domains and are presented in Supplemental Material S1.

The user performance domain was measured with a standardized scale indexing the degree of usability—the System Usability Scale (SUS; Brooke, 1996). The SUS is a standardized 10-item scale with a 5-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*; Bangor et al., 2009; Brooke, 1996). The SUS is most frequently

[3]Available through https://dab.sites.uu.nl.

Table 1. Participant characteristics of the usability study.

Variable	Value
Age (in years)	
M (*SD*)	33.50 (8.23)
Range	25.00–56.00
Gender	
Number of females	24
Work experience as an SLP (in years)	
M (*SD*)	9.88 (6.70)
Range	0.50–24.00
Work setting	
Number of SLPs in speech-language pathology practices	16
Number of SLPs in speech and hearing centers	8
Self-reported digital literacy	
M (*SD*)	8.23/10 (0.70)
Range	7.00/10–10.00/10

Note. SLP(s) = speech-language pathologist(s).

used for investigating usability since it is a short scale and is found to be an effective and reliable instrument (Bangor et al., 2008). However, the SUS is inadequate as a standalone tool, as it provides no information on why a particular score is achieved (Bangor et al., 2009; Broekhuis et al., 2019; Sauro & Lewis, 2016). Therefore, secondary parameters were added to investigate the domain of user performance. The intuitiveness of buttons in the digital environment and the intuitiveness of all the score sheets were collected with VAS questions from 1 (*negative*) to 10 (*positive*) per subtest of LITMUS-NL. SLPs were further asked to estimate the time in minutes to complete LITMUS-NL with a child in clinical practice.

The user satisfaction domain was also assessed with VAS questions. The VAS questions focused on the complexity and design of LITMUS-NL, the expected applicability in clinical practice with multilingual children, and the applicability of the duration of the test battery. Because SLPs cope with a high workload, time is an important factor for applicability in daily practice (Greenwell & Walsh, 2021). As the MAIN subtest was the only test that had to be scored manually, an additional question about the applicability of this manual scoring was included.

The acceptability domain was investigated by including a standardized scale assessing the added value of LITMUS-NL and VAS questions about the insight that the instrument could provide into children's language skills. The added value of LITMUS-NL was studied using the standardized Value/Usefulness subscale of the Intrinsic Motivation Inventory (IMI; Ryan, 1982). This subscale consists of seven questions that are scored on a 7-point Likert scale from *not at all true* to *very true* (Ryan, 1982).

Data Analysis

The audio recordings of the TAP were transcribed verbatim. All problems or questions in the transcriptions were coded to collect initial usability issues (Ehrler et al., 2018; Fonda et al., 2008). Similar issues mentioned by multiple SLPs were marked as one usability issue (following a procedure similar to that in Fonda et al., 2008). Finally, the usability issues were categorized into five issue categories.

The questionnaire data were anonymized after data collection. The answers to the scale questions were imported into SPSS 25, and the answers to the open-ended questions were imported into Microsoft Excel. Not all data were distributed normally; therefore, both mean and median scores were calculated for all questions with responses on a scale. As mentioned, two standardized scales, several VAS questions, and an open question about the test duration were included in the questionnaire. First, the standardized scale tapping into the degree of usability (user performance domain) was analyzed by calculating score contributions for each question (Brooke, 1996). The score contribution for positively stated items (odd questions) was scale position minus 1, and that for negatively stated items (even questions) was 5 minus scale position. The overall SUS score, ranging from 0 (*negative*) to 100 (*positive*), was calculated as the sum of the score contributions multiplied by 2.5 (Bangor et al., 2009; Ehrler et al., 2018). The mean SUS score was analyzed and then compared to the adjective rating of Bangor et al. (2009), stating that a score > 70 was acceptable (Bangor et al., 2008, 2009). In addition, median scores were calculated for each question. Second, the standardized scale tapping into added value (acceptability domain) was analyzed by calculating the mean score of the Value/Usefulness subscale of the IMI. The score was regarded positively if the score was > 4 (Gerber et al., 2019; Ryan, 1982). Median scores for each question of the scale were also calculated.

For all VAS questions, the mean and median were calculated. A score > 5 was interpreted as a positive score, as this score was higher than the middle point between the two extreme alternatives (Abascal et al., 2018). The open question on the duration of LITMUS-NL was analyzed by calculating the mean duration, which led to an expected time in minutes to complete LITMUS-NL.

Results

CL-NWR data were missing from one participant (MAIN, CLT, and SRep data were available), and two other participants forgot to audio-record the CLT (all other data were available). All available audio recordings and questionnaires (n = 24) were included in the analysis.

Usability Issues Resulting From the TAP

A total of 337 different usability issues were identified by analyzing the transcriptions of the TAP recordings. Table 2 presents these usability issues per LITMUS-NL test, which shows that most issues were related to the CLT and SRep (LITMUS-NL Additional). All usability issues were divided into five categories, some containing multiple subcategories. Most usability issues pertained to test construction, including issues with the length or difficulty of a test for the targeted population or issues with a specific item or picture within a test. In addition to issues related to test construction, technical issues were also frequently reported, which concerned the design of the online environment, user experience, and bugs. A total of seven out of 24 SLPs experienced technical difficulties and could not start the tests without help. The final three categories reflected issues related to test instructions, spelling errors, and issues emerging from the pronunciation of items that were read aloud by the digital testing environment.

Outcomes of the Questionnaire

The outcome parameters derived from the questionnaire are presented in Table 3. We found that the mean degree of usability, measured with the SUS, was 72.40 (out of 100). This degree of usability is rated "good" (Bangor et al., 2009). The mean perceived added value, measured with a subscale of the IMI, was 5.60 (out of 7). Since this score is > 4, the result is positive (Gerber et al., 2019; Ryan, 1982). The median scores for the individual items of both standardized scales are presented in Appendix A.

All other parameters measured with VAS questions were scored positively (all > 7 out of 10), although there was variation between SLPs as indicated by wide ranges in scores. Several SLPs specifically commented that some pictures were outdated and that they found LITMUS-NL difficult for 3-year-old children, addressing the need for stop rules, especially for the LITMUS-NL Additional measures. The mean time of the participants to complete the test battery was 45 min.

Adapting LITMUS-NL

LITMUS-NL was adapted based on the results of the usability study. The issues related to the test construction, representing 37.1% (125/337) of the usability issues, could not be solved by us, because they require the involvement of the working group that developed the test. Moreover, most of the construct-related issues concerned the LITMUS-NL Additional subtests. Solving issues related to the Core subtests was given a higher priority, as these subtests can contribute to a more reliable diagnosis of DLD in multilingual children and thereby fill an urgent need. Of the 212 non–construct-related issues, 69.3% (147/212) were solved, for example, by making adaptations in the online

Table 2. Categories of usability issues with frequencies of issues per LITMUS-NL (Dutch version of Language Impairment Testing in Multilingual Settings) subtest.

Issue category and subcategory		LITMUS-NL Core			LITMUS-NL Additional			
	General	CL-NWR	MAIN comp.	MAIN prod.	SRep	CLT comp.	CLT prod.	Total
Instructions								
General	5	8	3	7	14	9	14	60
Scoring	0	2	3	0	2	0	4	11
Technical								
Performance	5							5
Bugs	3	5	3	3	6	2	8	30
User experience	5	3	4	2	5	5	9	33
Notifications	3	0	0	0	1	3	9	16
User design	1	6	2	2	15	6	3	35
Features	0	0	0	1	0	0	0	1
Test construction								
General	3	2	3	1	6	11	8	34
Item-specific content	0	1	3	0	4	13	20	41
Item-specific presentation	0	0	2	0	2	26	20	50
Spelling errors	0	0	0	1	0	2	6	9
Pronunciation of items	0	10			2			12
Total	25	37	23	17	57	77	101	**337**

Note. CL-NWR = cross-linguistic nonword repetition task; MAIN = Multilingual Assessment Instrument for Narratives; comp. = comprehension; prod. = production; SRep = sentence repetition task; CLT = cross-linguistic lexical task.

Table 3. Summarized scores of standardized scales and visual analog scale (VAS) questions.

Domain	Parameter	*M* (*SD*)[a]	*Mdn*	Min	Max	Range
User performance	Degree of usability[b]	72.40 (8.13)	73.75	50.00	90.00	40.00
	Intuitiveness of buttons					
	Nonword repetition	7.50 (1.51)	8.00	4.00	10.00	6.00
	MAIN	7.36 (1.73)	7.75	2.40	10.00	7.60
	Sentence repetition	7.27 (1.38)	7.60	3.90	9.00	5.10
	CLT	7.06 (1.90)	7.80	3.00	10.00	7.00
	Intuitiveness of score sheet					
	Nonword repetition	8.04 (1.23)	8.00	6.00	10.00	4.00
	MAIN	7.08 (1.74)	8.00	2.00	9.00	7.00
	Sentence repetition	7.46 (1.74)	8.00	2.00	10.00	8.00
	CLT	7.38 (1.47)	8.00	4.00	9.00	5.00
	Duration in minutes	45.00 (11.83)	42.50	29.00	67.00	38.00
User satisfaction	Complexity	7.45 (1.30)	8.00	4.00	9.00	5.00
	Design	7.27 (1.27)	7.80	3.00	8.60	5.60
	Applicability					
	Test duration	7.96 (1.74)	8.10	1.00	10.00	9.00
	With children	7.17 (0.97)	7.00	5.00	9.20	4.20
	Manual scoring MAIN[c]	7.75 (2.27)	8.00	0.00	10.00	10.00
Acceptability	Added value[d]	5.60 (0.90)	5.86	3.57	7.00	3.43
	Insight into language skills	7.14 (1.13)	7.20	2.90	8.50	5.60

Note. Min = minimum; Max = maximum; MAIN = Multilingual Assessment Instrument for Narratives; CLT = cross-linguistic lexical task.
[a]All VAS questions were scored on a scale ranging from 0 to 10 and are stated positively, meaning a higher score is interpreted as a positive outcome. [b]The degree of usability was scored with the System Usability Scale. The maximum score is 100. [c]Not all subtests were developed online completely; the production task of the MAIN was scored on paper. [d]Added value was scored with the Value/Usefulness subscale of the Intrinsic Motivation Inventory. The maximum score of this scale is 7.

environment and by creating an elaborate test manual for the users to read before working with LITMUS-NL. The remaining issues were not solved because they were not deemed urgent or because of technical limitations, the inconsistency of bugs, or causes that could not be identified.

To enhance the insight that LITMUS-NL could provide into the children's language skills, a reports section was added. This reports section contained norm data for the Core subtests based on Dutch research data (Boerma, 2017) and included a score sheet to complete the questions of the risk index. The adapted version of LITMUS-NL was piloted with three SLPs before the start of the feasibility study, which is described in the following section.

Feasibility Study

Method

The feasibility study was conducted from February 2022 to May 2022 and involved using LITMUS-NL with multilingual children in daily clinical practice. The feasibility study was approved by the Ethics Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University.

Participants

Twenty-seven SLPs participated in the study. The SLPs who participated in the feasibility study did not participate in the usability study. The participants were recruited by promoting the study on social media and by contacting SLPs who already showed interest in the LITMUS-NL test battery. A total of 49 SLPs agreed to participate in the study; however, 45% dropped out before starting with the study procedures. All these dropouts indicated to have insufficient time to administer LITMUS-NL next to their daily work. Out of the 27 SLPs who started, two were excluded during the study period. One participant was not able to start the digital testing environment and lacked time to do the assessments after extra instructions. The other SLP did not permit the researchers to use her answers in the questionnaire, and therefore, her results were excluded. The final sample thus included 25 SLPs, who were all female and worked with multilingual children. Of these 25, two worked in speech and hearing centers, whereas 23 worked in speech-language pathology practices. The demographic data of the final sample of 25 SLPs are presented in Table 4.

Instruments and Procedure

Each SLP performed the Core and Additional subtests of LITMUS-NL in clinical practice with three different

Table 4. Participant characteristics of the feasibility study.

Variable	Value
Age (in years)	
M (*SD*)	39.60 (10.80)
Range	22.00–58.00
Gender	
Number of females	25
Work experience as an SLP (in years)	
M (*SD*)	15.90 (11.27)
Range	1.50–36.50
Work setting	
Number of SLPs in speech-language pathology practices	23
Number of SLPs in speech and hearing centers	2
Self-reported digital literacy	
M (*SD*)	7.40/10 (1.94)
Range	3.00/10–10.00/10

Note. SLP(s) = speech-language pathologist(s).

multilingual children between the ages of 5;0 (years;months) and 8;11. In the feasibility study, the three outcome measures of the MAIN (comprehension, production, and internal state terms) were distinguished to learn more about the different subcomponents of the test.

The feasibility of LITMUS-NL, comparable to the usability, was studied with multiple parameters within the domains of user performance, user satisfaction, and acceptability. The data were collected through two online questionnaires: a feasibility questionnaire, which was completed after all three assessments, and a questionnaire considering the reactions of children, which was completed after each assessment. The domains, parameters, and instruments are schematically displayed in Supplemental Material S2.

The feasibility questionnaire consisted of standardized scales, including the SUS and IMI (see above for a description), as well as additional VAS questions and open-ended questions. These questions were comparable to those used in the usability study but adapted to the difference in setting. Moreover, questions about the interpretation of the children's scores and the reports section were added. The questionnaire considering the observations of the reactions of children addressed the clarity of instructions for children, their motivation, their attention, and the engagement of the child for each subtest of LITMUS-NL.

Data Analysis

Questionnaire data were anonymized. The data from the standardized scales and VAS scales were imported into SPSS 28 (IBM Corporation, 2020), and the answers to the open-ended questions were imported into Microsoft Excel (Microsoft Corporation, 2020).

Not all data were distributed normally; therefore, both mean and median scores were calculated for all questions with responses on a scale and were interpreted as comparable to those in the usability study. The answers to the open-ended questions were used to detect any remaining issues with LITMUS-NL, the manual, and the digital environment. The answers to these questions turned out to be insightful. We therefore chose to analyze these answers to identify and categorize new issues that we could solve. Per open-ended question, all answers were labeled independently by two researchers. They labeled problems the SLPs addressed. These labels were subsequently merged into overarching problems through discussion. These overarching problems represented the feasibility issues of LITMUS-NL.

Results

The SLPs used LITMUS-NL with 74 multilingual children. All of these children experienced language problems. Information about potential diagnoses (e.g., DLD) of the children was unknown to us. Due to technical errors, a number of questionnaires were not completed. Therefore, the number of participants for the feasibility questionnaire ranged between 22 and 25, and that for the reactions of children ranged between 72 and 74.

Outcomes of the Feasibility Questionnaire

The outcomes on the standardized scales and VAS questions are presented per test in Table 5. The degree of usability received a mean score of 65.30 (out of 100). According to Bangor et al. (2009), this score is "OK." The added value of LITMUS-NL, assessed with the Value/Usefulness subscale of the IMI, was positive with a mean score of 5.44 (out of 7). The scores on the individual items of both scales are presented in Appendix B. The instructions and scoring procedures were overall clear, although scoring the MAIN production test and MAIN internal state terms was more difficult. Several SLPs not only expressed the need for extra instructions or instructional videos but also indicated that instructions and scoring became clearer after using LITMUS-NL multiple times. The parameters within the user satisfaction domain received overall positive scores. The SLPs gave a positive assessment of the applicability in terms of duration, with mean scores of 7.57 for the entire test battery and 8.26 (out of 10) for the Core subtests. The applicability with children was scored lower, with an *M* of 6.78 (out of 10).

Acceptability: Reactions of Children

The children who were assessed with the LITMUS-NL test battery were between the ages of 5;0 and 8;11. Most children (43%) were between the ages of 5;0 and 5;11. Of all LITMUS-NL assessments, 60.8% were split up into two

Table 5. Summarized scores of standardized scales and visual analog scale (VAS) questions in the feasibility questionnaire.

Domain	Parameter	*M* (*SD*)[a]	*Mdn*	Min	Max	Range
User performance	Degree of usability[b]	65.30 (12.51)	67.50	35.00	90.00	55.00
	Clarity of instructions					
	Nonword repetition	8.80 (1.12)	9.00	7.00	10.00	3.00
	MAIN comprehension	8.56 (1.39)	9.00	4.00	10.00	6.00
	MAIN production	7.72 (1.97)	8.00	3.00	10.00	7.00
	Risk index	8.60 (1.85)	9.00	1.00	10.00	9.00
	Sentence repetition	8.52 (1.30)	9.00	6.00	10.00	4.00
	CLT comprehension	9.00 (1.00)	9.00	7.00	10.00	3.00
	CLT production	8.88 (1.13)	9.00	6.00	10.00	4.00
	Clarity of scoring					
	Nonword repetition	8.70 (1.06)	9.00	6.00	10.00	4.00
	MAIN comprehension	8.30 (1.40)	8.00	5.00	10.00	5.00
	MAIN production	6.04 (2.46)	6.00	1.00	10.00	9.00
	MAIN internal state terms	5.87 (2.60)	6.00	1.00	10.00	9.00
	Risk index	8.78 (1.13)	9.00	6.00	10.00	4.00
	Sentence repetition	8.48 (1.31)	9.00	5.00	10.00	5.00
	CLT comprehension	8.78 (1.13)	9.00	6.00	10.00	4.00
	CLT production	8.70 (1.19)	9.00	6.00	10.00	4.00
	Intuitiveness of buttons					
	Nonword repetition	7.88 (2.03)	8.00	0.00	10.00	10.00
	MAIN comprehension	7.75 (2.05)	8.00	0.00	10.00	10.00
	MAIN production	7.29 (2.39)	8.00	0.00	10.00	10.00
	Risk index	7.58 (2.10)	8.00	0.00	10.00	10.00
	Sentence repetition	7.54 (2.09)	7.50	0.00	10.00	10.00
	CLT comprehension	7.71 (2.03)	8.00	0.00	10.00	10.00
	CLT production	7.67 (2.01)	8.00	0.00	10.00	10.00
	Interpretation of results					
	Nonword repetition	6.91 (2.61)	8.00	1.00	10.00	9.00
	MAIN comprehension	6.96 (2.64)	8.00	1.00	10.00	9.00
	MAIN production	6.65 (2.37)	8.00	1.00	10.00	9.00
	MAIN internal state terms	6.39 (2.62)	7.00	1.00	10.00	9.00
	Risk index	7.30 (2.46)	8.00	1.00	10.00	9.00
User satisfaction	Usefulness of report					
	Nonword repetition	6.57 (3.15)	8.00	1.00	10.00	9.00
	MAIN comprehension	6.96 (2.65)	8.00	1.00	10.00	9.00
	MAIN production	7.04 (2.72)	8.00	1.00	10.00	9.00
	MAIN internal state terms	6.78 (2.66)	7.00	1.00	10.00	9.00
	Risk index	6.57 (2.91)	7.00	1.00	10.00	9.00
	Applicability with children	6.78 (2.22)	7.00	1.00	10.00	9.00
	Applicability of duration					
	Core test battery	8.26 (1.45)	9.00	4.00	10.00	6.00
	Full test battery	7.57 (1.93)	8.00	2.00	10.00	8.00
Acceptability	Added value[c]	5.44 (1.47)	5.86	2.00	7.00	5.00

Note. Min = minimum; Max = maximum; MAIN = Multilingual Assessment Instrument for Narratives; CLT = cross-linguistic lexical task.

[a]Not all parameters were distributed normally; therefore, both the mean and median scores were calculated. All VAS questions were scored on a scale ranging from 0 to 10 and are stated positively, meaning a higher score is interpreted as a positive outcome. [b]The degree of usability was scored with the System Usability Scale. The maximum score is 100. [c]Added value was scored with the Value/Usefulness subscale of the Intrinsic Motivation Inventory. The maximum score of this scale is 7.

different test sessions, 27% were done in three sessions, and 12.2% were done in only one session. The mean and median scores on the observations of the reactions of the children per LITMUS-NL test are presented in Table 6.

The children needed extra stimulation for the successful completion of the MAIN production. SLPs commented that some children were hesitant and needed encouragement to tell a story. The SRep received low

Table 6. Observations of reactions of children.

Parameter	*M* (*SD*)[a]	*Mdn*	Min	Max	Range
Clarity of instructions					
Nonword repetition	7.97 (2.02)	8.00	0.00	10.00	10.00
MAIN comprehension	8.04 (1.80)	8.00	2.00	10.00	9.00
MAIN production	7.72 (1.98)	8.00	2.00	10.00	8.00
Sentence repetition	7.12 (2.33)	7.50	0.00	10.00	10.00
CLT comprehension	8.78 (1.48)	9.00	0.00	10.00	10.00
CLT production	8.73 (1.50)	9.00	0.00	10.00	10.00
Motivation					
Nonword repetition	7.68 (2.40)	8.50	1.00	10.00	9.00
MAIN comprehension	7.68 (2.31)	8.00	2.00	10.00	8.00
MAIN production	6.58 (2.64)	6.50	1.00	10.00	9.00
Sentence repetition	4.76 (3.07)	5.00	0.00	10.00	10.00
CLT comprehension	8.07 (2.29)	9.00	0.00	10.00	10.00
CLT production	7.69 (2.52)	8.50	0.00	10.00	10.00
Attention span during test					
Nonword repetition	8.15 (1.88)	9.00	1.00	10.00	9.00
MAIN comprehension	7.43 (2.07)	8.00	2.00	10.00	8.00
MAIN production	7.53 (2.02)	8.00	2.00	10.00	8.00
Sentence repetition	5.46 (2.58)	5.00	0.00	10.00	10.00
CLT comprehension	7.91 (2.16)	8.00	0.00	10.00	10.00
CLT production	7.82 (2.18)	8.00	0.00	10.00	10.00
Reaction to test					
Nonword repetition	7.92 (2.19)	8.00	0.00	10.00	10.00
MAIN comprehension	7.35 (1.59)	8.00	2.00	10.00	8.00
MAIN production	7.12 (1.85)	7.00	1.00	10.00	9.00
Sentence repetition	4.54 (2.45)	5.00	0.00	10.00	10.00
CLT comprehension	7.28 (1.93)	8.00	0.00	10.00	10.00
CLT production	7.14 (2.08)	8.00	0.00	10.00	10.00
Usefulness of report					
Nonword repetition	7.33 (2.40)	8.00	1.00	10.00	9.00
MAIN comprehension	7.53 (2.27)	8.00	1.00	10.00	9.00
MAIN production	7.47 (2.10)	8.00	2.00	10.00	8.00
Risk index	5.93 (2.54)	5.00	0.00	10.00	10.00

Note. Min = minimum; Max = maximum; MAIN = Multilingual Assessment Instrument for Narratives; CLT = cross-linguistic lexical task.
[a]Scales range from 1 (*negative*) to 10 (*positive*).

overall scores on attention span, motivation, and engagement. Instructions were clear, but SLPs addressed that this test was too long and too difficult for the children.

Feasibility Issues

The answers to the open-ended questions resulted in 186 different feasibility issues. Table 7 presents the issues per subtest and category. Most feasibility issues were related to technical issues, the clarity of instructions, and test construction, which correspond to the categories that were discussed in the usability study, as well as children's reactions to the tests. Technical difficulties were loss of connection and an error in the reports section. All but two participating SLPs ran into technical difficulties during test administration. Most feasibility issues concerned the SRep, which was found to be too long and too difficult for the children. Issues related to test construction concerned mostly the LITMUS-NL Additional subtests (SRep and CLT). Recurring issues were the lack of stop rules; difficulty of the test; and, relatedly, decreasing attention of the children when tests contained many items.

Adapting LITMUS-NL

LITMUS-NL was adapted based on the results of the feasibility study. Most feasibility issues related to test construction, representing 30.1% (56/186) of all issues, could not be fully solved by us, as previously mentioned. However, adaptations were made to solve 28.6% (16/56) of these issues. One of the solutions was shortening the

Table 7. Amount of different feasibility issues within categories per LITMUS-NL (Dutch version of Language Impairment Testing in Multilingual Settings) test.

Issue category and subcategory	General	LITMUS-NL Core					LITMUS-NL Additional			Total
		CL-NWR	MAIN comp.	MAIN prod.	Risk index	Report	SRep	CLT comp.	CLT prod.	
Technical										
Bugs	8	0	0	1	0	0	0	0	0	9
User experience	8	0	0	0	1	2	1	1	0	13
User design	1	0	0	2	0	0	3	0	1	7
Instructions										
General	5	1	0	3	0	1	2	1	3	16
Test manual	3	0	0	0	0	2	0	0	0	5
Online tool	5	0	0	0	0	0	0	0	0	5
Scoring										
Scoring of items	2	1	0	4	2	0	2	0	0	11
Interpretation	2	0	0	1	0	5	1	1	1	11
Test construction and materials										
Test construction	11	1	2	6	2	0	8	10	10	50
Audio recordings	0	2	0	0	0	0	2	0	0	4
Expectations	2	0	0	0	0	0	0	0	0	2
Recommendations										
General	8	0	1	2	0	1	1	1	1	15
Technical features	2	0	0	0	0	0	0	0	0	2
Website	5	0	0	0	0	0	0	0	0	5
Test performance	3	2	0	2	0	0	2	0	0	9
Reactions of children	4	5	4	1	0	0	6	1	1	22
Total	69	12	7	22	5	11	28	15	17	**186**

Note. CL-NWR = cross-linguistic nonword repetition task; MAIN = Multilingual Assessment Instrument for Narratives; comp. = comprehension; prod. = production; SRep = sentence repetition task; CLT = cross-linguistic lexical task.

SRep based on the study of van Barreveld et al. (2023). In this study, items were removed taking into account, among others, difficulty, discriminative ability, and internal reliability, resulting in a 16-item version with eight different sentence structures. Within the other categories, 88.5% (115/130) of the issues were solved. For example, we shortened the instruction manual, made instruction videos, and simplified the instructions in the online environment. We also composed one overall index score that combines the scores of the Core subtests, making the interpretation of test scores easier. Technical issues were solved by enhancing technical features in the digital testing environment, resolving bugs, and managing the expectations of users concerning technical possibilities. Finally, issues concerning the responses of the children were solved, mostly by describing different possibilities for stimulation and including a shorter version of the SRep.

Discussion

There is an urgent need for a more reliable identification of DLD in multilingual children, and research aiming to support adequate assessment of these children has increased in the past years. The development of the LITMUS tests (Armon-Lotem et al., 2015) significantly contributed toward achieving this goal, with some of the tests successfully distinguishing language delay from language disorder (e.g., Boerma et al., 2015, 2016; Chiat & Polišenská, 2016). In the current study, we strove toward the implementation in daily clinical practice of an online instrument—that is, LITMUS-NL—consisting of the Dutch versions of a number of these LITMUS tests. The steps needed for the successful implementation of such a new instrument, including the investigation of usability and feasibility in the early stages of the implementation process, are often forgotten in research but are essential for actual progress in the field. Many newly developed instruments that show promise for clinical practice are mainly used in research settings, possibly leading to research waste. Isaacs and Chalmers (2023) discuss how research waste can be avoided in the field of applied linguistics and, among others, emphasize the involvement of stakeholders. The current study did exactly that by involving SLPs, who are the end users of LITMUS-NL. We took essential steps toward implementation and investigated the usability and feasibility of LITMUS-NL according to SLPs. We identified what is deemed important to SLPs and were able to solve many

issues that would have interfered with successful implementation in daily clinical practice if these issues had not been identified through this iterative process of evaluation and adaptation. With this research, we provide a model for how the early stages of implementation of a new product can be approached in a systematic and stepwise manner.

The starting point of LITMUS-NL and the investigations of usability and feasibility was research that showed that several language-independent LITMUS tests had high clinical value (Boerma & Blom, 2017). These Core subtests were, together with two additional language-specific tests, embedded in an online test environment to make them accessible for professionals. This first version of LITMUS-NL was tested during a usability study to assess whether LITMUS-NL would be used as intended by SLPs, to understand the clinical setting in which LITMUS-NL would be used, and to find usability issues (Bowen et al., 2009; O'Cathain et al., 2019; Sauro & Lewis, 2016; WHO, 2016). Results showed a good degree of usability and a high added value of LITMUS-NL but also many usability issues.

Based on the feedback of participating SLPs, LITMUS-NL was adapted, a test manual was created, and norm scores for the Core subtests were developed together with a psychometrician. The new and improved version of LITMUS-NL was subsequently examined in a feasibility study to assess whether LITMUS-NL will be used as intended in daily clinical practice and to help identify feasibility issues (Bowen et al., 2009; Mulkey et al., 2019; WHO, 2016). Results of the feasibility study showed an acceptable degree of usability, but scores were lower than those found in the usability study. The added value was high. Despite the adaptations made after the usability study, the SLPs still addressed many feasibility issues, showing that the administration of LITMUS-NL in an uncontrolled setting (i.e., daily clinical practice) posed new and other challenges than its use in a controlled setting and confirming the importance of iteration. After the feasibility study, LITMUS-NL was further modified. Note that modifying LITMUS-NL based on the usability and feasibility studies is necessary but not sufficient for its implementation. Although several psychometric properties of the Dutch versions of the SRep, CL-NWR, and CLT have been investigated in previous research (Boerma et al., 2015; de Jong et al., 2021; Van Wonderen & Unsworth, 2021), an important next step is establishing the validity and reliability of LITMUS-NL as a test battery, building up to the actual implementation of the instrument.

Implications and Recommendations

LITMUS-NL is a unique test that supports the identification of DLD in multilingual children by measuring skills that are not dependent on exposure to the language of testing. As such, it can play an important role in the early phase of the diagnostic process, helping SLPs in deciding whether or not further assessment for DLD is necessary. This reduces misdiagnosis, improving the care for multilingual children with DLD. However, to reliably diagnose DLD, more information is necessary, which LITMUS-NL does not provide. LITMUS-NL is thus not a stand-alone instrument and should, in case DLD is suspected, be complemented with, among others, dual-language assessments, psychological evaluations, and hearing tests.

Our process of developing one online instrument of the LITMUS tests combined and taking the first steps toward implementation could provide a model for other LITMUS users. The usability issues regarding the online environment are specific to the Dutch (UU-DAB) context and, therefore, less relevant for other LITMUS users outside of the Netherlands. However, issues regarding length (e.g., SRep) are relevant for other LITMUS users as well.

The results of the current study revealed what aspects of an online instrument are deemed important for SLPs, which can be generalized beyond LITMUS-NL, particularly as there is a trend toward more online test assessments with children. First and unsurprisingly, new digital instruments should work seamlessly. In both studies, SLPs noted many technical issues, including bugs and loss of connection. This highlights the importance of multiple run-throughs, also in clinical practice settings, before the actual implementation of an online instrument. Instructions for a new digital product are also of great importance. Self-reported digital literacy varied between the participants, and instruments should be suitable for digitally literate SLPs and SLPs who are less used to working in an online environment. Therefore, the instructions on how to use the instrument need to be sufficient and clear. If a test manual or an instruction does not match the needs and preferences of the intended users, the validity of the instrument and test administration will decrease, possibly leading to poor clinical care (First et al., 2014; Mullins-Sweatt et al., 2016). In our study, the SLPs addressed the desire for instructional videos, allowing them to easily follow the steps needed to be taken and making the product accessible to a broad group of users. The use of instructional videos to facilitate user-friendly online test performances could be explored in future research.

The SLPs mentioned several issues concerning test construction, specifically concerning the length and complexity of tests. SLPs work with children on a daily basis and have valuable ideas about which pictures, sentences, or animations are suitable for children. Moreover, they know what motivates a child and how many items a child

can attentively respond to. By involving SLPs or other intended end users early in the process of test development, usability and feasibility can be enhanced, facilitating implementation (O'Cathain et al., 2019). However, it is important to involve participants who adequately represent the intended users (Gravitt, 2023). The current feasibility study was limited by an uneven distribution of SLPs working in speech and hearing centers and speech-language pathology practices. Twelve SLPs working in speech and hearing centers did apply for the feasibility study, but 10 of those dropped out before the start of the study, mainly due to insufficient time to perform LITMUS-NL together with their protocolled tests. The current study thus illustrates not only the value of involving end users in the early stages of the implementation process but also the difficulties.

Conclusions

To overcome the challenges of identifying DLD in multilingual children, the LITMUS tests were developed. Although these tests have proven to be successful in distinguishing DLD from language delay, their implementation in daily clinical practice is still limited. In this study, the first steps toward implementation of the digital Dutch version of several combined LITMUS tests—LITMUS-NL—were taken. By conducting a usability study followed by a feasibility study, we could unveil the key considerations essential to SLPs when implementing a new digital language assessment battery in daily clinical practice. Our findings emphasize the importance of involving the intended end users of a digital test battery throughout the entire process of development and implementation, evaluating and adapting the test battery in multiple cycles. Involvement enabled us to track down issues in test construction, instructions, and technical requirements and to develop a product that is usable, feasible, and of value to intended users. In addition, by working toward and reporting on implementation in a systematic and stepwise manner, we strive to inspire other researchers so that valuable products, developed in research settings, are not wasted but rather effectively utilized in daily clinical practice.

Author Contributions

Linda Wouda: Formal analysis (Lead), Methodology (Lead), Investigation (Lead), Project administration (Lead), Visualization (Lead), Writing – original draft (Equal). **Tessel Boerma:** Conceptualization (Lead), Funding acquisition (Equal), Supervision (Equal), Validation (Lead), Visualization (Equal), Writing – original draft (Equal), Writing – review & editing (Lead). **Ellen Gerrits:** Conceptualization (Supporting), Funding acquisition (Equal), Supervision (Supporting), Validation (Supporting), Writing – review & editing (Supporting). **Elma Blom:** Conceptualization (Lead), Funding acquisition (Lead), Resources (Lead), Supervision (Equal), Validation (Supporting), Writing – review & editing (Supporting).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, Elma Blom, upon reasonable request.

Acknowledgments

This study was supported by the K.F. Hein Fund, awarded to Elma Blom; the Damsté-Terpstra Fund, awarded to Linda Wouda; and a Dynamics of Youth grant from Utrecht University, awarded to Elma Blom, Tessel Boerma, and Ellen Gerrits. This study could not have been performed without the work of the members of European Cooperation in Science and Technology Action IS0804. We further express gratitude to Daan Asscheman and Eva van de Weijer-Bergsma for developing the subtests in the Utrecht University Developmental Assessment Battery, Dave Hessen for creating the norm scores for the LITMUS-NL (Dutch version of Language Impairment Testing in Multilingual Settings) Core subtests, and Esther Kroese and Anna de Graaf for their work on the online instructions and test manual.

References

Abascal, E., Díaz De Rada, V., García Lautre, I., & Landaluce, M. I. (2018). Analysis of the response structure to a set of questions with large number of scale points: A new combined metric and categorical approach. *International Journal of Social Research Methodology, 21*(4), 395–407. https://doi.org/10.1080/13645579.2017.1399620

American Speech-Language-Hearing Association. (2004). Knowledge and skills needed by speech-language pathologists and audiologists to provide culturally and linguistically appropriate services. *ASHA Supplement, 24,* 152–158.

Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.). (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Multilingual Matters. https://doi.org/10.21832/9781783093137

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction, 24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*(3), 114–123. https://

uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11*(1), 1–29. https://doi.org/10.2167/beb392.0

Bevan, N., Kirakowski, J., & Maissel, J. (1991). What is usability? In H. J. Bullinger (Ed.), *Proceedings of the 4th International Conference on Human–Computer Interaction*. Elsevier.

Boerma, T. D. (2017). *Profiles and paths: Effects of language impairment and bilingualism on children's linguistic and cognitive development* [Doctoral dissertation, Utrecht University]. Utrecht University Repository. https://dspace.library.uu.nl/handle/1874/356269

Boerma, T., & Blom, E. (2017). Assessment of bilingual children: What if testing both languages is not possible? *Journal of Communication Disorders, 66,* 65–76. https://doi.org/10.1016/j.jcomdis.2017.04.001

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research, 58*(6), 1747–1760. https://doi.org/10.1044/2015_JSLHR-L-15-0058

Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: Implications for clinical practice. *International Journal of Language & Communication Disorders, 51*(6), 626–638. https://doi.org/10.1111/1460-6984.12234

Bowen, D. J., Kreuter, M., Spring, B., Cofta-Woerpel, L., Linnan, L., Weiner, D., Bakken, S., Kaplan, C. P., Squiers, L., Fabrizio, C., & Fernandez, M. (2009). How we design feasibility studies. *American Journal of Preventive Medicine, 36*(5), 452–457. https://doi.org/10.1016/j.amepre.2009.02.002

Broekhuis, M., van Velsen, L., & Hermens, H. (2019). Assessing usability of eHealth technology: A comparison of usability benchmarking instruments. *International Journal of Medical Informatics, 128,* 24–31. https://doi.org/10.1016/j.ijmedinf.2019.05.001

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 207–212). CRC Press. https://doi.org/10.1201/9781498710411-35

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 125–150). Multilingual Matters. https://doi.org/10.21832/9781783093137-008

Chiat, S., & Polišenská, K. (2016). A framework for crosslinguistic nonword repetition tests: Effects of bilingualism and socioeconomic status on children's performance. *Journal of Speech, Language, and Hearing Research, 59*(5), 1179–1189. https://doi.org/10.1044/2016_JSLHR-L-15-0293

de Jong, J., Blom, E., & van Dijk, C. (2021). LITMUS SRep – een zinsherhaaltaak voor het Nederlands [LITMUS SRep—A sentence repetition task for Dutch]. *Stem-, Spraak- en Taalpathologie, 26,* 96–116. https://doi.org/10.21827/32.8310/2021-96

Dietrich, S., & Hernandez, E. (2022). *Language use in the United States: 2019*. U.S. Census Bureau. https://www.census.gov/library/publications/2022/acs/acs-50.html

Ehrler, F., Weinhold, T., Joe, J., Lovis, C., & Blondon, K. (2018). A mobile app (BEDSide Mobility) to support nurses' tasks at the patient's bedside: Usability study. *JMIR mHealth and uHealth, 6*(3), Article e57. https://doi.org/10.2196/mhealth.9079

First, M., Bhat, V., Adler, D., Dixon, L., Goldman, B., Koh, S., Levine, B., Oslin, D., & Siris, S. (2014). How do clinicians actually use the *Diagnostic and Statistical Manual of Mental Disorders* in clinical practice and why we need to know more. *The Journal of Nervous and Mental Disease, 202*(12), 841–844. https://doi.org/10.1097/NMD.0000000000000210

Fonda, S. J., Paulsen, C. A., Perkins, J., Kedziora, R. J., Rodbard, D., & Bursell, S.-E. (2008). Usability test of an internet-based informatics tool for diabetes care providers: The Comprehensive Diabetes Management Program. *Diabetes Technology & Therapeutics, 10*(1), 16–24. https://doi.org/10.1089/dia.2007.0252

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2019). MAIN: Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics, 56,* Article 155. https://doi.org/10.21248/zaspil.56.2019.414

Gerber, S. M., Schütz, N., Uslu, A. S., Schmidt, N., Röthlisberger, C., Wyss, P., Perny, S., Wyss, C., Koenig-Bruhin, M., Urwyler, P., Nyffeler, T., Marchal-Crespo, L., Mosimann, U. P., Müri, R. M., & Nef, T. (2019). Therapist-guided tablet-based telerehabilitation for patients with aphasia: Proof-of-concept and usability study. *JMIR Rehabilitation and Assistive Technologies, 6*(1), Article e13163. https://doi.org/10.2196/13163

Gillham, B. (2008). *Developing a questionnaire* (2nd ed.). Bloomsbury Publishing.

Gravitt, P. E. (2023). How to turn evidence into policy in resource-limited settings. *Nature Medicine, 29*(9), Article 2166. https://doi.org/10.1038/s41591-023-02349-w

Greenwell, T., & Walsh, B. (2021). Evidence-based practice in speech-language pathology: Where are we now? *American Journal of Speech-Language Pathology, 30*(1), 186–198. https://doi.org/10.1044/2020_AJSLP-20-00194

Grimm, A., & Schulz, P. (2014). Specific language impairment and early second language acquisition: The risk of over- and underdiagnosis. *Child Indicators Research, 7*(4), 821–841. https://doi.org/10.1007/s12187-013-9230-6

Grosjean, F., & Pavlenko, A. (2021). *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press. https://doi.org/10.1017/9781108975490

Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication, 2,* Article 16. https://doi.org/10.3389/fcomm.2017.00016

IBM Corporation. (2020). *IBM SPSS Statistics for Windows* (Version 27.0).

Isaacs, T., & Chalmers, H. (2023). Reducing 'avoidable research waste' in applied linguistics research: Insights from healthcare research. *Language Teaching*. Advance online publication. https://doi.org/10.1017/S0261444823000411

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders, 43*(6), 456–473. https://doi.org/10.1016/j.jcomdis.2010.02.002

Kulkarni, A. A., Chadd, K. E., Lambert, S. B., Earl, G., Longhurst, L. M., McKean, C., Hulme, C., McGregor, K. K., Cunniff, A., Pagnamenta, E., Joffe, V., Ebbels, S. E., Bangera, S., Wallinger, J., & Norbury, C. F. (2022). Editorial perspective: Speaking up for developmental language disorder—The top 10 priorities for research. *The Journal of Child Psychology*

and Psychiatry, 63(8), 957–960. https://doi.org/10.1111/jcpp.13592

Léglise, I. (2017). Multilinguisme et hétérogénéité des pratiques langagières: Nouveaux chantiers et enjeux du Global South [Multilingualism and heterogeneous language practices: New research areas and issues in the Global South]. *Langage et Société, 160–161*(2), 251–266. https://doi.org/10.3917/ls.160.0251

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., Al-Shahi Salman, R., Chan, A., & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet, 383*(9912), 101–104. https://doi.org/10.1016/S0140-6736(13)62329-6

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–122). Multilingual Matters. https://doi.org/10.21832/9781783093137-007

Microsoft Corporation. (2020). *Microsoft Excel*. https://office.microsoft.com/excel

Mulkey, M. A., Hardin, S. R., & Schoemann, A. M. (2019). Conducting a device feasibility study. *Clinical Nursing Research, 28*(3), 255–262. https://doi.org/10.1177/1054773818803171

Mullins-Sweatt, S. N., Lengel, G. J., & DeShong, H. L. (2016). The importance of considering clinical utility in the construction of a diagnostic manual. *Annual Review of Clinical Psychology, 12*(1), 133–155. https://doi.org/10.1146/annurev-clinpsy-021815-092954

O'Cathain, A., Croot, L., Duncan, E., Rousseau, N., Sworn, K., Turner, K. M., Yardley, L., & Hoddinott, P. (2019). Guidance on how to develop complex interventions to improve health and healthcare. *BMJ Open, 9*(8), Article e029954. https://doi.org/10.1136/bmjopen-2019-029954

Orgassa, A., & Weerman, F. (2008). Dutch gender in specific language impairment and second language acquisition. *Second Language Research, 24*(3), 333–364. https://doi.org/10.1177/0267658308090184

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics, 31*(2), 227–252. https://doi.org/10.1017/S0142716409990373

Pearson, B. Z. (2013). Distinguishing the bilingual as a late talker from the later talker who is bilingual. In L. Rescorla & P. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 67–87). Brookes. https://brookespublishing.com/wp-content/uploads/2021/06/intervention-approaches-for-young-late-talkers.pdf [PDF]

Peña, E. D., Bedore, L. M., Lugo-Neris, M. J., & Albudoor, N. (2020). Identifying developmental language disorder in school age bilinguals: Semantics, grammar, and narratives. *Language Assessment Quarterly, 17*(5), 541–558. https://doi.org/10.1080/15434303.2020.1827258

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Elsevier Science. https://doi.org/10.1016/c2010-0-65192-3

Schmeets, H., & Cornips, L. (2021). *Talen en dialecten in Nederland: Wat spreken we thuis en wat schrijven we op sociale media?* [Languages and dialects in the Netherlands: What do we speak at home and what do we write on social media?]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/talen-en-dialecten-in-nederland?onepage=true

Simonsen, H. G., & Haman, E. (2017). LITMUS-CLT: A new way to assess bilingual lexicons. *Clinical Linguistics & Phonetics, 31*(11–12), 811–817. https://doi.org/10.1080/02699206.2017.1307454

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., Boyd, K. A., Craig, N., French, D. P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. *BMJ Online, 374,* Article n2061. https://doi.org/10.1136/bmj.n2061

Stefano, F., Borsci, S., & Stamerra, G. (2010). Web usability evaluation with screen reader users: Implementation of the partial concurrent thinking aloud technique. *Cognitive Processing, 11*(3), 263–272. https://doi.org/10.1007/s10339-009-0347-y

Tuller, L. (2018). Clinical use of parental questionnaires in multilingual contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 301–330). Multilingual Matters. https://doi.org/10.21832/9781783093137-013

van Barreveld, M., Schaeffer, J., & Scheper, A. (2023). LITMUS-SR-NL-16: A short sentence repetition task to identify children with DLD. In P. Gappmayr & J. Kellogg (Eds.), *Proceedings of the 47th Annual Boston University Conference on Language Development* (pp. 29–42). Cascadilla Press. http://www.lingref.com/bucld/47/BUCLD47-03.pdf [PDF]

van Gemert-Pijnen, J. E. W. C. (2022). Implementation of health technology: Directions for research and practice. *Frontiers in Digital Health, 4,* Article 1030194. https://doi.org/10.3389/fdgth.2022.1030194

van Velsen, L., van der Geest, T., & Klaassen, R. (2011). Identifying usability issues for personalization during formative evaluations: A comparison of three methods. *International Journal of Human–Computer Interaction, 27*(7), 670–698. https://doi.org/10.1080/10447318.2011.555304

Van Wonderen, E., Blom, E., Boerma, T., Janssen, B., Unsworth, S., & Van Dijk, C. (2017). *Cross-linguistic lexical task – Dutch*. http://psychologia.pl/clts/

Van Wonderen, E., & Unsworth, S. (2021). Testing the validity of the cross-linguistic lexical task as a measure of language proficiency in bilingual children. *Journal of Child Language, 48*(6), 1101–1125. https://doi.org/10.1017/S030500092000063X

World Health Organization. (2016). *Monitoring and evaluating digital health interventions: A practical guide to conducting research and assessment*. https://iris.who.int/bitstream/handle/10665/252183/9789241511766-eng.pdf?sequence=1 [PDF]

Appendix A

Results Per Item of Standardized Scales for the Usability Study

System Usability Scale[a]	*M* (*SD*)	*Mdn*	Min	Max	Range
SUS1: I think I would like to use LITMUS-NL frequently.	3.21 (0.51)	3.00	2	4	2
SUS2: I found LITMUS-NL unnecessarily complex.	3.12 (0.80)	3.00	1	4	3
SUS3: I thought LITMUS-NL was easy to use.	2.67 (0.70)	3.00	1	4	3
SUS4: I think I would need technical support to use LITMUS-NL.	3.38 (0.65)	3.00	2	4	2
SUS5: I found that the various functions were well integrated.	2.75 (0.61)	3.00	1	4	3
SUS6: I thought that there was too much inconsistency in LITMUS-NL.	3.00 (0.66)	3.00	1	4	3
SUS7: I would imagine that most people would learn to use LITMUS-NL very quickly.	2.92 (0.72)	3.00	1	4	3
SUS8: I found LITMUS-NL cumbersome to use.	2.92 (0.58)	3.00	2	4	2
SUS9: I felt very confident using LITMUS-NL.	2.67 (0.82)	3.00	1	4	3
SUS10: I needed to learn a lot of things before I could get going with LITMUS-NL.	2.33 (0.82)	2.00	1	4	3
Value/Usefulness Subscale of the Intrinsic Motivation Inventory					
IMI1: I believe LITMUS-NL could be of some value for me.	5.83 (1.13)	6.00	3	7	4
IMI2: I think that doing LITMUS-NL is useful for identifying DLD in multilingual children.	5.50 (1.18)	5.00	3	7	4
IMI3: I think LITMUS-NL is important because it can determine language abilities of multilingual children.	5.17 (1.24)	5.00	1	7	6
IMI4: I would be willing to use LITMUS-NL again because it has value to me.	5.57 (1.11)	6.00	3	7	4
IMI5: I think LITMUS-NL would help me to determine language abilities of multilingual children.	5.46 (0.98)	6.00	4	7	3
IMI6: I believe LITMUS-NL could be beneficial to me.	5.79 (1.02)	6.00	4	7	3
IMI7: I think LITMUS-NL is important.	5.67 (1.09)	6.00	3	7	4

Note. Min = minimum; Max = maximum; SUS = System Usability Scale; LITMUS-NL = Dutch version of Language Impairment Testing in Multilingual Settings; IMI = Intrinsic Motivation Inventory; DLD = developmental language disorder.

[a]Odd items were positively stated questions; even items were stated negatively. The scores are calculated in a manner that the higher the score, the better the usability.

Appendix B

Results Per Item of Standardized Scales for the Feasibility Study

System Usability Scale[a]	*M* (*SD*)	*Mdn*	Min	Max	Range
SUS1: I think I would like to use LITMUS-NL frequently.	2.92 (0.91)	3.00	1	4	3
SUS2: I found LITMUS-NL unnecessarily complex.	2.32 (1.18)	3.00	0	4	4
SUS3: I thought LITMUS-NL was easy to use.	2.48 (1.05)	3.00	0	4	4
SUS4: I think I would need technical support to use LITMUS-NL.	3.08 (1.19)	3.00	0	4	4
SUS5: I found that the various functions were well integrated.	2.72 (0.94)	3.00	1	4	3
SUS6: I thought that there was too much inconsistency in LITMUS-NL.	2.80 (0.87)	3.00	1	4	3
SUS7: I would imagine that most people would learn to use LITMUS-NL very quickly.	2.72 (0.84)	3.00	0	4	4
SUS8: I found LITMUS-NL cumbersome to use.	2.24 (1.09)	3.00	0	4	4
SUS9: I felt very confident using LITMUS-NL.	2.60 (0.91)	3.00	1	4	3
SUS10: I needed to learn a lot of things before I could get going with LITMUS-NL.	2.24 (1.05)	3.00	0	3	3
Value/Usefulness Subscale of the Intrinsic Motivation Inventory					
IMI1: I believe LITMUS-NL could be of some value for me.	5.68 (1.34)	6.00	2	7	5
IMI2: I think that doing LITMUS-NL is useful for identifying DLD in multilingual children.	5.16 (1.80)	6.00	2	7	5
IMI3: I think LITMUS-NL is important because it can determine language abilities of multilingual children.	5.60 (1.50)	6.00	2	7	5
IMI4: I would be willing to use LITMUS-NL again because it has value to me.	5.36 (1.63)	6.00	2	7	5
IMI5: I think LITMUS-NL would help me to determine language abilities of multilingual children.	5.48 (1.66)	6.00	2	7	5
IMI6: I believe LITMUS-NL could be beneficial to me.	5.32 (1.63)	6.00	2	7	5
IMI7: I think LITMUS-NL is important.	5.48 (1.48)	6.00	2	7	5

Note. Min = minimum; Max = maximum; SUS = System Usability Scale; LITMUS-NL = Dutch version of Language Impairment Testing in Multilingual Settings; IMI = Intrinsic Motivation Inventory; DLD = developmental language disorder.

[a]Odd items were positively stated questions; even items were stated negatively. The scores are calculated in a manner that the higher the score, the better the usability.