Usability of the Language Test for Multilingual Children (LITMUS_NL)

Name:	L. (Linda) Wouda
Student number:	6356982
Status:	Master Thesis
Date:	25-06-2021
Studies:	Utrecht University Master Clinical Health Sciences Master's program Health Sciences for Healthcare Professionals UMC Utrecht
Teacher:	dr. Rob Zwitserlood
Supervisor:	Prof. dr. Ellen Gerrits
Institution:	HU University of Applied Sciences Utrecht Research group: Speech and Language Therapy – Participation is communication
Project Team:	Prof. dr. Elma Blom, dr. Tessel Boerma and Anna de Graaf Utrecht University – Faculty of Social Sciences Development & Education of Youth in Diverse Societies
Protocol:	STROBE Statement
Intended journal:	International Journal of of Language and Communication Disorders
Wordcount:	3791/3800
Wordcount Abstract:	299/300
Wordcount Abstract (NL):	280/300

Abstract

Background: There is a lack of suitable language tests for multilingual children to distinguish a developmental language disorder from a language delay, caused by insufficient input in the second language. Therefore, the Language Impairment Testing in Multilingual Settings (LITMUS) test has been developed and adapted into an online version. To implement the Dutch version, LITMUS_NL, the test must be user-friendly and valuable for speech and language therapists. Therefore, the usability needs to be investigated.

Aim: To investigate the degree of usability, usability issues, added value, clinical first impression and duration of the digital version of LITMUS_NL for identifying developmental language disorders in multilingual children.

Method: This study was a descriptive usability study. The sample consisted of speech and language therapists working with multilingual children in speech and language therapy practices or speech hearing centres. The degree of usability was measured with the system usability scale. Secondary parameters were usability issues, measured with a think-aloud protocol, added value, clinical first impression and time to complete LITMUS_NL, measured with an online-questionnaire.

Results: The 24 participants assessed the degree of usability as good, with a mean score of 72/100. 337 usability issues were found in the categories of ‘Instructions’, ‘Technical’, ‘Pronunciation of items’, ‘Spelling errors’, and ‘Development of tests’. The added value was rated 5.60/7. The mean duration was 45 minutes and all domains of clinical first impression were scored 7/10 or higher.

Conclusion: This evaluation of usability of LITMUS_NL found a good degree of usability and a positive clinical first impression. A range of usability issues was identified; technical issues that can be solved, but also issues related to the construct of the test which are difficult to adapt.

Recommendations: We recommend adapting LITMUS_NL to the feedback of the users and to investigate the usability of LITMUS_NL in daily clinical practice.

Keywords:

Multilingualism; Developmental Language Disorder; Language Assessment; Usability

Samenvatting

Achtergrond: Er is een gebrek aan geschikte taaltesten voor meertalige kinderen om onderscheid te maken tussen een taalontwikkelingsstoornis en een taalachterstand door onvoldoende taalaanbod in de tweede taal. Zodoende is de *Language Impairment Testing in Multilingual Settings* (LITMUS) testbatterij ontwikkeld en aangepast als een onlineversie. De Nederlandse versie, LITMUS_NL, moet gebruiksvriendelijk en waardevol zijn voor logopedisten om de test te implementeren. Hiervoor moet de bruikbaarheid onderzocht worden.

Doel: Het onderzoeken van de mate van bruikbaarheid, bruikbaarheidsproblemen, toegevoegde waarde, klinische eerste indruk en tijdsduur van de digitale versie van de LITMUS_NL-test om taalontwikkelingsstoornissen bij meertalige kinderen te identificeren.

Methode: Dit onderzoek was een beschrijvende bruikbaarheidsstudie. De steekproef bestond uit logopedisten werkzaam met meertalige kinderen in logopediepraktijken of in audiologische centra. De mate van bruikbaarheid werd gemeten met de *system usability scale*. Secondaire uitkomstmaten waren bruikbaarheidsproblemen, gemeten met hardop denken, toegevoegde waarde, klinische eerste indruk en tijdsduur van LITMUS_NL, gemeten met een online vragenlijst.

Resultaten: De mate van bruikbaarheid was als goed beoordeeld, met een score van 72.40/100. Er zijn 337 bruikbaarheidsproblemen gevonden, in de categorieën 'Instructies', 'Technisch', 'Spelfouten', 'Uitspraak van de items' en 'Testontwikkeling'. De toegevoegde waarde werd gescoord met 5.60/7, de gemiddelde tijdsduur was 45 minuten en alle domeinen binnen klinische eerste indruk scoorden 7/10 of hoger.

Conclusie: Deze evaluatie van de bruikbaarheid van LITMUS_NL vond een goede mate van bruikbaarheid en een positieve klinische eerste indruk. Er zijn vele bruikbaarheidsproblemen gevonden: technische problemen die op te lossen zijn, maar ook problemen betreffende het construct van de test, die moeilijk zijn aan te passen.

Aanbevelingen: We bevelen aan LITMUS_NL aan te passen aan de opmerkingen van de gebruikers en om de bruikbaarheid van LITMUS_NL in de dagelijkse praktijk te onderzoeken.

Trefwoorden:

Meertaligheid; Taalontwikkelingsstoornis; Taaltesten; Bruikbaarheid

Introduction

Developmental language disorder (DLD) is a neurobiological developmental disorder characterized by persistent problems in language comprehension and/or production[1,2]. DLD occurs in 5% to 12% of children at the age of 0-7 years old[3]. Language problems can have negative effects on social interactions, school performance and behaviour[2-4]. Early diagnosis of DLD is important, to start speech and language therapy (SLT) at an early phase, preventing participation problems[2,3,5]. To diagnose DLD, intelligence and hearing difficulties must be ruled out, and the language problems are confirmed by language tests[2,3].

DLD is difficult to diagnose in multilingual children[2]. The language tests that are used to diagnose DLD only assess one language, often the majority language of a country, which is Dutch in the Netherlands. Language problems in multilingual children are usually caused by DLD or language delay. This language delay means that the problems are caused by insufficient input in Dutch since it is their second language[2]. It is important to differentiate between multilingual children with DLD and language delay to offer appropriate care. However, this differentiation is not possible with the current language tests.

Multilingual children with DLD have language difficulties in all languages they speak. Therefore, all their languages must be assessed to diagnose DLD[3,5]. In the Netherlands, this is carried out by either trained translators consulted by speech hearing centres or multilingual speech and language therapists (SLTs)[3]. However, the results of these home-language assessments are difficult to interpret, since validated reference scores are lacking for multilingual children[6-8].

Defining reference scores for multilingual children is difficult due to heterogeneity within the population[8]. The multilingual population is heterogeneous in multiple aspects, for instance; variation in first languages, duration, and context of second language acquisition[9-11]. Armon-Lotem et al.[9] suggested applying a larger margin of error on cut-off scores for multilingual children. However, diagnoses remain invalid since the reference scores are based on results of monolingual children[3].

Orellana et al.[12] suggested conducting dynamic assessments for diagnosing DLD in multilingual children. Dynamic assessments determine a child's learning potential by performing test-teach-retest for different language domains[12]. Suggestive evidence regarding diagnostic accuracy was found. However, the results are based on studies with small, heterogeneous samples, devaluing their reliability[12].

To fulfil the lack of standardized language tests for multilingual children, the Language Impairment Testing in Multilingual Settings (LITMUS) test was developed by participants of

COST-ACTION IS0804[1][9]. LITMUS contains subtests for nonword repetition, narratives, sentence repetition and expressive and receptive vocabulary[9]. These subtests assess universal language skills that are not specific to one language and therefore, applicable for multilingual children[9]. Although the subtests were found to be valuable and suitable to identify DLD in multilingual children, the validity has not been assessed[7,11,13]. However, the combination of the Multilingual Assessment Instrument for Narratives (MAIN), the Quasi-Universal nonword repetition task and a parental questionnaire was found to be highly sensitive (97%) and specific (97%) for classifying DLD in multilingual children[14]. These results indicate the added value of the combined subtests of LITMUS for identifying DLD in multilingual children.

The Dutch version of the test, LITMUS_NL, has been developed as a digital test. This digitalisation will enhance the efficiency of language assessments. To implement LITMUS_NL in daily clinical practice, the test needs to work optimally for SLTs. By improving and then implementing LITMUS_NL, SLTs will use the test to better differentiate between DLD and language delay. This results in better identification and thus, better care for multilingual children with DLD. LITMUS_NL should be user-friendly, time-efficient, and valuable to the intended users[15]. SLTs need to evaluate LITMUS_NL to collect feedback for further improvement[15].

Aim

We aim to investigate the degree of usability, usability issues, added value, clinical first impression and duration of the digital version of the LITMUS_NL test for identifying DLD in multilingual children.

Methods

The degree of usability, usability issues, added value, clinical first impression, and duration of LITMUS_NL were studied in a descriptive usability study. A usability study assesses whether a program is used as intended and evaluates the domains of acceptability, user-satisfactory, intention to use a product, perceived appropriateness, and expected applicability in daily practice[16].

[1] COST-ACTION IS0804 is a European project coordinating research on linguistic and cognitive abilities of multilingual children with DLD across different migration communities. For more information see: https://www.bi-sli.org and Armon-Lotem et al.[9]

Sample

The study population consisted of SLTs working with multilingual children in SLT practices or speech hearing centres. SLTs that performed language assessments on multilingual children between the age of 3-9 years old were eligible to participate.

The study population was homogeneous in terms of gender and work setting. Based on similar usability studies, the sample size was determined at 25 SLTs[16-20].

SLTs participated voluntarily and were recruited by promoting the study on social media in February 2021. Participants received written information and signed a consent form.

This study was conducted following the principles of the Declaration of Helsinki (2013) and is not subject to the Medical Research Involving Human Subjects Act (WMO), as stated by the Ethics Committee of the Faculty of Social and Behavioural Sciences of Utrecht University (reference number: 20-546).

LITMUS_NL test

The LITMUS_NL test assesses language skills in multilingual children[9]. The instrument contains four subtests: nonword repetition (NWRT), sentence repetition (SR), narratives, and vocabulary[9]. NWRT assess phonological short-term memory[13]. The child is asked to repeat nonwords, read aloud by the program. The subtest contains visual rewards of an alien moving on the screen, as presented in Figure 1. SR is measuring the imitation of sentences. The child is asked to repeat sentences read aloud by the program and receives visual rewards of a bear moving on the screen after finishing a sentence. The narrative subtest assesses story comprehension and elicit storytelling skills with two picture sequence narratives from the MAIN[11,21]. The SLT reads a story aloud, while showing corresponding pictures, and asks questions afterwards. In a second picture sequence narrative, the child is asked to tell the story by him or herself. The vocabulary tests are testing expressive and receptive vocabulary of verbs and nouns. In the two expressive tests, the child has to name pictures. In the two receptive subtests, the child is asked to choose one picture out of four after the SLT reads the word in a short question.

LITMUS_NL was recently developed as an online program by the IT services of Utrecht University. Answers and scores can be entered in the online score form. The online test has not been used yet and is evaluated for the first time.

[insert Figure 1]

Data collection and analysis

Data regarding degree of usability, usability issues, perceived value, clinical first impression and duration of LITMUS_NL were collected. Data were anonymised after collection and then imported into SPSS 25 and Excel. If data were missing, available case analyses were conducted. Data were checked for normality using the Kolmogorov-Smirnov test. When distributed normally, p>.05, means and standard deviations were calculated[22].

The degree of usability was measured by filling out the System Usability Scale (SUS)[23]. SUS is a standardized 10-item scale with a 5-point Likert scale from 1(strongly disagree) to 5(strongly agree). The items address complexity, integration, consistency, convenience, confidence, need for support, and willingness to use the product[24]. The score contributions of the questions were calculated using the strategy of Brooke[23]. The score contribution for positively stated items (odd questions) was scale position minus 1 and for negative items (even questions) 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[16,24]. The mean SUS score, including standard deviation, was calculated, and compared to the adjective rating of Bangor et al.[24], stating a score >70 is acceptable[24,25]. For each question, median scores were calculated.

SUS is most frequently used for investigating usability since it is a short scale with an unambiguous degree of usability and is found to be an effective and reliable instrument[18,24,26]. However, SUS is inadequate as a stand-alone tool, since it provides no information about why the score is achieved[18,24,26]. Therefore, secondary parameters were added to explain the degree of usability.

Usability issues were collected by performing a synchronised concurrent think-aloud protocol (TAP)[27]. TAP was used to identify unsatisfactory features by expressing the users' thoughts while working with LITMUS_NL[28]. TAP collects information for further improvement of the program. While performing the tests of LITMUS_NL, the SLTs verbalized and audio-recorded their thoughts. These audio recordings were transcribed verbatim. Reading out loud of the instructions or items was excluded from the transcriptions. The transcriptions were coded to collect initial usability errors[16]. The errors of all transcriptions were compared, and similar errors were merged as usability issues[29]. All issues and their frequencies were categorized, and individually prioritized by the principal researcher, working as an SLT, and a linguist, highly experienced in multilingualism. Priority was scored on a 5-point scale from 1(not at all urgent) to 5(very urgent) to identify the major and minor usability issues of LITMUS_NL[29]. When differing more than one scale position on an issue, the two researchers discussed until consensus was reached.

The added value of LITMUS_NL was studied using the subscale Value/Usefulness of the Intrinsic Motivation Inventory (IMI)[30]. This subscale consists of seven questions that are scored on a 7-point Likert scale from 'not at all true' to 'very true'[30]. The mean score of the subscale was calculated and was regarded positively if the score was >4[19,30]. The median score was calculated for each question.

Clinical first impression concerning intuitiveness of buttons, complexity, applicability, feasibility, design, and insight into language skills were collected with visual-analogue scales (VAS) from 1(negative) to 10(positive). Opinions about the design and additional feedback were asked in open-ended questions. The VAS scores were calculated to median scores and answers to the open-ended questions were used to explain these scores. A score >5.5 was regarded positively. All questions regarding clinical first impression are included in appendix 1.

The estimated time to complete LITMUS_NL, in minutes, was collected through an open-ended question in the questionnaire. An additional scale question was included regarding the acceptability of this duration. The mean time was calculated and led to an expected time to complete LITMUS_NL. Since SLTs cope with a high workload, time is an important factor for applicability in daily practice[31].

An online questionnaire was composed, containing SUS, the subscale value/usefulness, and the questions regarding clinical first impression. Questionnaires are the most preferred method for studying the usability of a product[20].

Demographic data regarding gender, age, work setting, work experience and self-reported digital literacy were collected in the questionnaire to describe the participants. Self-reported digital literacy was assessed with a VAS from 1(low) to 10(high), together with self-reported use of other digital assessment instruments. LITMUS_NL should be usable for SLTs with a range of digital literacy, as well as SLTs in all work settings with variating work experience.

Study procedure

The instructions to start LITMUS_NL and the usability questionnaire were piloted and evaluated by an independent group of five SLTs. They gave individual, written feedback on the clarity of questions and instructions. Their feedback was then used to adjust the instructions and the questionnaire.

Data collection and analysis was conducted from February 2021 until May 2021. The participating SLTs received a written protocol in which the goal and set-up of the study were explained, together with instructions to start the online program of LITMUS_NL and the audio recording. The SLTs performed the tests of LITMUS_NL on their own computer, tablet, or smartphone at their workplace or home. The SLTs were asked to verbalize and audio-record their thoughts on a phone, computer, or audio recorder during the entire test.

After starting the audio recording, the SLTs followed the instructions to open LITMUS_NL together with the response form on another device or tab. The SLTs completed all subtests while following the TAP. They could contact the main researcher if they encountered difficulties in starting LITMUS_NL.

Since LITMUS_NL will eventually be performed with children in daily practice, the SLTs had to take response time into account. They waited a few seconds for each item before continuing to the next. When all tests were completed, the audio recording was ended and sent securely to the main researcher, using Surffilesender. Finally, the SLTs completed the online questionnaire, by clicking a hyperlink in the instructions.

Results

Sample

A total of 25 female SLTs participated in this study. One participant was excluded, since she did not follow the instructions properly, making her results invalid. Demographic data of the participants are displayed in Table 1. As seen in this table, the SLTs were working in both SLT practices and speech hearing centres with a wide range of work experience. Digital literacy is reported as high, ranging from 7.0 to 10.0, with a median of 8.0/10.0.

Data for NWRT was missing from one participant and two participants forgot to audio-record the vocabulary test. All audio recordings were included in the analysis.

[Insert Table 1]

Degree of Usability

The mean degree of usability, measured with SUS, was 72.40/100 (±8.125). When comparing this score to the adjective scale of Bangor et al., this degree of usability is rated 'good'[24]. The frequencies of the scores per item are displayed in Figure 2. As presented in this figure, positive-stated items were mostly scored 'agree' and negative-stated items varied between 'strongly disagree' and 'neutral'. SUS10 was scored lowest, with a median score contribution of 2/4, meaning the SLTs needed more information before using LITMUS_NL as intended. The other items were rated with a median of 3/4.

[Insert Figure 2]

Usability Issues

A total of 337 usability issues were found and divided into five categories, some containing multiple subcategories, as presented in Table 2. Next, all items were prioritised using a 5-point scale. 154 items differed more than one scale position and were discussed until consensus was reached.

Usability issues mentioning clarity of instructions, stop rules and test protocol were merged in the category 'Instructions', with a sub-category for issues concerning scoring protocols.

The most frequent technical usability issues concerned user design, user experience and bugs. User design issues were mainly related to the placement and size of buttons, pictures, and text. Common user experience issues were inconveniences such as much scrolling and unclear buttons. Bugs consisted of issues regarding not being able to complete the subtests as intended, due to a not-responding second screen, not-functioning audio, or inactive buttons. As presented in Table 3, these items were judged as very urgent, since the tests could not be completed properly. 7/24 SLTs had technical difficulties and could not start the tests without help, which was also evaluated as very urgent. The most frequent usability issue, mentioned by 15/24 SLTs, was a 'Notifications' issue for the expressive vocabulary subtest regarding the notification of 'not all items are filled out'. This notification was shown when all items were scored without filling out the optional answering fields.

A total of 9 spelling errors were found. Some of the most frequent issues were spelling errors in the expressive vocabulary subtest. However, these issues were evaluated as not at all urgent, although they need to be corrected.

The items of NWRT and SR were read aloud by the program. Therefore, issues regarding pronunciation, such as improbable melodies in sentences and unclear phonemes in nonwords, were categorised for these specific subtests.

7/24 SLTs finished their TAP with a positive, final judgement, meaning they were positive about LITMUS_NL despite the usability errors.

Besides issues concerning the technical design of LITMUS_NL, the participants mentioned a range of issues regarding the development of the original tests. Some issues were general, such as the length or difficulty of a subtest, other issues were related to the selection of a specific item or picture. For instance, ambiguous pictures or inappropriate words in the vocabulary test and illogical sentences in SR.

[Insert Table 2]

[Insert Table 3]

Added Value

The mean perceived value was scored at 5.60/7 (±.897). Since this score is >4, the result is positive, meaning SLTs are positive about the added value of LITMUS_NL in daily practice[19,30]. The frequencies of the scale positions per item are displayed in Figure 3. IMI3, regarding the importance to determine language abilities, received one score of 'not at all true'. This participant commented that LITMUS_NL cannot determine general language abilities, only Dutch language abilities.

[Insert Figure 3]

Duration

The mean time of the SLTs to complete the test was 45.0 (±11.83) minutes. The acceptability of this duration in daily practice was evaluated positively, with a median score of 8/10 (±1.74).

Clinical First Impression

All median VAS scores and their ranges of the questions regarding clinical first impression are presented in Table 4.

The intuitiveness of buttons and score sheets varied between the subtests of LITMUS_NL. The NWRT was evaluated the most positive. The buttons of the vocabulary subtests and the score sheet of the MAIN were rated least intuitive, resulting in the lowest scores. However, all median scores were >7/10, meaning an overall positive score. Complexity was rated positively with a score of 8/10 (±1.30).

The applicability in daily practice with children in the age of 3-9 years old was rated positively**.** However, SLTs found LITMUS_NL difficult for 3-year-old children and addressed the need for stop rules.

Not all subtests were developed online completely. The scoring and interpretation of the storytelling subtest were not digital. Some SLTs disliked scoring manually, they stated it takes more time. However, on average it was evaluated positively with a median of 8/10.

The design was evaluated positively, however, SLTs commented that some pictures seem outdated. Finally, the relevance of LITMUS_NL and insight in language abilities of children was evaluated positively, with a median of 7.2/10 (±1.13).

[Insert Table 4]

Discussion

LITMUS_NL is a result of a European project regarding language testing in multilingual settings. However, none of the subtests has been studied for usability. We investigated the degree of usability, usability issues, added value, clinical first impression and duration of the digital version of the LITMUS_NL test for identifying DLD in multilingual children. The degree of usability was rated as good. However, the usability issues show a need for improvement, mainly regarding technical issues and instructions. The added value and the domains of clinical first impression of LITMUS_NL were evaluated positively. The mean duration was 45 minutes and was evaluated as acceptable.

SUS was scored above the cut-off point of 70/100, making LITMUS_NL acceptable as a digital product[25]. However, the SLTs found usability issues, indicating improvement is necessary. The combination of an acceptable degree of usability together with usability issues was also found in the study of Ehrler et al.[16]. They state resolving these issues will help using the product more smoothly[16].

Many usability issues regarding the clarity or lack of instructions were mentioned, some categorised as very urgent. These results are in line with the lowest scored item of SUS, "I needed to learn a lot of things before I could get going with LITMUS_NL"[23], indicating the need for improvement of the instructions. The instructions will be elaborated, and a manual is being developed.

Spelling errors were identified as minor usability issues since they are easily adaptable and do not influence the language assessment. These issues are probably caused by the conversion into the digital version and will be rectified.

Not all usability issues are easily resolved. The issues on the development of the original test, especially regarding the selection of items, cannot be solved by changing online technology. Adapting the content of subtests can only be conducted by the developers of each subtest and will change the construct of the test. Unfortunately, these issues do have a negative impact on the implementation of LITMUS_NL. In future research, we recommend also incorporating the needs of SLTs in the original constructs of tests.

We used multiple methods to investigate the usability of LITMUS_NL, as recommended in several studies[28,32]. The qualitative results of TAP were used to explain and give meaning to the quantitative scores. SUS was scored good, and IMI showed the value of LITMUS_NL. However, TAP showed that improvements are necessary. The combination of these methods uncovered all potential usability problems and opinions of the users, to anticipate in implementation[28].

This study was performed by a multidisciplinary research team consisting of researchers, a researcher/SLT, linguists and SLTs. The prioritising of usability issues was performed by an expert group of a researcher/SLT, and a linguist experienced in multilingualism. Together they could give funded opinions, based on their experience, on the usability issues, mentioned by the SLTs.

In the phase of prioritising, some usability issues appeared to be unclear and ambiguous for the second researcher. Secondly, the scale of urgency could be described more elaborately. The principal researcher composed the scale and had more insight into the meaning of scale positions, for instance; very urgent, meaning the inability to finish a subtest as intended. In the phase of discussion, these ambiguities were resolved, resulting in consensus of urgency for the usability issues. A discussion beforehand, regarding the usability issues and scale positions would have reduced these ambiguities, resulting in fewer differences in the evaluation of urgency.

The SLTs completed LITMUS_NL by themselves, without assessing it together with children. Therefore, information regarding children's opinions and applicability with children is missing. Although many SLTs tried taking reaction time into account, the measured length of time could be less accurate in daily practice. Applicability was measured by opinions based on the SLTs' clinical expertise and could deviate when involving children in the usability study.

This study adds information about the opinions of SLTs towards LITMUS_NL. Since these opinions are part of the barriers and facilitators regarding LITMUS_NL, it offers crucial information to anticipate when implementing the test[33]**.** The usability issues show factors of importance to SLTs. Instructions must be clear and complete, and the user design should be convenient without bugs. Future research regarding the development of language tests should take these factors into account and should involve SLTs in the development of tests.

LITMUS_NL will be further adapted to improve the quality, enhancing the usability in daily practice. The participating SLTs were positive about the instrument, indicating that implementation might be successful. Further research investigating the feasibility in daily practice and the reactions of children to LITMUS_NL is recommended. Once updated and changed according to recommendations of SLTs, LITMUS_NL can be expected to have a significant impact on identifying multilingual children with DLD.

Conclusion

The LITMUS_NL test was developed to identify DLD in multilingual children. Our study investigated usability and found a good degree of usability and positive evaluations of clinical first impression. However, the SLTs found solvable usability issues regarding technology and instructions, but also issues related to the construct of the test, which are difficult to adapt. By resolving the usability issues, LITMUS_NL can be implemented as a useful test to identify DLD in multilingual children.

References

1. Gerrits E, Beers M, Bruinsma G and Singer I. Handboek taalontwikkelingsstoornissen. Bussum: Coutinho, 2017.

2. Bishop DVM, Snowling MJ, Thompson PA and Greenhalgh T. CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. PloS one. 2016 Jul 8; 11(7): e0158753.

3. Nederlandse Vereniging voor Logopedie en Foniatrie. Richtlijn Logopedie bij Taalontwikkelingsstoornissen. Woerden: NVLF, 2017.

4. Wiefferink K, van Beugen C, Wegener Sleeswijk B and Gerrits E. Children with language delay referred to Dutch speech and hearing centres: caseload characteristics. Int. J. Lang. Commun. Disord. 2020 May 27; 55(4): 573-582.

5. Taylor-Goh S. Royal College of Speech & Language Therapists Clinical Guidelines. London: Routledge, 2017.

6. Thordardottir E, Rothenberg A, Rivard M and Naves, R. Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages?. J. Multiling. Commun. Disord. 2009 Sep29; 4(1): 1-21.

7. Hamann C and Abed Ibrahim L. Methods for Identifying Specific Language Impairment in Bilingual Populations in Germany. Frontiers in Communication. 2017 Oct 25; 2.

8. Kohnert K. Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. J. Commun. Disord. 2010 Feb; 43(6): 456-473.

9. Armon-Lotem S, de Jong J and Meir N. Assessing multilingual children. Bristol: NBN International, 2015.

10. Armon-Lotem s. Introduction: Bilingual children with SLI – the nature of the problem. Bilingualism. 2012 Jan; 15(1): 1-4.

11. Boerma TD, Leseman PPM, Timmermeister M, Wijnen FNK and Blom WBT. Narrative abilities of monolingual and bilingual children with and without language impairment: implications for clinical practice. Int. J. Lang. Commun Disord. 2016 Mar 15; 51(6): 626-638.

12. Orellana CI, Wada R and Gillam RB. The Use of Dynamic Assessment for the Diagnosis of Language Disorders in Bilingual Children: A Meta-Analysis. American J. of speech-language pathology. 2019 Jun 13; 28(3): 1-20.

13. Boerma T, Chiat S, Leseman P, Timmermeister M, Wijnen FNK and Blom WBT. A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children Learning Dutch as a Second Language. JSLHR. 2015 Dec; 58(6): 1747-1760.

14. Boerma T and Blom E. Assessment of bilingual children: What if testing both languages is not possible? J. Commun. Disord. 2017 Apr; 66: 65-76.

15. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Monitoring and Evaluating Digital Health Interventions. Geneva: World Health Organization, 2016, p.843.

16. Ehrler F, Weinhold T, Joe J, Lovis C and Blondon K. A Mobile App (BEDSide Mobility) to Support Nurses' Tasks at the Patient's Bedside: Usability Study. JMIR mHealth and uHealth 2018; 6(3): e57.

17. De Luca R, Maggio MG, Naro A, Portaro S, Canavò A and Calabrò RS. Can patients with severe traumatic brain injury be trained with cognitive telerehabilitation? An inpatient feasibility and usability study. J of clinical neuroscience. 2020 Sep; 79: 246-250.

18. Broekhuis M, van Velsen L and Hermens H. Assessing usability of eHealth technology: A comparison of usability benchmarking instruments. Int. J. of medical informatics 2019; 128.

19. Gerber SM, Schütz N, Uslu AS, Schmidt N, Röthlisberger C, Wyss P, et al. Therapist-Guided Tablet-Based Telerehabilitation for Patients With Aphasia: Proof-of-Concept and Usability Study. JMIR rehabilitation and assistive technologies 2019; 6(1): e13163.

20. Klaassen B, van Beijnum, B. J. F and Hermens HJ. Usability in telemedicine systems—A literature survey. Int. J. of medical informatics. 2016 Sep; 93: 57-69.

21. Gagarina N, Klop D, Kunnari S, Tantele K Välimaa T, Balciüniené I, et al. MAIN: Multilingual Assessment Instrument for Narratives. ZASPIL. 2012 Dec; 56.

22. Baldi B and Moore DS. Practice of Statistics in the Life Sciences. New York: Macmillan Learning, 2018.

23. Brooke. SUS: A 'Quick and Dirty' Usability Scale. In: Usability Evaluation in Industry: Boca Raton FLO: CRC Press; 1996, p.207.

24. Bangor A, Kortum P and Miller J. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Journal of Usability Studies. 2009 May; 4: 114-123.

25. Bangor A, Kortum PT and Miller JT. An Empirical Evaluation of the System Usability Scale. Int. J. of human-computer interaction. 2008 Jul 30; 24(6): 574-594.

26. Sauro J and Lewis J. Quantifying the User Experience. Burlington MA: Morgan Kaufmann; 2012.

27. Federici S, Borsci S and Stamerra G. Web usability evaluation with screen reader users: Implementation of the partial concurrent thinking aloud technique. Cogn Process. 2010 Aug 1; 11(3): 263-272.

28. van Velsen L, van der Geest T and Klaassen R. Identifying Usability Issues for Personalization During Formative Evaluations: A Comparison of Three Methods. Int. J. of human-computer interaction 2011 Jul 1; 27(7): 670-698.

29. Fonda SJ, Paulsen CA, Perkins J, Kedziora RJ, Rodbard D and Bursell SE. Usability Test of an Internet-Based Informatics Tool for Diabetes Care Providers: The Comprehensive Diabetes Management Program. Diabetes technology & therapeutics 2008 Feb; 10(1): 16-24.

30. Ryan RM. Control and information in the interpersonal sphere: an extension of cognitive evaluation theory. Journal of Personality and Social Psychology 1982; 43(3): 450-461.

31. Paul den Boer and Jos Frietman. Arbeidsmarktonderzoek logopedie 2015. Nijmegen: KBA, 2016.

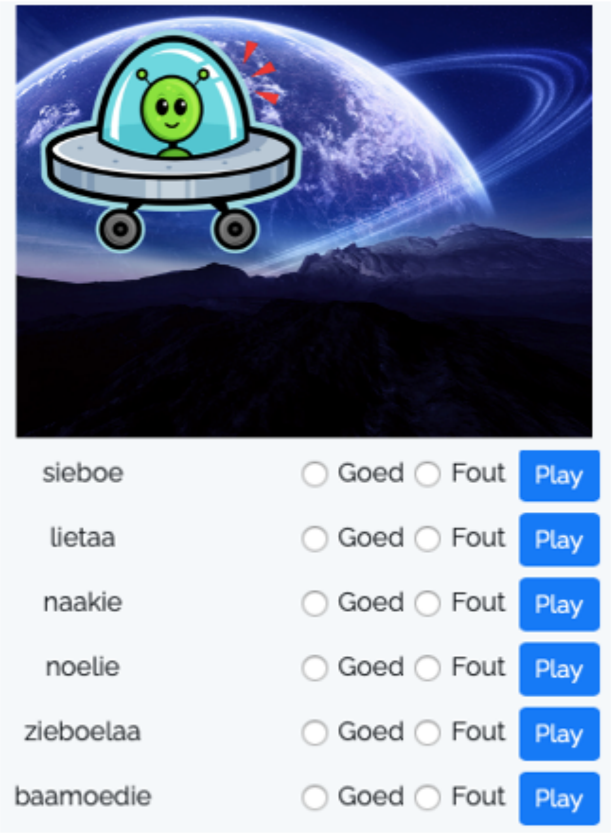
32. Peleg M, Shachak A, Wang D and Karnieli E. Using multi-perspective methodologies to study users' interactions with the prototype front end of a guideline-based decision support system for diabetic foot care. Int. J. of medical informatics 2009; 78(7): 482-493.

33. Wensing M and Grol R. Implementatie. Houten: Bohn Stafleu en van Loghum; 2017.

Tables and Figures

Figure 1

Print screen of the nonword repetition test



sieboe	○ Goed	○ Fout	Play
lietaa	○ Goed	○ Fout	Play
naakie	○ Goed	○ Fout	Play
noelie	○ Goed	○ Fout	Play
zieboelaa	○ Goed	○ Fout	Play
baamoedie	○ Goed	○ Fout	Play

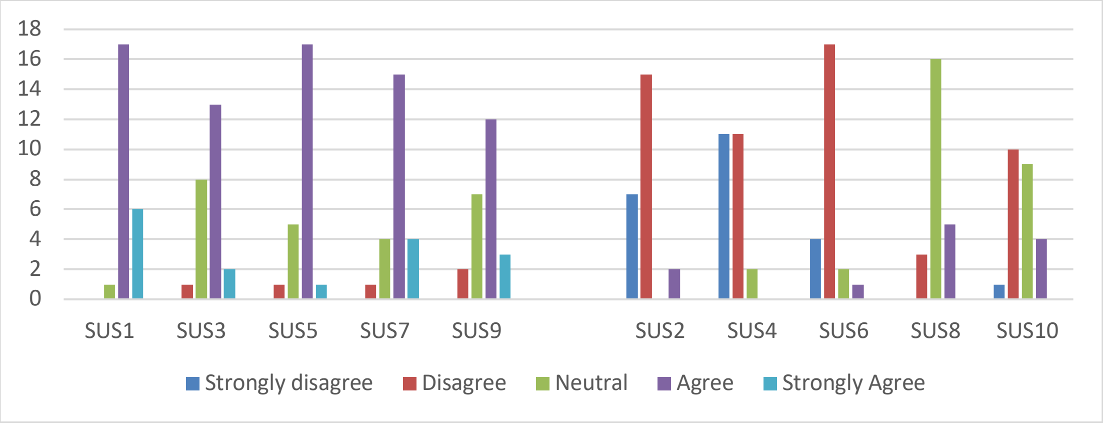
Table 1

Sample characteristics

Age (in years)	
Median (IQR)	31.5 (13)
Range	*25-56*
Gender	
Female	N= 24
Male	N= 0
Work Experience as SLT (in years)	
Median (IQR)	7.8 (11.6)
Range	*.5 - 24*
Work setting	
SLT practice	N= 16
Speech Hearing Centre	N= 8
Self-reported digital literacy (0-10)	
Median (IQR)	8.0 (.7)
Range	*7.0-10.0*
Self-reported use of other digital instruments	
Yes	N= 15
No	N= 9

Figure 2

Frequencies of SUS scores per item[a]

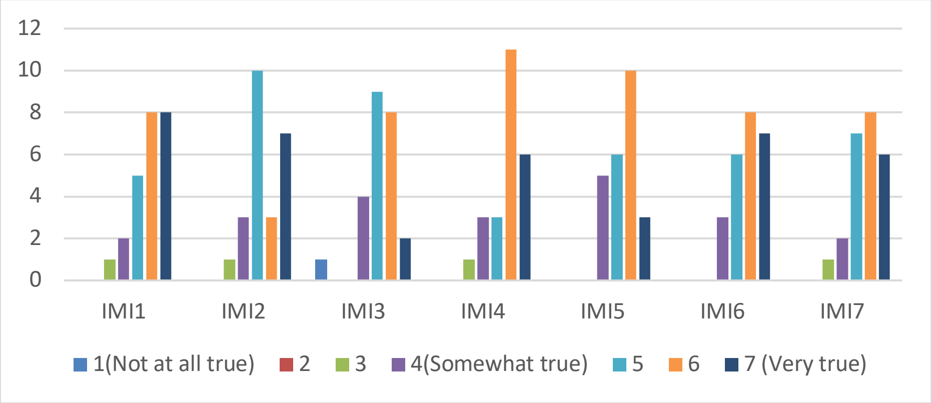


Note. SUS1: I think I would like to use LITMUS_NL frequently
SUS3: I thought LITMUS_NL was easy to use
SUS5: I found that the various functions were well integrated
SUS7: I would Imagine that most people would learn to use LITMUS_NL very quickly
SUS9: I felt very confident using LITMUS_NL
SUS2: I found LITMUS_NL unnecessarily complex
SUS4: I think I would need technical support to use LITMUS_NL
SUS6: I thought that there was too much inconsistency in LITMUS_NL
SUS8: I found LITMUS_NL cumbersome to use
SUS10: I needed to learn a lot of things before I could get going with LITMUS_NL

[a]Odd items were positive-stated questions, meaning ‘Agree’ is a positive score. Even items were stated negatively, meaning ‘Disagree’ is a positive score.

Figure 3

Frequencies of IMI scores per item



Note. IMI1: I believe LITMUS_NL could be of some value for me.
IMI2: I think that doing LITMUS_NL is useful for diagnosing DLD in multilingual children.
IMI3: I think LITMUS_NL is important because it can determine language abilities of multilingual children.
IMI4: I would be willing to use LITMUS_NL again because it has value to me.
IMI5: I think LITMUS_NL would help me to determine language abilities of multilingual children.
IMI6: I believe LITMUS_NL could be beneficial to me.
IMI7: I think LITMUS_NL is important.

Table 2

Categories of usability issues with frequencies of issues per LITMUS_NL subtest

Issue Category	Gen	NWRT	Rec	Expr	Sen	MAIN	MAIN	**Total**
Subcategory			Voc	Voc	Rep	Comp.	S.T	
Instructions	5	8	9	14	14	3	7	60
Scoring	0	2	0	4	2	3	0	11
Technical								
Performance	5	n/a	n/a	n/a	n/a	n/a	n/a	5
Bugs	3	5	2	8	6	3	3	30
User Experience	5	3	5	9	5	4	2	33
Notifications	3	0	3	9	1	0	0	16
User Design	1	6	6	3	15	2	2	35
Features	0	0	0	0	0	0	1	1
Development of tests	3	2	11	8	6	3	1	34
Item specific Content	0	1	13	20	4	3	0	41
Item specific Presentation	0	0	26	20	2	2	0	50
Spelling errors	0	0	2	6	0	0	1	9
Pronunciation of items	0	10	n/a	n/a	2	n/a	n/a	12
Total	25	37	77	101	57	23	17	**337**

Note. Gen= General; NWRT= Nonword repetition test; Rec Voc= Receptive vocabulary test; Expr Voc= Expressive vocabulary test; MAIN comp= MAIN story comprehension test; MAIN S.T.= MAIN storytelling test; n/a= not applicable.

Table 3

Usability issues prioritized as very urgent by 2/2 researchers

Usability issue	LITMUS_NL Test	Category	Frequency
Need for multiple correct answers	Expressive vocabulary	Instructions: Scoring	14
Starting second screen is not working	General	Technical: Performance	7
Audio is not working	Nonword repetition	Technical: Bugs	5
Notification: 'wait on score sheet/participation screen': filling out answers is not possible	Sentence Repetition	Technical: Notifications	4
Notification not all items are filled out: cancelling does not work	Expressive vocabulary	Technical: Notifications	3
Not able to fill out answers	Expressive vocabulary	Technical: Bugs	2
Audio recording is not in instructions	MAIN- storytelling	Instructions	2
Program locks, all items need to be scored again	Receptive vocabulary	Technical: Bugs	1
Participant screen is not showing anything	General	Technical: Bugs	1
Scoring correct/incorrect is not working anymore	Expressive vocabulary	Technical: Bugs	1
Participant screen is not showing pictures	MAIN- storytelling	Technical: Bugs	1
Participant screen does not respond to score sheet	Nonword repetition	Technical: Bugs	1

Table 4

Median VAS scores and ranges of the questions regarding clinical first impression

	Median (IQR)	***Range***
Buttons		
Nonword Repetition Test	8.00 (1.1)	*4.0-10.0*
Vocabulary Test	7.80 (1.3)	*3.0-10.0*
Sentence Repetition Test	7.60 (1.7)	*3.9-9.0*
MAIN	7.75 (1.5)	*2.4-10.0*
Score sheet		
Nonword Repetition Test	8.00 (2.0)	*6.0-10.0*
Vocabulary Test	8.00 (1.0)	*4.0-9.0*
Sentence Repetition Test	8.00 (2.0)	*2.0-10.0*
MAIN	8.00 (2.0)	*2.0-9.0*
Complexity	8.00 (1.2)	*4.0-9.0*
Applicability	7.00 (.8)	*5.0-9.2*
Feasibility		
Length of time	8.10 (1.5)	*1.0-10.0*
Manual scoring MAIN	8.00 (2.8)	*.0-10.0*
Design	7.80 (1.1)	*3.0-8.6*
Relevance language assessment	7.20 (1.0)	*2.9-8.5*

Note. All items were scored on a scale ranging from 0-10 and are stated positively, meaning a higher score is interpreted as a positive outcome.

APPENDIX I: VAS-questions regarding clinical first impression

1. To what extend is the placement and naming of buttons of LITMUS_NL logical and intuitive for each subtest?

Illogical and unclear Logical and intuitive

0 10

a. Nonword repetition test

b. Vocabulary (receptive and expressive)

c. Sentence repetition

d. Storytelling tests

2. To what extend are the score forms logical and intuitive for each subtest?

Illogical and unclear Logical and intuitive

0 10

a. Nonword repetition test

b. Vocabulary (receptive and expressive)

c. Sentence repetition

d. Storytelling tests

3. To what extend are the instructions complex?

Very complex Not at all complex

0 10

4. To what extent is the instrument acceptable to use in daily clinical practice regarding the duration?

Not acceptable: duration is too long Acceptable: duration is good

0 10

5. The storytelling test (in which the child tells a story) is scored manually with an audio recording. To what extend is this feasible in daily clinical practice?

Not feasible: Feasible

Too much time investment/ live scoring is better for me/...

0 10

6. To what extent is LITMUS_NL applicable in multilingual children in the age of 3 to 9 years old?

Not applicable Somewhat applicable Very applicable

0 10

7. To what extend does this combination of subtests give insight in the language skills of multilingual children?

No insight in language skills Sufficient insight:
offers leads for treatment

0 10

8. To what extent is the presentation and design of the subtests of LITMUS_NL cohesive?

Incohesive Cohesive

0 1

9. What do you think about the design?

Do you have any other remarks regarding the LITMUS_NL language test or this questionnaire?