Acta Logopaedica (2025), Vol. 2 | 36–57
DOI: 10.58986/al.2025.34243

Acta Logopaedica

ORIGINAL ARTICLE

Development of a Swedish classroom screening tool for grammatical comprehension (LegiLexi GRA) in young school-age children

André Kalmendal[1*], Amlin Al Emara[2], Anna Eva Hallin[2]

[1]Department of Psychology, Linnaeus University, Växjö, Sweden; and [2]Division of Speech and Language Pathology, Karolinska Institutet, CLINTEC, Stockholm, Sweden

***Corresponding author**: André Kalmendal, Department of Psychology, Linnaeus University, Växjö, Sweden. Email: andre.kalmendal@lnu.se

Publication date: 9 May 2025

Abstract

Background: Early identification of grammatical comprehension difficulties is critical for supporting language and reading development and academic achievement in young learners. There is a lack of robust whole-class screening tools of grammatical listening comprehension, particularly in linguistically diverse contexts such as Swedish primary education.

Objective/Aim: This study aimed to develop and evaluate the psychometric properties of the LegiLexi GRA-10, a brief classroom screening tool designed to detect students with challenges in grammatical comprehension among first-year students in Swedish compulsory school, including monolingual and multilingual students with Swedish as their first language (L1) and second language (L2) learners of Swedish.

Material and Methods: An initial 16-item version of the test was administered to a large sample of first-year students (N=8245). Item Response Theory (IRT) analyses were employed to select the ten best performing items to form the GRA-10. The tool's discriminatory power and reliability were evaluated across both L1 and L2 student populations.

Results: The psychometric evaluation demonstrated that the GRA-10 reliably detects students with grammatical comprehension challenges, using a cut-off score of eight points or lower. The tool exhibited strong internal consistency and high discriminatory power across both L1 and L2 groups. These results indicate that GRA-10 is effective in identifying students who may benefit from early targeted language support.

Conclusion: The LegiLexi GRA-10 offers a valid, reliable, and efficient method for classroom-based screening of grammatical comprehension in young Swedish learners. Its strong psychometric properties across diverse language backgrounds support its use

as an early identification tool, facilitating timely and targeted educational support. The GRA-10 represents a valuable addition to the resources available for promoting language development and academic success in the early years of schooling.

Keywords: grammatical comprehension; test development; early literacy screening

Abstract

Bakgrund: Tidig identifiering av svårigheter med grammatisk förståelse är viktigt för att stödja språk- och läsutveckling samt skolframgång tidigt i grundskolan. Det saknas robusta verktyg för helklass-screening av grammatisk förståelse, inte minst i språkligt heterogena sammanhang som svensk grundskola.

Syfte: Denna studie syftade till att utveckla och utvärdera de psykometriska egenskaperna hos LegiLexi GRA-10, ett kort screeningverktyg som används i helklass. GRA-10 är utformat för att identifiera elever med utmaningar i grammatisk förståelse i svensk grundskolas första år, både hos enspråkiga och flerspråkiga elever som följer kursplanen för svenska (L1) och elever som följer kursplanen för svenska som andraspråk (L2).

Material och Metod: En ursprunglig version av testet med 16 uppgifter administrerades till ett stort urval av elever i förskoleklass (N=8245). Item Response Theory (IRT)-analyser användes för att välja ut de tio bäst presterande uppgifterna till GRA-10. Verktygets diskriminerande förmåga och tillförlitlighet utvärderades för både L1- och L2-elever.

Resultat: Den psykometriska utvärderingen visade att GRA-10 på ett tillförlitligt sätt identifierar elever med svårigheter i grammatisk förståelse, med resultat på åtta poäng eller lägre. Verktyget uppvisade stark intern konsistens och hög diskriminerande förmåga i både L1- och L2-grupperna. Resultaten indikerar att GRA-10 är effektivt för att identifiera elever som kan ha nytta av tidiga och riktade stödinsatser för språk.

Slutsats: LegiLexi GRA-10 erbjuder en tillförlitlig och effektiv metod för klassrumsbaserad screening av grammatisk förståelse hos svenska elever i grundskolans första år. De starka psykometriska egenskaperna över elever med olika språkbakgrunder stöder dess användning som ett verktyg för tidig identifiering, vilket möjliggör tidig och riktad pedagogisk support. GRA-10 utgör ett värdefullt tillskott till de resurser som finns för att främja språk- och läsutveckling samt skolframgång i de tidiga skolåren.

Keywords: grammatisk förståelse; testutveckling; tidig screening av läsförmåga

Introduction

Many children and adolescents in the Organisation for Economic Cooperation and Development (OECD, 2023: PISA 2022 Results) have language and literacy challenges, which may affect academic achievement and carry personal and societal costs (Cronin *et al.*, 2020). In Sweden, as in many other countries, this negative trend is already seen in elementary schools: In Spring 2023, 27% of all students who follow the curriculum for Swedish as an L1 failed one or more of the eight subtests in the national assessments of reading and writing in grade 3, and this number increases to almost 61% of students who follow the curriculum for Swedish as an L2 (Statistics

Sweden, 2024a). To provide early and targeted intervention, it is central to detect students at risk for reading difficulties in the first years of school and give appropriate support. Researchers have called for universal screening for language and literacy skills, preferably in a response-to-intervention (RTI) framework, where early detection and intervention is in focus, not diagnosis (e.g. Adlof and Hogan, 2019; Ebbels *et al.*, 2021; Hulme *et al.*, 2024). Studies are showing that early oral language difficulties are not discovered to the same extent as early written language difficulties, with one reason being that oral language skills are not always being screened as systematically as decoding skills (Adlof and Hogan, 2019). Bao *et al.* (2024) did a systematic review of published screening tools to identify students with developmental language disorder (DLD), a neurodevelopmental condition characterised by difficulties in learning, using, and understanding one's native language(s) despite adequate language exposure (Bishop *et al.*, 2016). They found 13 commercially available English DLD screeners suitable for young school-age children (5 years old and above). None of the tests was for screening by a teacher in a group/class setting, which the authors mean likely limits universal implementation of systematic classroom screening.

In Sweden, the non-profit LegiLexi foundation (www.legilexi.org) offers a free digital whole-class screening tool to screen both oral language and literacy skills and monitor student progress for grades 0–3 (since Fall 2017) and grades 4–6 (since Fall 2024). The tool has been created in collaboration with a group of independent volunteer researchers in language, literacy, and pedagogy from several universities around Sweden and has been continuously improved through analyses of de-identified student data (e.g., by beta-testing new items, and omitting poorly performing items). The tool has a starting point in the reading model: *The Simple View of Reading* (Hoover and Gough, 1990; Nordström *et al.*, 2025), which states that reading comprehension is the product of word recognition and language (listening) comprehension, with both components affected by and associated with many underlying cognitive skills (Hoover and Tunmer, 2022). Thus, the tool includes subtests for reading comprehension, word recognition, word recognition-related skills (knowledge of grapheme–phoneme connections and phonemic awareness) and language comprehension. Initially, subtests for language comprehension only included one subtest targeting receptive vocabulary and one subtest targeting listening comprehension for grades 0–3. In Fall 2022, a third language comprehension subtest was added to the tool following an initiative of one researcher in the researcher group (Anna Eva Hallin, the last author of this paper). This subtest, called LegiLexi GRA (*Sw: GRAmmatisk språkförståelse* [GRAmmatical langauge comprehension]), screens grammatical comprehension in Swedish grade 0 (*Sw: förskoleklass*). Most children begin this grade in August of the calendar year they turn 6 years old. With the increasing popularity and widespread use of the LegiLexi screening tool (more than 37,000 Swedish teachers are active users as on March 1st, 2025 according to the foundation), and a new statistician in the volunteer researcher group (André Kalmendal the first author of this paper), the group decided to assess the measurement validity of all LegiLexi subtests using the Item Response Theory (IRT). If warranted, they aimed to improve the test item structure based on the results and publish the findings to make this information publicly available for this widely used screening tool.

Recently, Hulme *et al.* (2024) conducted a similar evaluation of a British app-based language screening tool called LanguageScreen, designed for children

aged 3–9 years. LanguageScreen is administered individually by the children's teachers or teachers' assistants, and using data from more than 8,000 schools and almost 3,49,000 students, the researchers fitted the data to a Rasch model to investigate the tool's psychometric properties. They concluded that the test's total score was sufficient to estimate a child's language ability. Similar to Hulme *et al.* (2024), we aimed to leverage large datasets generated by widely used digital tools to calculate the psychometric properties of language and literacy screening tests.

This is the first paper in a planned series of such publications investigating the screening tool provided by the LegiLexi foundation, focusing on the development and the measurement validity of LegiLexi GRA.

Grammatical (or syntactic) comprehension is one of the oral linguistic skills that support reading comprehension through language comprehension (Hoover and Tunmer, 2022), and research confirms that children with difficulties with grammatical comprehension perform lower on reading comprehension tasks (e.g. Poulsen *et al.*, 2022). Challenges with grammatical or wider language comprehension, regardless of the underlying cause, are difficult to detect by both teachers and parents, unless structured screening is implemented (Adlof *et al.*, 2017; Hendricks *et al.*, 2019). One group of students having difficulties with grammatical comprehension had DLD, affecting 7–10% of all students (Norbury *et al.*, 2016). Another group that is at risk for grammatical comprehension difficulties and might need extra classroom support are the students who are second language (L2) learners of the language of instruction. In Sweden, L2-learners of Swedish form a considerable minority of 13.9% of the student body in primary schools (Statistics Sweden, 2024b).

There are different ways to operationalise and test grammatical comprehension, where some test designs arguably also tap into vocabulary knowledge, fluid reasoning, linguistic awareness, and in some cases expressive abilities (Nielsen *et al.*, 2024). Clinically, the most used standardised norm-referenced test for grammatical comprehension in Sweden is the *Test for Reception of Grammar*, version 2 (TROG-2; Bishop and Garsell, 2009), an 80-item test where the student listens to one sentence at a time and chooses one picture out of four as their response. The picture alternatives are designed so that the listener needs to understand the syntactic relationship between the words (e.g. subject and object) to choose the correct picture. Hendricks *et al.* (2019) used a subset of 16 items from the English version of TROG-2 to develop a classroom screening for grammatical comprehension in US grades 1 and 2, reporting an overall sensitivity of 76% and a 25% false positive rate for DLD in a study in their sample of 97 children, and concluded that this measure had a promise in detecting children with language comprehension difficulties due to DLD.

The initial development of LegiLexi GRA

Encouraged by the results from Hendricks *et al.* (2019), and knowing that language comprehension difficulties often go undetected in the early school years, the LegiLexi foundation decided to support the development of LegiLexi GRA, with an item structure modelled after TROG-2 and having 16 items similar to Hendricks *et al.* (2019) in the first version. The development was initiated and led by Anna Eva Hallin, the last author of this paper, a speech-language pathologist (SLP), and a researcher in language and literacy development and disorders, who also created

https://doi.org/10.58986/al.2025.34243

the items. The LegiLexi foundation recorded the sound for all items, provided an illustrator, and had developers incorporate GRA in the online LegiLexi test platform. In addition, the LegiLexi foundation provided anonymised data from pilot and beta-testing, and assisted with data analysis and visualisation during the first phases of GRA development.

Like most other LegiLexi subtests, GRA is administered by a teacher in a group/classroom setting, with each child doing the self-paced test on a computer or tablet. The recorded items are presented once in headphones and the child makes their response choice on the screen. In contrast to other subtests in LegiLexi's tool, GRA was not designed to track development over time, but rather to detect grammatical difficulties in 5- to 7-year olds. Thus, the test was designed so that most students in the first year of Swedish compulsory school would find the test relatively easy and ceiling effects were expected, which are similar to the planned ceiling effects in a whole-class language screening task, including vocabulary, grammar, and higher-level language comprehension (e.g. inferences) used in Adlof *et al.* (2017).

Initially, 24 items were created. The starting point was syntactic structures that Swedish primary school children should master (based on TROG-2 normative data and knowledge about typical language development in Swedish), and those that Swedish-speaking children with DLD may have in particular difficulties (Reuterskiöld *et al.*, 2021). All 24 sentences contained only high-frequency concrete nouns and verbs that would be known by Swedish primary school children. The author also specified three foils for each item to make sure that the child had to rely on grammatical comprehension to choose the right picture. For an item example, including the foils, see Appendix A. For the first pilot test (October 2021–January 2022) to arrive at a 16-item test, the 24 items were divided into two 12-item tests and piloted on 994 L1 and 114 L2 students in grades 0–1 (5- to 7-year olds), with the help of volunteer teachers among LegiLexi users. In addition, a pre-vocabulary test was added to ensure that L2 learners would not get low results at GRA due to low vocabulary knowledge only. Based on the response patterns of those students and surface characteristics of the items, a 16-item test was created: GRA-16 (for all items with translations; see Appendices A and C). To investigate whether GRA-16 explained variance in LegiLexi listening comprehension above and beyond LegiLexi vocabulary scores, data from 2,465 students in grades 0–3 (2,050 L1 students and 415 L2 students, 5–9 years old) were analysed in a degree project (Kaya, 2022). GRA-16 scores explained around 7% unique variance in listening comprehension for both L1 and L2 students in models, including age, vocabulary, and GRA-16 scores, results similar to a recent study that found that the Danish version of TROG-2 explained 8% variance in reading comprehension in 4th graders after decoding, vocabulary, and fluid reasoning were included in a regression model (Nielsen *et al.*, 2024).

In the second half of 2023, the LegiLexi foundation decided to decrease the number of items in GRA from 16 to 10. This decision was based on feedback from their teacher users, with the explicit end goal to decrease the total test time for the full LegiLexi tool. The selection of the 10 items for GRA-10 was again based on surface characteristics and the responses of students in grades 0 and 1 who had participated in voluntary beta-testing. The aim was to keep good performing items based on participant responses, but no advanced statistical methods were used. All items included in the first version of GRA-10 and their translations are presented

in Appendix A. The pre-vocabulary test was limited to five items, and students had to have three vocabulary items correct to proceed to do GRA-10. To investigate the external validity of this first version of GRA-10, another degree project compared the results of GRA-10 and TROG-2 in a sample of 87 students in grade 0 (Al Emara, 2024). Correlation between TROG-2 (block score) and GRA was moderate to good for the whole sample (rho = 0.66), and moderate when split into the L1 group (rho = 0.56, n = 60) and the L2 group (rho = 0.53, n = 27). In addition, Al Emara (2024) found that GRA-10 reliably detected L1 students with grammatical comprehension difficulties according to TROG-2: only 3% of L1-students had a GRA-10 score of 9 or higher while identified as having grammatical difficulties by TROG-2, defined as a standard score of 80 or below (i.e. false negatives). In addition, only 8% of all students were false positives, with a score of 8 or lower on GRA-10 but a standard score above 80 on TROG-2 (9% of L1 students and 6% of L2 students). The conclusion was that GRA-10 had an acceptable external validity, especially for L1 students, and that a cut-off score of 8 or lower indicating grammatical difficulties in L1 students was reasonable. Al Emara (2024) suggested that an item analysis through statistical methods would complement her results to further develop GRA-10, especially for L2 students, because the data set used in her study was small, and TROG-2 was a less suitable test for this group due to TROG-2 stop criteria and the normative sample. In a follow-up conference presentation, Hallin *et al.* (2024) used a subset of the data from Al Emara (2024) to investigate the external validity of the new GRA-10, whose development and statistical properties are described in the present paper, and showed an improvement in both correlation with TROG-2 block scores and classification accuracy of students with and without grammatical difficulties: these results are further outlined in the Discussion section.

To summarise, GRA-16 was first developed based on knowledge of grammatical development in Swedish children with and without DLD and results from initial pilot data testing, and the subsequent GRA-10 was developed based on user request for a shorter total screening time, using student beta-test results and surface characteristics and validated against TROG-2 (Al Emara, 2024). Because the selection of items in the first version of GRA-10 was not statistically driven, the present study aims to extend the above work and use statistical models to create the best possible version of GRA-10 based on beta-test results from GRA-16. The IRT framework (Hambelton and Jones, 1993) is used to examine the psychometrics of all possible 10-item combinations derived from GRA-16 to find the best 10-item set for GRA-10. The IRT is a general statistical theory about how performance relates to the abilities that are measured based on the included examinee item and test performance. The statistics generated by the IRT framework share many similarities with factor analysis and are transformed into factor intercepts and loadings and *vice versa* (Zhang *et al.*, 2023).

Aims

As previously stated, this study is an effort by independent volunteer researchers collaborating with the LegiLexi foundation to assess and, if possible, increase the measurement validity of LegiLexi GRA-10 to provide an efficient and reliable tool to support the identification of grammatical comprehension challenges in Swedish students in grade 0 (5- to 7-year olds).

Specifically, we:

1. Select the optimal 10 items that maximise the tool's precision in identifying students with or without risk of grammatical comprehension difficulties.
2. Evaluate the psychometric properties of the new 10-item test, such as reliability, validity, and test measurement accuracy.
3. Examine differences in item performance for students following Swedish as a first (L1) or second language (L2) curriculum.

Method

The GRA-16 results were provided by the LegiLexi foundation to the first author (André Kalmendal), where all identifying personal information was omitted. The data consisted of results from 8,245 Swedish students in grade 0 from 356 different schools across Sweden, collected during January–June 2023. The majority (88%) were mono- and multilingual students enrolled in the regular Swedish curriculum for students with Swedish as a first language (L1). The remaining 12% were enrolled in the curriculum for Swedish as a second language (L2). Because the data is anonymised and the LegiLexi tool does not require teachers to give the date of birth of their students, the exact ages of the students are unknown, but almost all students in Sweden start grade 0 in August, the year they turn 6 years old. This means that the ages of the students in our sample should vary between 6 years, 0 months and 7 years, 5 months, depending on their date of birth and when in the semester they participated in GRA testing.

Procedure

GRA-16 consisted of 16 multiple-choice items (see description in the Introduction). For each item, four pictures were presented accompanied by a corresponding spoken sentence (see Appendices A and C). GRA-16 was administered in a classroom setting by teachers or special educators aided by instructions provided by LegiLexi. Each student completed the test individually using a computer/tablet with headphones. The test begins with a five-item practice vocabulary set, where one spoken everyday word is presented at a time along with four pictures, and the student has to make the correct choice. To proceed to the grammatical comprehension items, students must answer three out of five vocabulary items correctly. Students could listen to the vocabulary items an unlimited number of times to familiarise themselves with the test structure and response format; however, they could only listen to each sentence in GRA-16 once before making a choice. Because GRA-16 items consist of grammatical structures that should be mastered in the target age group, allowing the students to listen more than once would increase the risk of missing students with grammatical comprehension challenges.

Ethical considerations

According to Swedish law (2003:460), ethical approval is needed for research that involves physical intervention as well as biological, genetic, biometric, or sensitive personal data. Because LegiLexi data is collected by the students' ordinary teachers to aid teaching, does not contain any sensitive personal information, and cannot be connected to any individual student, no ethical approval was sought for this study.

A statement from the Swedish Ethical Review Authority has confirmed that studies including only de-identified LegiLexi data do not need additional ethical approval.

Data Analysis

The data was modelled using a 3PL model (three-parameter logistical model) within the IRT framework, using the R package *mirt* (Chalmers, 2012). The 3PL model allowed us to estimate three key parameters: item difficulty, discrimination, and the guessing parameter.

The 3PL model, in comparison to the more commonly used Rasch (1960) model (Debelak *et al.*, 2022), allows more variables free to vary. For example, the Rasch model assumes the slope to be fixed and treats all items as discriminant, while the 3PL model allows each item to have a unique discrimination parameter. Another strength of the 3PL model is that the intercept on the probability axis is free to vary, while under the Rasch model, it is fixed. This allows the model to calculate a guessing parameter ($P(\Theta)$) for each item. Students tend to guess on items they do not know and the model calculates a probability for that guessing to occur; this is especially important in the subgroup analysis where we want to examine whether there are any differences between L1 and L2 students, because the varying linguistic knowledge in the L2 group might increase the use of guessing as a strategy.

In Appendix B, the statistical calculations for the Rasch- and the 3PL-model are available as well as a comparison between the two models. In Appendix C, item difficulty and Cronbach's alpha for GRA-16 are available. All codes and data presented in this paper are available on OSF: https://osf.io/8zt3f/

Results

All items from GRA-16 were modelled and examined through IRT framework (see Appendix C). Ten items were selected based on the best item statistics to create a new GRA-10. The selection of items was based upon Cronbach's alpha values in combination with the correlation between an item score and the sum of the rest of the test items index (RIR); the RIR is used due to the use of binary data (Martinková and Hladká, 2023). Items that had a small or no contribution to the internal consistency in combination with the lowest RIR value were removed. Compared to the first version of GRA-10, which was developed based on student responses and surface characteristics, three items were exchanged based on the results from IRT (see Appendix A for all items included in GRA-16 and items in the previous and new GRA-10). Thus, all students in the dataset used for analyses participated in GRA-16, but the psychometric data presented here is for the new GRA-10, which has been used in LegiLexi's digital tool since Fall 2024.

Descriptives

In Table 1, descriptives of the new GRA-10 are presented. As expected, the mean score of L2 students is lower than for L1 students, and the standard deviation is also larger. In educational research, Cronbach's alpha value of 0.75 is generally considered to indicate good internal consistency (Taber, 2018) for a set of

Table 1. Descriptives of GRA-10 (min = 0, max = 10).

Sample	n	Mean	SD	Cronbach's alpha
All students	8,245	8.66	1.85	0.75
L1 students	7,254	8.84	1.72	0.74
L2 students	991	7.33	2.19	0.69

L1 = Swedish curriculum; L2 = Swedish as second language curriculum.

10 items measuring a single construct, such as grammatical comprehension. As shown in Table 1, the alpha value is slightly higher for L1 students, compared to L2 students.

Item descriptives

In Tables 2 and 3, item difficulty, using classical test theory (CTT), for the whole sample and L1 and L2 students are presented, respectively, the higher the score, the higher the proportion of right answers. The easiest item was "Hunden sover men inte katten" (*The dog is sleeping but not the cat*) and the hardest item was "Kaninen i lådan är vit" (*The rabbit in the box is white*). All items were more difficult for L2 students, and there were some slight differences between L1 and L2 students in the level of difficulty with minor deviations in the order of items between the hardest and the easiest item (Table 3). The L2 group had a slightly higher standard deviation on each item, compared to the L1 group, which is to be expected because there is a higher variance in performance among the L2 students.

IRT Model

Assumptions

Local dependence was tested with Yen's Q3 test with a threshold of 0.2 (Christensen *et al.*, 2017) and no item revealed any sign of dependency. To examine

Table 2. Item difficulty (CTT) with mean and standard deviation listed from the easiest to the hardest item.

Item (for translations, see Appendix A)	Difficulty	SD
05. Hunden sover men inte katten	0.937	0.243
06. Blomman är gul men inte stor	0.933	0.250
09. Flickan springer men pekar inte	0.930	0.256
08. Pojken blir jagad av hunden	0.910	0.287
15. Hunden som har mörk päls bär kaninen	0.890	0.312
10. Boken som är stor ligger på blomman.	0.856	0.351
11. Kaninen sitter bakom katten	0.839	0.367
03. Flickanserkaninenhoppa	0.798	0.402
13. Flickan ska ge hunden mat	0.795	0.404
02. Kaninen i lådan är vit	0.768	0.422

Table 3. Item difficulty (CTT) for L1 and L2 students separately.

	L1 Students		L2 Students	
Item (for translations, see Appendix A)	Difficulty	SD	Difficulty	SD
05. Hunden sover men inte katten	0.943	0.231	0.893	0.309
09. Flickan springer men pekar inte	0.940	0.237	0.855	0.353
06. Blomman är gul men inte stor	0.939	0.239	0.887	0.317
08. Pojken blir jagad av hunden	0.937	0.244	0.712	0.453
15. Hunden som har mörk päls bär kaninen	0.912	0.283	0.733	0.443
10. Boken som är stor ligger på blomman.	0.880	0.324	0.680	0.467
11. Kaninen sitter bakom katten	0.860	0.347	0.685	0.465
03. Flickan ser kaninen hoppa	0.811	0.392	0.701	0.458
13. Flickan ska ge hunden mat	0.811	0.391	0.674	0.469
02. Kaninen i lådan är vit	0.804	0.397	0.506	0.500

L1 = Swedish curriculum; L2 = Swedish as second language curriculum.

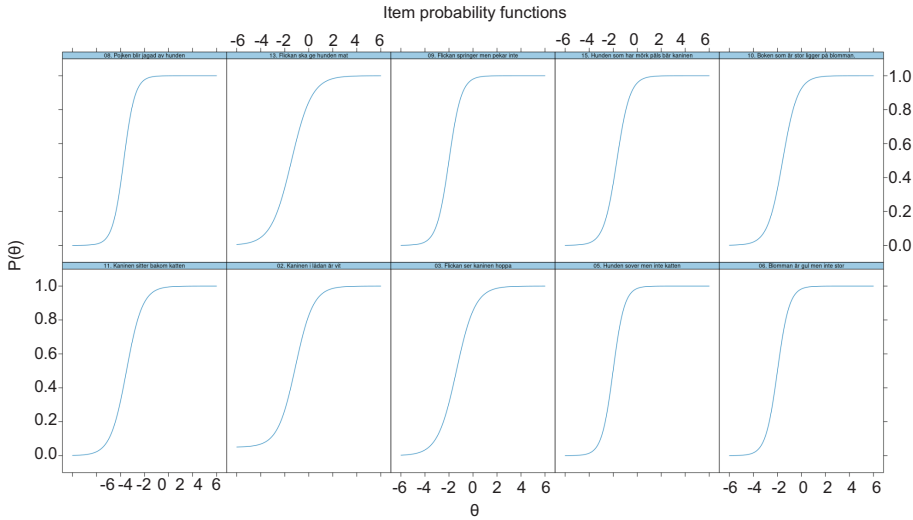
unidimensionality, an exploratory factor analysis was conducted. The data was calculated with tetrachoric correlation. Both models used the minimal residual method and the two-factor model, an oblimin rotation was selected. The one-factor model explained 49% of the proportion variance, RMSR = 0.03, TLI = 0.959, and RMSEA = 0.065 (90% CI [0.062, 0.068]). The two-factor model explained 52% of the variance, first factor 49% and the second factor 3%, RMSR = 0.01, TLI = 0.989, and RMSEA = 0.033 (90% CI [0.030, 0.037]). Due to a large sample size (n = 8,245), the Chi square statistic was not interpreted because a large sample size generates a significant result. According to Navarro and Foxcroft (2025), both models have a good fit TLI > 0.95 and satisfactory RMSEA for the one-factor model (< 0.08) and a good fit for the two-factor model (< 0.05). The results point to a strong case for unidimensionality. The small increase in explained variance with the two-factor model and the robust fit of the one-factor model suggest that the simpler one-factor solution provides a satisfactory representation of the data. Therefore, it is reasonable to conclude that the data is best described by a unidimensional structure based on the given analysis and fit statistics.

3PL Model

The 3PL model describes the data well; M2(25) = 43.32, p = .013, RMSEA = 0.001 (90% CI = [0.004, 0.014]), CFI = 0.999, TLI = 0.998, and SRMSR = 0.029. According to Martinková and Hladká (2023), a RMSEA value of less than 0.05 indicates a close fit, while CFI and TLI values of above 0.95 suggest a good fit. The 3PL model meets all these fit criteria, demonstrating a sufficiently good representation of the data.

As shown in Figure 1, all items effectively differentiate between students with and without grammatical comprehension difficulties. In the full model, the item slopes are positioned to the left of 0, which represents the mean ability level of the sample, further confirming that most items are relatively easy. Additionally, the guessing

Figure 1. ICC of LegiLexi GRA 10-item test (*N* = 8,245).



parameter P(Θ) is close to zero, indicating that the likelihood of answering correctly by chance is minimal.

Items with a low P(Θ) tend to offer better discrimination between ability levels, as they are less influenced by random guessing, an important factor, especially for multiple-choice items. Furthermore, a steeper slope on the Item Characteristic Curve (ICC) indicates that the item effectively distinguishes between students with lower and higher grammatical comprehension, reinforcing the test's ability to assess varying proficiency levels accurately.

Test accuracy

As described in the Introduction, LegiLexi GRA was designed to be a test that most students in grade 0 find easy (with the youngest students being 5 years and 9 months old at start), and the item descriptives presented above are in line with that. The Person-item map based on the 3PL model in Figure 2 further confirms this: all items are located below 0 on the difficulty parameter and the majority of the respondents are located close to or above 0. In general test development, this is usually not desired because the test might end up with a ceiling effect for students performing at or above average. We argue that this is not a psychometric issue in this test because the purpose of the test is to find low-performing students at risk and not to assess students' performance if they already acquired an average, age-adequate, or above-average level of grammatical comprehension. Instead, what this test achieves is very high accuracy in detecting children with grammatical comprehension difficulties. As we can see in Figure 3, most information and the lowest standard errors are derived from students performing below average. This increases the validity of Legilexi GRA as a good screening tool for grammatical comprehension difficulties.

https://doi.org/10.58986/al.2025.34243 Acta Logopaedica is published according to open access under a Creative Commons CC-BY 4.0 license

Figure 2. Person-item map.***02***. *Kaninen i lådan är vit*, ***03***. *Flickan ser kaninen hoppa*, ***05***. *Hunden sover men inte katten*, ***06***. *Blomman är gul men inte stor*, ***08***. *Pojken blir jagad av hunden*, ***09***. *Flickan springer men pekar inte*, ***10***. *Boken som är stor ligger på blomman*, ***11***. *Kaninen sitter bakom katten*, ***13***. *Flickan ska ge hunden mat*, ***15***. *Hunden som har mörk päls bär kaninen.*

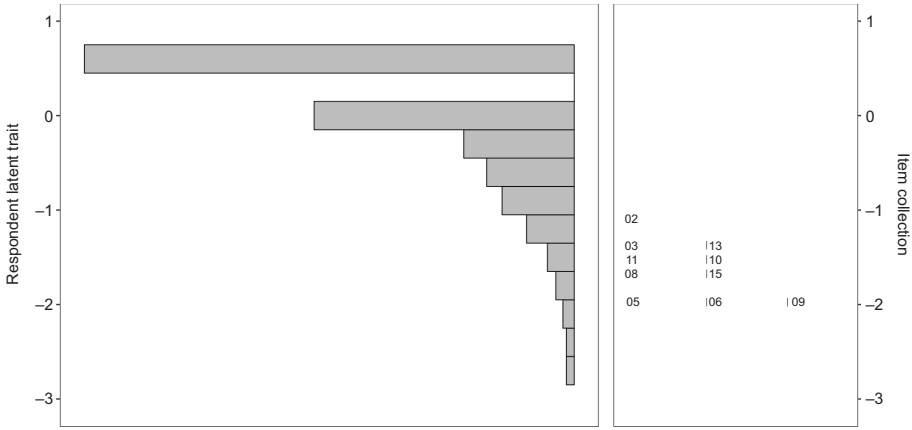
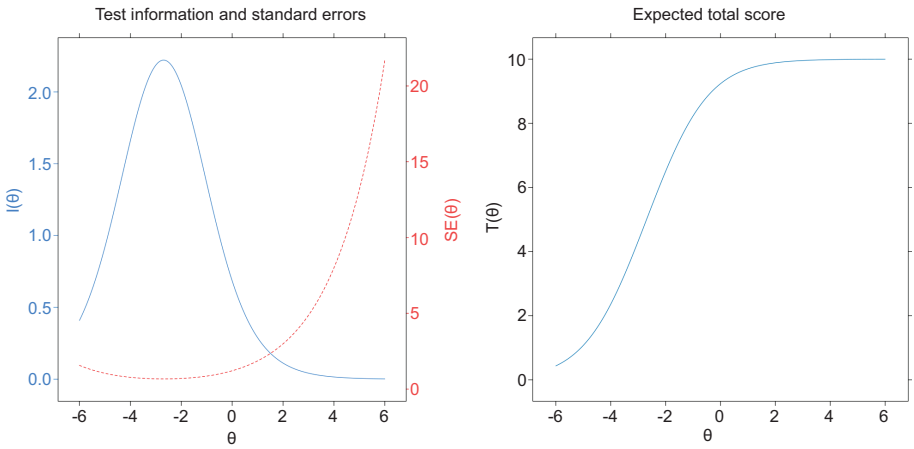


Figure 3. Test information (with SE) and expected total score with performance.

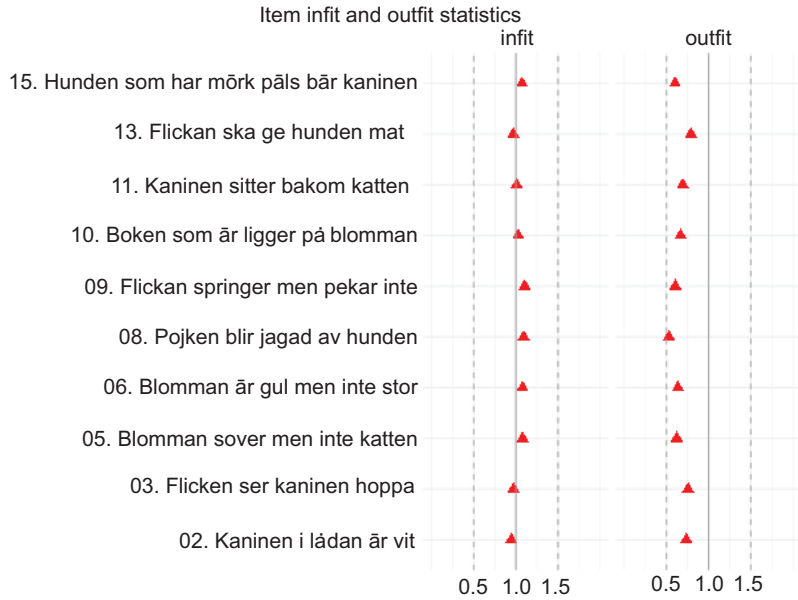


Fit diagnosis

The infit and outfit mean-square statistics are the Chi-square statistics divided by its degrees of freedom; this means that its expected value is close to 1. A common interpretation is that mean-square values between 0.5 and 1.5 are considered productive for measurement and values close to 1.0 indicate little or no distortion in the measurement (Wright *et al.*, 1994). GRA-10 infit and outfit values are within the accepted range, as shown in Figure 4. The outfit statistics show more variation in comparison to the infit value; this might depend on a large sample size (Müller, 2020) or that the study uses a 3PL model instead of a Rasch model. In general, letting

https://doi.org/10.58986/al.2025.34243

Figure 4. Infit and outfit statistics.



Note: Items with values within 0.5 and 1.5 are considered to be productive for measurement.

more variables vary in the model increases the risk for overfitting; however, the results show that this is not a concern.

Subgroup-analysis

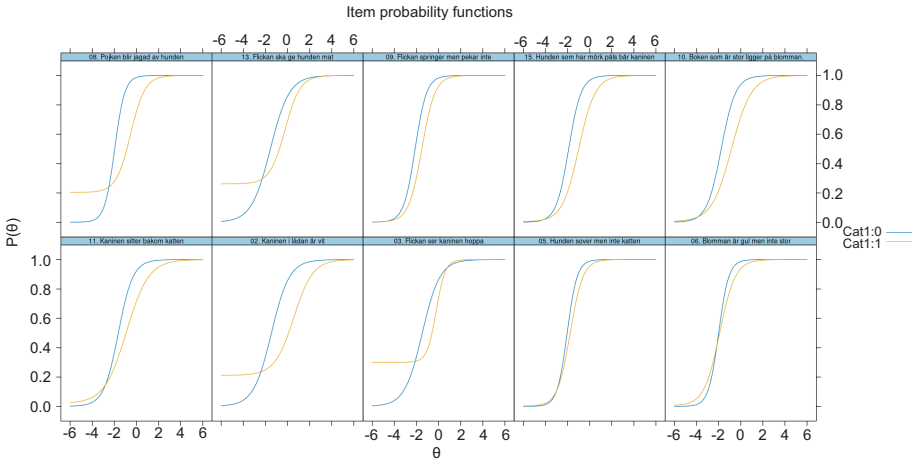
To examine whether the model detects any differences between L1 and L2 students, the 3PL model was used with curriculum as a category. As we see in Figure 5, for L1 students (blue line), all items discriminate well between students with/without grammatical comprehension difficulties. The yellow line represents L2 students and on four questions (02, 03, 08, and 13), the probability (P(Θ)) for a correct answer is higher in comparison to the L1 students, which means that L2 students guess on these four questions to a higher degree. In addition, the slope is not as steep for L2 as for L1 students, indicating that the test is slightly better at distinguishing between students with/without grammatical comprehension difficulties in the L1 group, compared to the L2 group. Lastly, we can see that the yellow slope is situated more to the right for all items, which again shows that these items are considered harder for L2 students in comparison to L1 students.

Discussion

The purpose of the digital test tool from the LegiLexi foundation is the early identification of students with difficulties in language and literacy to better support them in the classroom (www.legilexi.org). The subtest GRA was designed to find students in Swedish grade 0 with grammatical comprehension difficulties. Using the IRT

https://doi.org/10.58986/al.2025.34243

Figure 5. Subgroup analysis of L1 (n = 7,254) and L2 students (n = 991). *Blue line = L1, yellow line = L2.



framework (Hambleton *et al.*, 1991; Martinková and Hladká, 2023) and test scores from a 16-item version of LegiLexi GRA from more than 8,000 students in Swedish grade 0 (6–7 years old), we created a 10-item subset with the best possible psychometric qualities (new GRA-10). The 3PL model described the new GRA-10 data well without overfitting or underfitting the data, and analyses of internal consistency reached acceptable levels. The analyses further confirmed that the items in GRA-10 are easy for most students in this age group but have a good ability to discriminate between students with and without grammatical comprehension difficulties in both L1 and L2 groups.

Oral language comprehension difficulties in primary school-age children often go undetected (Adlof and Hogan, 2019), even though it is an important language skill that supports reading comprehension (Hoover and Tunmer, 2022). Most existing published screeners of oral language skills are individually administered (Bao *et al.*, 2024). LegiLexi GRA-10, a widely used whole-class screener in Swedish, could thus serve an important role in enabling early detection of students with grammatical comprehension difficulties to be able to give them the right support in the classroom. Previous studies, including whole-class screening of language comprehension, have had a different focus than GRA: identifying children who meet criteria for DLD (Adlof *et al.*, 2017; Hendricks *et al.*, 2019). Similar to the British LanguageScreen evaluated with the Rasch model by Hulme *et al.* (2024), the purpose of LegiLexi GRA is to identify students with language difficulties and signal to teachers that they may need extra support, not to try and predict a binary diagnostic outcome. Thus, diagnostic classification accuracy has not been investigated for the new GRA-10. As mentioned in the Introduction, Hallin *et al.* (2024) compared scores from the new GRA-10 with TROG-2 block scores in a sample of 55 students in grade 0, using a subsample of the same students as in Al Emara (2024), which investigated the external validity of the first version of GRA-10. Hallin *et al.* (2024) showed that the new

https://doi.org/10.58986/al.2025.34243

GRA-10 showed slightly improved identification of students with/without grammatical comprehension difficulties with TROG-2 results as a benchmark, compared to the first version of GRA-10 (Al Emara, 2024). Using cut-offs for GRA at 8 points or lower and TROG-2 at a standard score of 80 or lower, only 7.5% of the L1 students were misclassified as having grammatical difficulties by the new GRA-10 (false positives, compared to 9% in the previous GRA-10) while the rate of false negatives was low, and similar to the previous GRA-10 (2.5% vs. 3%). Similar to Al Emara (2024), Hallin *et al.* (2024) noted that a subgroup analysis on the small L2 sample (n = 13) based on TROG-2 standard scores was not suitable.

The IRT-model

We chose the 3PL model for item analyses instead of the more commonly used Rasch model. A multilevel model is always going to fit the data better than a restraint model, but we believe that because we have a big data set, the 3PL model can be justified. To obtain accurate estimations for a 3PL model, a sample size of 1,000 respondents is adequate (Yen, 1987; De Ayala, 2009), which is fulfilled with our L2 sample (992 respondents) and well above with our L1 sample (7,254 respondents). The 3PL model enabled us to go in-depth with sub-group analyses and investigate potential differences between L1 and L2 groups. The model revealed that some items had a higher guessing parameter in the L2 group (see below) but also revealed that the results for L1 students are very stable. These discrepancies would not have been detected if we used the more commonly used Rasch model.

Based on the 3PL-model, three items were exchanged in the new GRA-10, compared to the first version of old GRA-10. Surprisingly, two out of three items that were brought back were excluded in the old GRA-10 due to surface characteristics, for example preposition knowledge ("the rabbit is *behind…*" – arguably more vocabulary knowledge than grammar knowledge; see discussion in Nielsen *et al.*, 2024), and future tense ("*will* feed"). But given the item performance and characteristics, and the increase in internal consistency from 0.70 for L1 students and 0.66 for L2 students (old version of GRA-10) to 0.74 for L1 students and 0.69 for L2 students (new GRA-10), we are fairly confident that items in the new GRA-10 do measure the same latent construct. In the study done by Nielsen *et al.* (2024), including Danish speaking 10-year olds, TROG-2 scores did not correlate more strongly with vocabulary scores than a sentence repetition task and a "whodunnit" task, where the child had to identify the agent of the sentence with two alternatives. This went against the researchers' expectations because TROG-2 included items where knowledge of, for example, prepositions was tested. It is worth noting here that it is impossible to completely isolate grammatical comprehension skills from other linguistic skills (e.g. vocabulary) and non-linguistic skills (e.g. fluid reasoning and executive functioning needed to aid picture selection; Nielsen *et al.*, 2024). Similarly to many other grammatical comprehension tasks, the included nouns and verbs in GRA are chosen to be well known by children at an early age, and the inclusion of a pre-vocabulary test where all students have to answer at least 3 out of 5 questions correctly ensures that students who do not have these basic vocabulary skills do not proceed with GRA-10, because the results would be difficult to interpret.

Subgroup analysis

A considerable number of students in Swedish schools have Swedish as their L2, and in 2023–2024, 13.9% followed the curriculum for Swedish as a second language, similar to the proportion of 12% in the sample analysed in this study (Statistics Sweden 2024b). Even more students in Swedish schools are multilingual, with Swedish as one of their first languages: 28.9% of students in compulsory school in Sweden are entitled to lessons in a native language other than Swedish (Statistics Sweden 2024b). In the present study, multilingual students with Swedish as an L1 are included in the L1 sample because the only information we have about the students included in the dataset is which curriculum for Swedish they follow. The psychometric data of GRA-10 for monolingual and multilingual students with Swedish as an L1 showed good internal consistency and all items performed well in discriminating between students with/without grammatical comprehension difficulties. Furthermore, all test items were easy for the group taken as a whole. The picture was slightly different for L2 students, which can be expected because students in primary school who follow Swedish as a L2 curriculum form a very heterogeneous group, from those who have recently arrived in Sweden to those who have lived in Sweden for some years, but have not had regular or enough exposure to Swedish to be fully proficient. Our subgroup analyses showed that on four of the ten items, the L2 students had a higher probability of answering correctly while guessing than the L1 students. A higher guessing parameter could be due to less effective distractors (too easily neglected alternatives) or correct answers that are more intuitive. Because the guessing parameter P(Θ) did not show any signs of variance in the full model, we argue that heterogeneity in language knowledge in the L2 student group is the reason. This can also explain why the L2 students had higher standard deviations on the discrimination parameter, and slightly lower Cronbach's alpha than the L1 group. Importantly, because students need to answer nine out of ten items correctly to pass the GRA-10 screening, guessing correctly on four items won't result in a pass on the whole screening task, and we conclude that GRA-10 effectively finds students with grammatical comprehension difficulties in both L1 and L2 groups despite slight differences in the guessing parameter values.

Practical implications

Our results show that the majority of young children in the second semester of Swedish compulsory school can accurately and independently complete a 10-item grammatical listening comprehension task in the classroom setting with a forced-picture-choice response, with each sentence presented only once, when they have the linguistic knowledge to choose the correct answer. There are, of course, students who get a score of 8 or lower (the cut-off) due to other intrinsic or extrinsic factors than language difficulties, for example, concentration, a less-than-optimal test environment, forgetting instructions, tiredness, or accidental error responses, but the fact that most students do score 9 or 10 points on GRA-10, and the external validation using TROG-2 (Hallin *et al.*, 2024) does not indicate that this occurs to a large extent. The brief duration of the task (five practice vocabulary items and 10 test items) together with the digital and focused presentation also mitigates some of these risks. Importantly, a low

result on the new GRA-10 cannot (and should not) be interpreted as an indication of a disability, but rather as an indication of grammatical comprehension difficulties that warrants further investigation (similar to Hulme *et al.* 2024). Thus, students who do not reach 9–10 points on GRA-10 need to be followed up by their teacher/special education teacher or school SLP and be given appropriate support, especially if they also score low on other LegiLexi language comprehension subtests (vocabulary and listening comprehension). In addition, a passing result on GRA-10 (a score of 9 or 10) should be interpreted as an age-adequate ability in grammatical comprehension for grade 0 students, but this does not say anything about *how* good the student's grammatical comprehension is compared to their peers, which is important to remember when using the test. If GRA-10 is used for older students, it is also important to remember that the test has been created and validated for students in the first year of Swedish compulsory school (5- to 7-year olds), which means that an older student who scores 9–10 points on GRA-10 still can have difficulties in comprehending more complex syntactic structures not included in this task.

Conclusion

We conclude that the IRT-modeling and analyses of internal consistency described in the present study, together with the results from Hallin *et al.* (2024), strongly indicate that the new LegiLexi GRA-10 work well to identify students with grammatical comprehension difficulties in the first year of Swedish compulsory school, both for those who follow the Swedish curriculum and for those who follow the curriculum for Swedish as a second language. Because most published oral language screeners are administered individually (Bao *et al.*, 2024), it is promising that a group-administered measure achieves this, and an important step to better detect students with early oral language difficulties to give them appropriate support.

References

Adlof, SM and Hogan, TP. 2019. If we don't look, we won't see: Measuring language development to inform literacy instruction. *Policy Insights from the Behavioral and Brain Sciences* 6(2), 210–217. https://doi.org/10.1177/237273221983907

Adlof, SM, Scoggins, J, Brazendale, A, Babb, S and Petscher, Y. 2017. Identifying children at risk for language impairment or dyslexia with group-administered measures. *Journal of Speech, Language, and Hearing Research* 60(12), 3507–3522. https://doi.org/10.1044/2017_JSLHR-L-16-0473

Al Emara, A. 2023. *Grammatical Comprehension in Monolingual and Multilingual Children in Nursery Class: A Comparison Between TROG-2 and LegiLexi*. Degree project in speech and language pathology, Division of Speech and Language Pathology, Karolinska Institutet, CLINTEC, Stockholm, Sweden.

Bao, X, Komesidou, R and Hogan, TP. 2024. A review of screeners to identify risk of developmental language disorder. *American Journal of Speech-Language Pathology* 33(3), 1548–1571. https://doi.org/10.1044/2023_AJSLP-23-00286

Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability. In: Lord, F and Novick, N (eds.), *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley, pp. 397–479.

Bishop, DVM and Garsell, M. 2009. *Test for Reception of Grammar, Version 2: TROG-2 Manual (Swedish Version)*. Stockholm: Pearson Assessment.

Bishop, DV, Snowling, MJ, Thompson, PA, Greenhalgh, T and Consortium, C. 2016. CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS One* 11(7), e0158753. https://doi.org/10.1111/jcpp.12721

Chalmers, RP. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* 48, 1–29. https://doi.org/10.18637/jss.v048.i06

Christensen, KB, Makransky, G and Horton, M. 2017. Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement* 41(3), 178–194. https://doi.org/10.1177/0146621616677520

Cronin P, Reeve, R, McCabe, P, Viney, R and Goodall, S. 2020. Academic achievement and productivity losses associated with speech, language and communication needs. *International Journal of Language & Communication Disorders* 55(5), 734–750. https://doi.org/10.1111/1460-6984.12558

De Ayala, RJ. 2009. *The Theory and Practice of Item Response Theory.* New York, NY: Guilford Press.

Debelak, R, Stobl, C and Zeigenfuse, MD. 2022. *An introduction to the Rasch MODEL with examples in R* (1st ed.). Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315200620

Ebbels, SH, McCartney, E, Slonims, V, Dockrell, JE and Norbury, CF. 2019. Evidence-based pathways to intervention for children with language disorders. *International Journal of Language & Communication Disorders* 54(1), 3–19. https://doi.org/10.1111/1460-6984.12387

Hallin, AE, Al Emara, A and Kalmedal, A. 2024. *Utveckling och validering av ett screening test för grammatisk språkförståelse i förskoleklass (LegiLexi GRA)* (Development and validation of a screening test for grammatical comprehension in grade 0 [LegiLexi GRA]). Oral presentation, Nov. 14–15. Nationell Konferens i Logopedi, Göteborg, Sweden.

Hambleton, RK, Swaminathan, H and Rogers, HJ. 1991. *Fundamentals of Item Response Theory.* Washington DC: Sage.

Hendricks, AE, Adlof, SM, Alonzo, CN, Fox, AB and Hogan, TP. 2019. Identifying children at risk for developmental language disorder using a brief, whole-classroom screen. *Journal of Speech, Language, and Hearing Research* 62(4), 896–908. https://doi.org/10.1044/2018_JSLHR-L-18-0093

Holcombe, AO, Kovacs, M, Aust, F and Aczel, B. 2020. Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS ONE* 15(12), e0244611. https://doi.org/10.1371/journal.pone.0244611

Hoover, WA and Gough, PB. 1990. The simple view of reading. *Reading and Writing* 2, 127–160. https://doi.org/10.1007/BF00401799

Hoover, WA and Tunmer, WE. 2022. The primacy of science in communicating advances in the science of reading. *Reading Research Quarterly* 57(2), 399–408. https://doi.org/10.1002/rrq.446

Hulme, C, McGrane, J, Duta, M, West, G, Cripps, D, Dasgupta, A., ... and Snowling, M. 2024. LanguageScreen: The development, validation, and standardization of an automated language assessment app. *Language, Speech, and Hearing Services in Schools* 55(3), 904–917. https://doi.org/10.1044/2024_LSHSS-24-0000

Kaya, A. 2022. *The Relationship between Listening Comprehension, Vocabulary and Grammatical Comprehension in Swedish Students in Grades F-3.* Degree project in speech and language pathology, Division of Speech and Language Pathology, Karolinska Institutet, CLINTEC, Stockholm, Sweden.

Martinková, P and Hladká, A. 2023. *Computational Aspects of Psychometric Methods: With R* (1 Uppl.). Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781003054313

Müller, M. 2020. Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications* 7(1), 5. https://doi.org/10.1186/s40488-020-00108-7

Navarro, D and Foxcroft, D. 2025. *Learning Statistics with Jamovi: A Tutorial for Beginners in Statistical Analysis*. Cambridge: Open Book Publishers. https://doi.org/10.11647/OBP.0333

Nielsen, JL, Christensen, RV and Poulsen, M. 2024. What's format got to do with it? A comparison of three syntactic comprehension measures. *Journal of Research in Reading* 47(1), 1–19. https://doi.org/10.1111/1467-9817.12438

Norbury, CF, Gooch, D, Wray, C, Baird, G, Charman, T, Simonoff, E, ... and Pickles, A. 2016. The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry* 57(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Nordström, T, Fälth, L and Danielsson, H. 2025. Evaluating the simple view of reading model: Longitudinal testing and applicability to the Swedish language. *Education Sciences* 15(3), 260. https://doi.org/10.3390/educsci15030260

Organization for Economic Cooperation and Development (OECD). 2023. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. PISA. Paris: OECD Publishing. https://doi.org/10.1787/53f23881-en

Poulsen, M, Nielsen, JL and Christensen, RV. 2022. Remembering sentences is not all about memory: Convergent and discriminant validity of syntactic knowledge and its relationship with reading comprehension. *Journal of Child Language* 49(2), 349–365. https://doi.org/10.1017/S0305000921000210

Rasch, G. 1960. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielsen & Lydiche, pp. xiii, 184.

Reuterskiöld, C, Hallin, AE, Nair, VK and Hansson, K. 2021. Morphosyntactic challenges for Swedish-speaking children with developmental language disorder in comparison with L1 and L2 peers. *Applied Linguistics* 42(4), 720–739. https://doi.org/10.1093/applin/amaa058

Statistics Sweden. 2024a. *Nationella prov I årskurs 3, Tabell 7: Fördelning i procent över antal delprov där eleverna uppnått kravnivån för de elever som deltagit i alla del prov i ämnet, läsåret 2022/23.* https://siris.skolverket.se/siris/sitevision_doc.getFile?p_id=552759

Statistics Sweden. 2024b. *Elever med undervisning i modersmål. Tabell 8A: Elever i grundskola med undervisning i modersmål och svenska som andraspråk (SVA) läsåren 2021/22–2023/24.* https://siris.skolverket.se/siris/sitevision_doc.getFile?p_id=553088

Taber, KS. 2018. The Use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education* 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Wright, BD, Linacre, JM, Gustafson J-E and Martin-Löf, P. 1994. Reasonable mean square fit values. *Rasch Measurement Transactions* 8(3), 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm. Accessed: February 25, 2025.

Yen, W. 1987. A comparison of the efficiency and accuracy of bilog and logist. *Psychometrika* 52 (2), 275–291. https://doi.org/10.1007/BF02294241

Zhang, R, Pek, J, Flake, JK and Chalmers, RP. 2023. Using the triad task as a measure of thought style: A validation study. PsyArXiv; 2023. PPR: PPR618550. https://doi.org/10.31234/osf.io/6jwd8

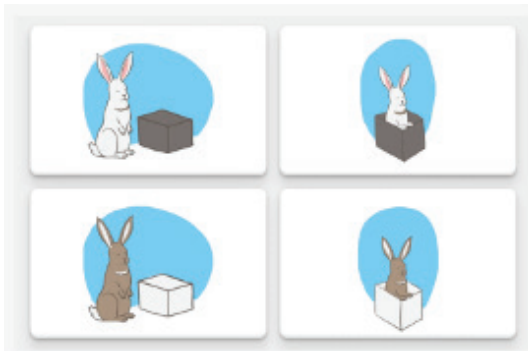
Appendix A

Items in GRA-16, old version of GRA-10, and new GRA-10, and an item example

Below are all sentences included in GRA-16 with translations. Items included in the new GRA-10 are in bold font, and bold and underlined items (n = 3) were not included in the previous version of GRA-10. Italicised items (n = 3) were included in the old GRA-10. GRA-10 structures marked with an asterisk overlaps with structures found in Swedish TROG-2 (item #9 is identical to an item in Swedish TROG-2; Bishop and Garsell, 2009).

01. Flickan får smaka på hans glass	*The girl is tasting his ice cream*
***02. Kaninen i lådan är vit**	**The rabbit in the box is white**
03. Flickan ser kaninen hoppa	**The girl sees the jumping rabbit**
04. *Här är pojken som inte klättrar.*	*Here is the boy that doesn't climb*
***05. Hunden sover men inte katten**	**The dog sleeps but not the cat**
***06. Blomman är gul men inte stor**	**The flower is yellow but not big**
07. *Katten jagar hunden och hoppar*	*The cat chases the dog and jumps*
***08. Pojken blir jagad av hunden**	**The boy is chased by the dog**
***09. Flickan springer men pekar inte**	**The girl runs but does not point**
***10. Boken som är stor ligger på blomman.**	**The book that is big is on the flower**
11. Kaninen sitter bakom katten	**The rabbit sits behind the cat**
12. Pojkens paket är litet	The boy's gift is small
13. Flickan ska ge hunden mat	**The girl will feed the dog**
14. Hunden som lyfter kaninen är stor	The dog lifting the rabbit is big
***15. Hunden som har mörk päls bär kaninen**	**The dog which has dark fur carries the rabbit**
16. Katten jagar hunden som hoppar	The cat chases the dog which jumps

Example of the picture choices for sentence #2 *"The rabbit in the box is white"* © *LegiLexi, 2024.*



https://doi.org/10.58986/al.2025.34243

Appendix B

Statistical equations

The Rasch model (Rasch, 1960; Debelak *et al.*, 2022)

$$\Pr(U_{pi} = 1 \mid \theta_p, \beta_i) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}.$$

U = 1 – means a correct response.
θ is the person's ability.
β shifts the curve horizontally (difficulty).

The three-parameter logistic (3PL) model (Birnbaum, 1968)

$$\Pr(U_{pi} = 1 \mid \theta_p, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1 - \gamma_i) = \frac{\exp\{\alpha_i \cdot (\theta_p - \beta_i)\}}{1 + \exp\{\alpha_i \cdot (\theta_p - \beta_i)\}}$$

U = 1 – means a correct response.
θ is the person's ability.
α controls the slope of the curve (discrimination).
β shifts the curve horizontally (difficulty).
y is the lower asymptote (guessing probability).

Model comparison

Model	AIC	SABIC	HQ	BIC	logLik	X2	df	P
Rasch	55,197	55,239	55,223	55,274	-27,587			
3PL	54,974	55,089	55,046	55,184	-27,457	260	19	0

Appendix C

Item difficulty (using CTT), standard deviation, and Cronbach's alpha if the item is removed for the initial GRA-16

#	Item	Difficulty	SD	Alpha
1.	**Flickan får smaka på hans glass**	**0.761**	**0.426**	**0.745**
2.	Kaninen i lådan är vit	0.768	0.422	0.740
3.	Flickan ser kaninen hoppa	0.798	0.402	0.740
4.	**Här är pojken som inte klättrar**	**0.870**	**0.336**	**0.742**
5.	Hunden sover men inte katten	0.937	0.243	0.741
6.	Blomman är gul men inte stor	0.933	0.250	0.741
7.	**Katten jagar hunden och hoppar**	**0.762**	**0.426**	**0.752**
8.	Pojken blir jagad av hunden	0.910	0.287	0.737
9.	Flickan springer men pekar inte	0.930	0.256	0.742
10.	Boken som är stor ligger på blomman.	0.856	0.351	0.737
11.	Kaninen sitter bakom katten	0.839	0.367	0.737
12.	**Pojkens paket är litet**	**0.650**	**0.477**	**0.754**
13.	Flickan ska ge hunden mat	0.795	0.404	0.742
14.	**Hunden som lyfter kaninen är stor**	**0.691**	**0.462**	**0.749**
15.	Hunden som har mörk päls bär kaninen	0.890	0.312	0.738
16.	**Katten jagar hunden som hoppar**	**0.412**	**0.492**	**0.767**

*Items in bold font were removed in GRA-10.

https://doi.org/10.58986/al.2025.34243