Construct validity of a process-oriented test assessing syntactic skills in German primary school children

Manuscript accepted for publication in the journal *Language Assessment Quarterly: An International Journal*

Julia Schindler and Tobias Richter
University of Würzburg

Maj-Britt Isberner
University of Kassel

Johannes Naumann
Goethe University Frankfurt

Yvonne Neeb
German Institute for International Educational Research

Julia Schindler
University of Würzburg, Department of Psychology
Röntgenring 10
97070 Würzburg, Germany
mail: julia.schindler@uni-wuerzburg.de
phone: +49 (0) 931 31-83879
fax: +49 (0) 931 31-84891

Abstract

Reading comprehension is based on the efficient accomplishment of several cognitive processes at the word, sentence, and text level. To the extent that each of these processes contributes to reading comprehension, it can cause reading difficulties if it is deficient. To identify individual sources of reading difficulties, tools are required that allow for a reliable and valid assessment of individual differences in specific cognitive processes of reading comprehension. The present study demonstrates the usefulness of this process-oriented approach to assessing reading comprehension skills using the example of a test for assessing syntactic skills in German primary school children. The test comprises a grammaticality-judgment task that contains items with carefully varied features which are known to facilitate or impede syntactic processes. By means of explanatory item-response models, we demonstrate that empirical item difficulties vary as a function of experimentally varied item features, indicating that the test indeed assesses syntactic skills. Moreover, the test measures individual differences in syntactic skills across the whole range of skill levels and is sensitive to developmental changes. We conclude that the test is a valid tool to assess individual differences and to detect deficits in this component process of reading comprehension.

Keywords: explanatory item-response model, grammaticality judgment task, reading comprehension, syntactic skills

Acknowledgments

The data presented in this article were collected as part of the project *Process-based assessment of reading and listening skills in primary school children*, which was funded by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF, grants 01 GJ 0985, 01 GJ 0986, 01 GJ 1206A, and 01 GJ 1206B). We would like to thank David Nitz for programming assistance and several student assistants for their help in constructing stimulus materials and collecting data. Researchers who are interested in the material used in the visual and auditory grammaticality judgment tasks are invited to send an e-mail to the first or the second author.

From a cognitive perspective, reading comprehension may be regarded as the outcome of efficient mastery of cognitive component processes at the word, sentence, and text level (Perfetti, Landi, & Oakhill, 2005; Schindler & Richter, in press; for a critical discussion see Alderson, 2000). These component processes of reading interact in various ways and make unique contributions to reading comprehension. Accordingly, each single component process is a potential cause for reading difficulties when it is deficient in an individual reader (Vellutino, Fletcher, Snowling, & Scanlon, 2004). Therefore, process-oriented reading tests are needed that address each of the major cognitive component skills of reading comprehension separately in order to construct successful and target-oriented reading interventions and remediation programs.

However, most psychological tests in German and English do not assess individual differences in cognitive component processes of reading comprehension in an adequate way. They neglect the complex structure of cognitive component processes of reading comprehension, they neglect processing time as indicator of the degree of their automatization, or they involve additional cognitive skills unrelated to reading. Most importantly, each cognitive component skill must be assessed by well-defined process-specific tasks with carefully constructed, theoretically-based test items to allow for a meaningful interpretation of individual test scores (Hartig & Frey, 2012; Hartig, Frey, Nold, & Klieme, 2012).

The aim of this study was to illustrate the feasibility of a process-oriented approach to reading comprehension by using the example of a carefully designed grammaticality judgment task for assessing syntactic skills of German 3rd and 4th graders. The test is part of a more comprehensive test battery with six subtests, each addressing specific cognitive component skills of reading comprehension at the word, sentence, and text level selectively (Richter, Naumann, Isberner, Neeb, & Knoepke, 2017). The core of the paper are results from explanatory item response models examining the construct validity of the grammaticality judgment task, i.e. the

contribution of the targeted syntactic skills for solving the test items. In the following, we will first discuss the limitations of existing diagnostic approaches to reading comprehension and, then, turn to the grammaticality judgment task as a concrete example of a test that assesses one cognitive component skill of reading comprehension in a process-oriented manner and by use of carefully constructed, theoretically-based test items with systematically varied item features.

Reading comprehension is based on efficient cognitive processes

Successful comprehension of written texts is based on the efficient accomplishment of several cognitive component processes at the word, sentence, and text level (Perfetti et al., 2005). On the word level, the reader has to identify written word forms indirectly by *phonological recoding* of written words or directly by comparing written word forms to orthographical representations in the mental lexicon (*orthographical decoding*) (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). In addition, word meanings have to be retrieved from the mental lexicon (*access to word meanings*). On the sentence level, readers need to derive a coherent sentence meaning (*semantic analysis*) in consideration of the sentence's grammatical structure (*syntactic analysis*) (e.g. Müller & Richter, 2014; Pylkkänen & McElree, 2006). Although syntactic and semantic analysis are both necessary for successful sentence comprehension and interact in several ways to derive a coherent sentence interpretation (e.g., Rayner, Carlson, & Frazier, 1983; Taraban & McClelland, 1988), evidence suggests that they are psychometrically discernible cognitive processes (Carpenter, Miyake, & Just, 1995; Richter, Isberner, Naumann, & Kutzner, 2012). On the text and discourse level, readers need to construct and continuously update a situation model (mental model) of the text content by *establishing local and global coherence relations* between adjacent and distant statements and by integrating text information with prior knowledge (Johnson-Laird, 1981; van Dijk & Kintsch, 1983). It is noteworthy that many approaches to reading competence are based on a coarser distinction of component skills or focus on global text

comprehension processes such as conscious text reflection and interpretation (e.g., Sabatini et al., 2013; for a critical discussion see Alderson, 2000). Unlike these approaches, the cognitive view adopted here focuses on the aforementioned basal cognitive processes at the word, sentence, and text level (e.g., Müller & Richter, 2014; Schindler & Richter, in press).

Each of these cognitive processes makes a unique contribution to reading comprehension. Consequently, individual differences in one or more of these cognitive processes result in individual differences in reading comprehension (Schindler & Richter, in press). In order to create successful and adaptive intervention programs, it is important to determine the exact source and severity of individual reading deficits. However, most available tests assessing reading comprehension skills in German are subject to several restrictions (which apply to tests in English as well).

First, most of the available reading tests either focus on visual word recognition skills (e.g. SLRT-II, Moll & Landerl, 2010; WLLP-R, Schneider, Blanke, Faust, & Küspert, 2011) or on global product-oriented measures of reading comprehension (e.g. FLVT 5-6, Souvignier, Trenk-Hinterberger, Adam-Schwebe, & Gold, 2008; HLP 1-4, May & Arntzen, 2000; SLS 2–9, Wimmer & Mayringer, 2014; VSL, Walter, 2013), thereby neglecting the complex structure of cognitive component processes of reading comprehension at the word, sentence, and text level. A second restriction concerns the use of tasks which involve cognitive skills that are not specific to reading comprehension, such as finding pictures that match target words (e.g. ELFE 1-6, Lenhard & Schneider, 2005; HAMLET 3-4, Lehmann, Peek, & Poerschke, 1997; WLLP-R, Schneider, Blanke, Faust, & Küspert, 2011) or selecting a correct answer among several alternatives in a single-choice task (e.g. ELFE 1-6, Lenhard & Schneider, 2005; FLVT 5-6, Souvignier et al., 2008; LESEN 6-7, Bäuerlein, Lenhard, & Schneider, 2012; VSL, Walter, 2013). Their diagnostic value,

notwithstanding, these measures of reading ability might be influenced, at least to a certain degree, by individual differences in reading-unrelated cognitive skills.

Finally, reading comprehension skills are often operationalized as the number of successfully solved test items (e.g. HAMLET 3-4, Lehmann et al., 1997; HLP 1-4, May & Arntzen 2000; VSL, Walter, 2013). While this approach may be informative with regard to the quality of the mental representations constructed during reading, it does not tell us much about the efficiency or degree of routinization of the cognitive processes involved (Perfetti, 1985). This component of efficiency is captured in the time it takes to carry out reading-related tasks. Despite the fact that time-based measures are a common indicator of automaticity of cognitive processes in experimental studies on reading in L1 and L2 and on L2 acquisition (e.g. Hulstijn, 2015; Hulstijn, van Gelderen, & Schoonen, 2009; Lim & Godfroid, 2015; Trapman, van Gelderen, van Steensel, van Schooten, & Hulstijn, 2014), such measures are very rarely used in standardized tests for the assessment of reading comprehension skills (e.g. ELFE 1-6, Lenhard & Schneider, 2005), with the exception of speed tests such as one minute of reading (e.g. Deno, 1985; SLRT-II, Moll & Landerl, 2010; SLS 2–9, Wimmer & Mayringer, 2014). However, speed tests usually focus on the fluency of single-word reading or reading of simple sentences and do not assess individual differences in specific cognitive component skills of reading comprehension.

Assessing syntactic skills in German

Syntactic skills are a case in point to illustrate the feasibility of a process-based approach to assess individual differences in component skills of reading. Syntactic skills comprise the ability to reliably and efficiently access the syntactic knowledge of a specific language (such as knowledge of word order constraints and inflectional morphology) and derive an appropriate syntactic structure when processing spoken or written sentences (see e.g. Müller & Richter, 2014; Schindler & Richter, in press). Several studies suggest a strong relationship between syntactic skills and

reading comprehension (Byrne, 1981; Casalis & Louis-Alexandre, 2000; Nation & Snowling, 2000; Plaza & Cohen, 2003; Stothard & Hulme, 1992; Tunmer et al., 1987; for a critical discussion on the relationship between syntactic skills and reading comprehension, see Oakhill & Cain, 1997). They indicate that individual differences in syntactic skills account for a unique portion of variance in reading comprehension skills and imply that poorly developed or deficient syntactic skills can contribute to poor reading comprehension. To detect such a deficit, a psychological test is required that reliably and selectively measures individual differences in syntactic skills.

In German, several tests assessing dimensions of language development in children include a subtest that assesses syntactic skills. However, these tests differ from the instrument examined here in several ways. First, the majority of the available tests focus on language production skills or involve language production as part of the task, preventing selective assessment of comprehension skills (e.g. P-ITPA, Esser, Wyschkon, Ballaschk, & Hänsch, 2010; PET, Angermaier, 1977; SET 5-10, Petermann, Metz, & Fröhlich, 2010). Second, most of the available tests involve cognitive skills that are not specific to language processing such as picture identification, the memorization of multiple response alternatives, acting out sentence contents etc. (e.g. HSET, Grimm & Schöler, 2001; MSVK, Elben & Lohaus, 2001; TROG-D, Fox, 2011). Third, to our knowledge, none of these tests includes measures of processing latencies (e.g. ADST, Steinert, 2011; P-ITPA, Esser et al., 2010; SET 5-10, Petermann et al., 2010; TROG-D, Fox, 2011) which are necessary to assess the efficiency of syntactic processes (see Jeon & Yamashita, 2014 for further discussion). Finally, as far as we know, there is no German test that assesses syntactic skills in both spoken and written language processing by means of strictly parallel tasks and materials, which allows a direct comparison of syntactic skills in both modalities.

In the following, we will present a grammaticality judgment task for German primary school children that assesses syntactic skills in auditory and written language processing in a

process-oriented fashion. A distinctive feature of this test is its systematic and theoretically-based item construction which is a necessary precondition for assessing individual differences in syntactic skills selectively and for being able to interpret individual test scores in a meaningful way. Explanatory item response models were estimated to examine whether systematically varied item features that, according to psycholinguistic research, affect the ease of syntactic processes actually predict the empirical difficulties of the test items (Wilson & De Boeck, 2004). If the item features predicted empirical item difficulties this would not only show that the task actually assesses individual differences in syntactic skills, but it would also highlight the necessity to base item construction on a sound theoretical background to design a valid test in terms of construct validity (Hartig & Frey, 2012).

Assessing syntactic skills: Principles of test construction

To assess syntactic skills in primary school children, we constructed a visual grammaticality judgment task for children in Grades 3 and 4 and an auditory grammaticality judgment task which can be used with children from Grade 1 to 4. The visual grammaticality judgment task is part of a more comprehensive German-language test battery that consists of six subtests, each of them assessing a cognitive component process of reading comprehension at the word, sentence, or text level selectively (ProDi-L: Prozessbezogene Diagnostik des Leseverstehens bei Grundschulkindern [Process-based assessment of reading skills in primary school children], Richter et al., 2017). The auditory grammaticality judgment task is part of a strictly parallel German-language test battery for measuring the corresponding component processes of listening comprehension (ProDi-H: Prozessbezogene Diagnostik von Hörverstehensfähigkeiten bei Grundschulkindern [Process-based assessment of listening skills in primary school children]). In the grammaticality judgment tasks, children are presented with written (ProDi-L) and spoken (ProDi-H) grammatical sentences such as (1) and ungrammatical sentences such as (2):

(1) *Die Arbeiter bauen das Haus* / The workmen build the house.

(2) **Die Hexe auf einem Besen reitet* / * The witch on a broom rides.

By pressing one of two response keys children indicate whether the spoken and written sentences are grammatically correct or not. Both response accuracy (whether the sentence was correctly judged as grammatical or ungrammatical) and response latency (time from the beginning of sentence presentation to pressing a response button) are recorded. Thus, the test assesses not only whether grammatical knowledge can be retrieved reliably, but also how efficiently this knowledge can be retrieved and applied to derive a syntactic structure for each sentence. Moreover, both grammaticality judgment tasks involve solely skills that are specific to language processing.

Ability tests should include items of varying difficulty that allow selective measurements on different levels of ability. However, to interpret individual test scores of the grammaticality judgment task in a meaningful and process-oriented way, item difficulties must reflect the demands different items make on syntactic processing. In other words, easy items should be easy because they make low demands on syntactic processing, whereas difficult items should be difficult because they make high demands on syntactic processing. A straightforward and theoretically sound way to create items of varying difficulty is to vary specifically those item features that – according to psycholinguistic theory and research – may be expected to facilitate or impede syntactic processing.

One of the features varied in the present grammaticality judgment task to generate items of different difficulty is *syntactic complexity* (operationalized, for example, as the number of syntactic nodes or phrases or embedded vs. non-embedded sentence structures), which has been found to be associated with difficulties in sentence and text comprehension. Earlier studies demonstrated that reading times for whole sentences (Graesser, Hoffmann, & Clark, 1980) or single words (Ferreira, Henderson, Anes, Weeks, & McFarlane, 1996) increased and response accuracy to comprehension questions decreased with increasing syntactic complexity, indicating increased processing

difficulty (Marton, Schwartz, & Braun, 2005). Accordingly, error rates and response latencies in our grammaticality judgment task are expected to increase with syntactic complexity (here operationalized as the number of syntactic phrases) of the test items.

Another feature relevant for the difficulty of grammaticality judgments is the *grammaticality* (grammatically correct vs. grammatically incorrect) of sentences themselves. Participants showed slower processing times (Flores d'Arcais, 1982, Exp. 4) and slower detection of words (Baum, 1991) in ungrammatical than in grammatical contexts and made more errors in a grammaticality judgment task when judging ungrammatical as compared to grammatical sentences (Friederici, Hahne, & Saddy, 2002). On account of this, we expected ungrammatical items to be more difficult than grammatical items, resulting in longer response latencies and lower response accuracy for ungrammatical items.

Finally, the *type of grammatical violation* was varied systematically to create items of varying difficulty. There are several grammatical markers or *cues* (e.g. Bates & MacWhinney, 1989) such as word order or case marking that help a reader or listener to comprehend a sentence. Recent findings suggest that German children exhibit more difficulties in using case-marking information than word order information to assign thematic roles in sentence comprehension (Dittmar, Abbot-Smith, Lieven, & Tomasello, 2008; Schipke, Knoll, Friederici, & Oberecker, 2012). Against this background, we expected detecting case-marking violations such as (4) to be more difficult for primary school children than detecting word-order violations such as (2), here repeated as (3). Moreover, the test included sentences with violations of verb-tense form such as (5), which consist of a finite auxiliary (*hat* /has) in V2-position and another finite simple past tense verb form (e.g. *schrieb*) in sentence final position instead of the required infinite past participle form (*ge-schrieb-en* / written). Thus, these sentences actually contain two types of violation, a violation of word order (the finite simple past tense verb form occurs in sentence final position)

and a morphological violation (the verb form in sentence final position is incorrectly conjugated). This might enable even faster and more accurate recognition of tense-form violations as compared to violations of word order or case marking.

(3) **Die Hexe auf einem Besen reitet* / * The witch on a broom rides

(4) **Die Schafe fressen dem*-DAT *Gras* / *The sheep eat the-DAT grass

(5) **Lisa hat einen Brief schrieb* / *Lisa has wrote a letter

All items of the visual and the auditory grammaticality judgment task were systematically varied with respect to syntactic complexity, grammaticality, and (for the ungrammatical sentences) type of violation. A grammatically correct sentence of low complexity, for example, should be least difficult, whereas an ungrammatical sentence of high complexity with a case-marking violation should be most difficult. Consequently, children with low syntactic skills should be able to make fast and accurate judgments on easy sentences but not on difficult ones, whereas children with high syntactic skills should be able to make fast and accurate judgments on both easy and difficult items. If the items of the grammaticality judgment tasks are well constructed and indeed assess individual differences in syntactic skills, these item features should have a measurable effect on empirical item difficulties (De Boeck & Wilson, 2004). In other words, empirical item difficulties should be predictable from the experimentally varied item features. The core of the present study was to demonstrate that the grammaticality judgment task with its carefully designed and theoretically-based test items actually assesses individual differences in syntactic skills. **Method**

Participants

Participants were 1,380 primary school children (678 boys and 658 girls, gender information for 44 children was missing) recruited from 26 schools (95 classes) in Cologne, Kassel, and Frankfurt am Main (Germany). One-thousand-one-hundred-and-eighteen children (548 boys,

531 girls, for 39 children gender information was missing) from Grades 1 to 4 completed the auditory version of the grammaticality judgment task, and 691 children (330 boys, 354 girls, for 7 children gender information was missing) from Grades 3 and 4 completed the visual grammaticality judgment task. Of these children, 429 children in Grades 3 and 4 participated in both the auditory and the visual grammaticality judgment task. Characteristics of the study sample are provided in Table 1. Socio-demographic data such as age and sex of the children as well as information about the parents' graduation and academic achievement (which were unrelated to the purposes of the present study) were collected via a parent questionnaire and supplemented by information from a teacher questionnaire when information from the parent questionnaire was missing. Only children with written parent consent participated in the study.

---------- Table 1 about here ----------

Instruments

The visual grammaticality judgment task contained 38 written sentences and the auditory grammaticality judgment task contained 38 spoken sentences. Two additional practice sentences preceded the test sentences in each task as a means of familiarizing students with the task. Both practice sentences were repeated until they were answered correctly. These items were excluded from the analysis. Half of the test sentences in each version (19 sentences) were grammatically well-formed German sentences such as (1), whereas the other half contained a grammatical violation. Of these 19 ungrammatical sentences, 10 sentences contained violations of word order such as (3), 4 sentences contained violations of case marking such as (4), and 5 sentences contained violations of the verb-tense form such as (5). Written and spoken sentences were comparable with respect to length (number of characters, spoken sentences: $M = 37.97$; $SD = 12.01$; $Min = 17$; $Max = 59$; written sentences: $M = 37.82$; $SD = 11.60$; $Min = 21$; $Max = 56$), syntactic complexity (number of syntactic phrases, spoken sentences: $M = 4.16$; $SD = 1.37$; $Min = 2$; $Max = 7$; written

sentences: *M* = 4.21; *SD* = 1.42; *Min* = 2; *Max* = 7), and number and types of violations. For ungrammatical sentences, which contained one of the three types of violation, the relative position of the violation within the sentences (mostly towards the end of the sentence) was comparable in the parallel versions. For both indicators of reliability and efficiency of syntactic processes, i.e. accuracy and response latencies, high reliabilities have been obtained in an earlier study for the visual grammaticality judgment task (accuracy: Cronbach's α = .91, response latencies: Cronbach's α = .98, Richter et al., 2012) and the auditory grammaticality judgment task (accuracy: Cronbach's α = .79, response latencies: Cronbach's α = .93, Richter & Naumann, 2012). Half of the spoken sentences were recorded by a male and the other half by a female speaker.

Administration Procedures

Both grammaticality judgment tasks were part of a cross-sectional study investigating the processes of listening and reading comprehension with various measures on the word, sentence, and text level (Richter, Naumann, Isberner, & Neeb, 2013; Richter et al., 2012). The children were tested together in classrooms of the participating schools. The grammaticality judgment task was presented on notebook computers, embedded in a story of an extraterrestrial named Reli, who asked the children for their help in learning the earthlings' language by telling him when he did something wrong. Reli introduced the tasks in short animated video clips and walked the children through the tasks. In the auditory version of the grammaticality judgment task, children listened to the two practice items and the test sentences via headphones. In the visual version of the task, the sentences were presented in the center of the screen, one at a time (font: Verdana, visual angle: 1.5 degrees). Each child received all 38 test sentences of a version in randomized order. For each sentence, children were asked to judge whether it was grammatically correct or not by pressing one of two response buttons (green button on the keyboard for *yes, the sentence is correct* or a red button for

no, the sentence is not correct). Prior to the presentation of test items, the two practice sentences were presented, for which children received feedback from Reli. When they gave an incorrect response, the practice sentences were repeated. Log-transformed response latencies (measured from stimulus onset to the press of a response button) and response accuracy were recorded as dependent variables.

Construct validity of the grammaticality judgment task

The core of the present study was to demonstrate that the grammaticality judgment task with its carefully designed and theoretically-based test items actually assesses individual differences in syntactic skills (construct validity). This aim was accomplished in three steps. In Step 1, empirical item difficulties were estimated using a one-parameter logistic model (1PL or Rasch model) for the accuracy data and a Rasch-analogous model (van Breukelen, 2005) for the (log-)transformed response latencies. In Step 2, item difficulties were modeled as a function of the varied item features (syntactic complexity, grammaticality, and, for the ungrammatical sentences, type of violation) by estimating explanatory item response models, i.e. a logistic multilevel regression model was estimated for accuracy and a linear multilevel regression model for the (log-transformed) response latencies (e.g. Hartig et al., 2012) with item features as predictor variables. In Step 3, these item difficulties, which were predicted from the item features, were then correlated with the empirical item difficulties to obtain a test of construct validity. In the following, the three steps of analysis and their underlying logic are explained in more detail.

Step 1. The Rasch model for logit-transformed response accuracy (i.e. log-transformed odds ratios for correct responses) as dependent variable with items as dummy-coded fixed effects and subjects as random effect was as follows:

(1) $\text{logit}(accuracy_{ij}) = b_{0j} + b_{1j} X_{1ij} + b_{2j} X_{2ij} + ... + b_{(k-1)j} X_{(k-1)ij} + r_{ij}$.

$b_{0j} = g_{00} + u_{0j}$ *Random coefficient*, u_{0j}: person parameter

$b_{1j} = g_{10}$

...

$b_{(k-1)j} = g_{(k-1)0}$ } *Fixed coefficients*: item parameters

The one item that is not represented by a dummy-coded fixed effect is called the reference item. The empirical item difficulties can be computed from this model as the sum of the intercept of the model g_{00} and the coefficients $b_{(k-1)j}$ for each specific item $X_{(k-1)ij}$ (the item difficulty of the reference item is captured by the intercept itself). An analogous model was estimated for log-transformed response latencies as dependent variable (van Breukelen, 2005).

Step 2. A Logistic Linear Test Model (LLTM, Fischer, 1974; with a logit link function for response accuracy, Dixon, 2008) with the theoretically derived item features as fixed item-level predictors and a random effect of subjects was estimated. Syntactic complexity (number of syntactic phrases) was included as grand-mean centered predictor variable into the LLTMs and the LLT-analogous models. Violation of word order (no violation = 0, word-order violation = 1), violation of tense form (no violation = 0, tense-form violation = 1), and violation of case marking (no violation = 0, case-marking violation = 1) were included as dummy-coded predictor variables. Grammaticality is captured in the dummy-coded predictor variables of violations with the grammatically correct sentences being the reference category):

(2) $\text{logit}(accuracy_{ij}) = b_{0j} + b_{1j}\,(feature_1)_{ij} + b_{2j}\,(feature_2)_{ij} + \ldots + b_{qj}\,(feature_q)_{ij} + r_{ij}$.

$b_{0j} = g_{00} + u_{0j}$ *Random coefficient*, u_{0j}: person parameter

$b_{1j} = g_{10}$

...

$b_{(k-1)j} = g_{(k-1)0}$ } *Fixed coefficients*: effects of item features

Again, an analogous linear multilevel model was estimated for the (log-transformed) response latencies as dependent variable (van Breukelen, 2005). Based on the intercept g_{00} of the

LLTM and the LLT-analogous model, the specific feature characteristics of each item, and the regression coefficients for these features $b_{qj}(feature_q)_{ij}$, predicted item difficulties were computed. If, for example, a sentence contains no violation at all, its predicted item difficulty would be $g_{00} + b_{qj}(syntactic\ complexity_q)_{ij}$. If a sentence contains a word order violation, its predicted item difficulty would be $g_{00} + b_{qj}(syntactic\ complexity)_{ij} + b_{qj}(word\ order\ violation)_{ij}$.

Step 3. A substantial correlation between the empirical item difficulties from the Rasch-model (or the Rasch-analogous model) from Step 1 and the predicted item difficulties from the LLTM (or the LLT-analogous model) from Step 2 would show that the systematic variation of syntactic complexity, grammaticality, and type of violation affects empirical item difficulties. Consequently, individual differences in response accuracy and response latency in the visual and auditory grammaticality judgment task would reflect individual differences in syntactic skills. This may be regarded as evidence for the construct validity of the test.

Finally, in terms of criterion validity, a test assessing syntactic abilities should be sensitive to developmental changes and learning progress during the primary school years. Thus, response accuracy was expected to increase and response latency was expected to decrease from Grade 3 to 4 (visual grammaticality judgment task) and from Grade 1 to 4 (auditory grammaticality judgment task). To test this hypothesis, grade level was included as another fixed item-level predictor into the LLTM for response accuracy and the LLT-analogous model for the log-transformed response latencies.

Results

Responses that were unusually slow or fast given the typical item- or person-specific response latencies (3 *SD* or more below the item-specific mean and 2 *SD* or more below or above the person-specific mean after standardizing each item by its item-specific mean) were excluded from the analyses because these responses, most likely, were anomalous (i.e. influenced by task-

irrelevant processes). Due to this procedure, there was a loss of 5.7% of the children's data in the visual version of the task and 7.7% in the auditory version. Descriptive statistics for the visual and auditory grammaticality judgments of the total sample and of each grade are reported in Table 2.

All models were tested with the software packages *lme4* (Bates, Maechler, Bolker, Walker, Christensen, & Sigmann, 2014) and *lmerTest* for R (Kuznetsova, Brockhoff, & Christensen, 2014). Parameters were estimated with Restricted Maximum Likelihood (REML). All significance tests were based on a type-I error probability of .05. Separate models were estimated for the visual and the auditory grammaticality judgment task and for accuracy and response latency as dependent variables.

---------- Table 2 about here ----------

In the following sections, we report the results from the Rasch and Rasch-analogous models (Step 1), from the LLTMs and the LLT-analogous models (Step 2), and the correlations between empirical item difficulties and predicted item difficulties (Step 3) for the visual and the auditory grammaticality judgment task and for response accuracy and response latency separately. Finally, to demonstrate that both tasks are sensitive to learning progress during primary school years, we report developmental changes from Grades 3 to 4 (visual grammaticality judgment task) and from Grades 1 to 4 (auditory grammaticality judgment task).

Visual Grammaticality Judgment Task

Response Accuracy

Step 1. The Rasch model for logit-transformed response accuracy as dependent variable revealed that all items were relatively easy to solve for third and fourth graders. Empirical item difficulties (depicted as the log-odds ratios for solving a specific item across participants) are depicted in the left panel of the person-item map in Figure 1a. Values with a positive sign indicate a probability higher than 50% to solve a specific item (a value of 2 equals 88% correct responses).

Person abilities (log-odds for providing a correct response across items) are depicted in the right panel of the person-item map in Figure 1a. The distributions of item difficulties and person abilities overlapped substantially, suggesting that the set of items measures individual differences across the whole range of person abilities in Grades 3 and 4. Items differentiated person abilities best for children with average to high syntactic abilities.

---------- Figure 1 about here ----------

Step 2. The parameter estimates for the fixed and random effects of the LLTM with logit-transformed response accuracy as dependent variable are provided in in the first column of Table 3 (Model 1). All main effects reached significance. As expected, children's response accuracy decreased (β=-0.04; z=-3.03; p<.05) with increasing syntactic complexity. Their responses were significantly less accurate for sentences containing a violation of word order (β=-0.29; z=-5.83; p<.05), tense form (β=-0.25; z=-3.91; p<.05), or case marking (β=-0.95; z=-15.88; p<.05) as compared to grammatically correct sentences. As expected, this effect was strongest for sentences containing violations of case marking and weakest for violations of tense form.

---------- Table 3 about here ----------

Step 3. Based on the intercept, the specific feature characteristics of each item and the regression coefficients for these features from the LLTM, we computed the predicted item difficulties for the response accuracies. As expected, these predicted item difficulties correlated strongly and positively with the empirical item difficulties. Forty-five percent of the variance in the empirical item difficulties was explained by the predicted item difficulties (Figure 2a).

---------- Figure 2 about here ----------

Response latency

Step 1. Rasch-analogous models for log-transformed response latency as dependent variable revealed that the distributions of empirical item difficulties (log-transformed latencies for

providing a response to a specific item across participants) and person abilities (log-transformed latencies for providing a response across items) overlap substantially. This suggests that the set of items measures individual differences across the whole range of person abilities. The distribution of item difficulties is depicted in the left panel of the person-items map in Figure 1b and the distribution of person abilities in the right panel (a log-transformed response latency of 8 equals 2,98 seconds). Items differentiated person abilities best for children with average syntactic abilities.

Step 2. The parameter estimates for the fixed and random effects of the LLT-analogous model with log-transformed response latency as dependent variable are provided in the third column of Table 3 (Model 1). Again, all main effects reached significance. Children needed significantly more time (β=0.06; $t(686)=29.37$; $p<.05$) to respond to syntactically complex sentences as compared to syntactically less complex sentences. Furthermore, they needed more time to respond to sentences containing a violation of word order (β=0.05; $t(686)=7.03$; $p<.05$), tense form (β=0.09; $t(686)=10.55$; $p<.05$), or case marking (β=0.07; $t(686)=7.03$; $p<.05$) as compared to grammatically correct sentences. This effect was strongest for sentences containing violations of tense form and weakest for violations of word order.

Step 3. Based on the intercept, the specific feature characteristics of each item, and the regression coefficients for these features from the LLT-analogous model, predicted item difficulties for response latency were computed and correlated with empirical item difficulties. Thirty-eight percent of the variance in the empirical item difficulties was explained by the predicted item difficulties (Figure 2b).

In sum, analyses of response accuracy and response latencies indicate that the systematic variation of theoretically-based item features that we expected to facilitate or increase the difficulty of syntactic processes resulted in items of varying difficulty. The reported findings suggest that the

items of the visual grammaticality judgment task indeed assess individual differences in syntactic skills. Finally, a test assessing syntactic skills should be sensitive to developmental changes and learning progress during primary school years. To test this prediction, grade level was included as further dummy-coded fixed effect into the LLTM for response accuracy and the LLT-analogous model for response latencies. The regression coefficients for response accuracy are presented in Table 3, second column (Model 2) and for response latencies in Table 3, fourth column (Model 2). As expected, children at Grade 4 responded significantly faster (β=-0.20; $t(685)$=-7.68; p<.05) and with overall higher accuracy (β=0.20; z=2.77; p<.05) to all sentence types than children at Grade 3. This finding indicates that the visual grammaticality judgment task is sensitive to developmental changes in syntactic skills.

Auditory Grammaticality Judgment Task

Response accuracy

Step 1. The Rasch model for logit-transformed response accuracy as dependent variable revealed that most items were relatively easy to solve for students from Grades 1 to 4. Empirical item difficulties (log-odds for solving a specific item across participants) are depicted in the left panel of the person-item map in Figure 3a. Again, values with a positive sign indicate a probability higher than 50% to solve a specific item (a value of 2 equals 88% correct responses). Person abilities (log-odds for providing a correct response across items) are depicted in the right panel of the person-item map in Figure 3a. The distribution of item difficulties and person abilities overlapped substantially, indicating that the items are suitable to reveal individual differences in subjects' syntactic skills. Items differentiated person abilities best for children with average to high syntactic abilities.

---------- Figure 3 about here ----------

Step 2. The parameter estimates for the fixed and random effects of the LLTM with logit-transformed response accuracy as dependent variable are provided in the first column of Table 4 (Model 1). The main effects for violation of word order (β=-0.73; z=-19.90; p<.05), tense form (β=-0.43; z=-8.89; p<.05), and case marking (β=-3.06; z=-67.98; p<.05) reached significance. As expected, children responded less accurately to sentences containing one of the three types of violation as compared to grammatically correct sentences. This effect was strongest for sentences containing violations of case marking and weakest for violations of tense form. Somewhat unexpectedly, there was no significant main effect for syntactic complexity.

---------- Table 4 about here ----------

Step 3. Based on the intercept, the specific feature characteristics of each item, and the regression coefficients for these features from the LLTM, predicted item difficulties for response accuracy were computed. The predicted item difficulties correlated substantially and positively with the empirical item difficulties. Seventy-three percent of the variance in the empirical item difficulties was explained by the predicted item difficulties (Figure 4a).

---------- Figure 4 about here ----------

Response latency

Step 1. Rasch-analogous models for log-transformed response latency as dependent variable revealed that the distributions of empirical item difficulties (log-transformed latencies for providing a response to a specific item across participants) and person abilities (log-transformed latencies for providing a response across items) overlap substantially. Again, this suggests that the set of items measures individual differences across the whole range of person abilities. The distribution of item difficulties is depicted in the left panel of the person-item map in Figure 3b and the distribution of person abilities in the right panel (a log-transformed response latency of 8 equals

2,98 seconds). Items differentiated person abilities best for children with average to high syntactic abilities.

Step 2. The parameter estimates for the fixed and random effects of the LLT-analogous model for log-transformed response latency as dependent variable are provided in the third column of Table 4 (Model 1). The model revealed significant main effects for syntactic complexity (β=0.11; $t(1111)=100.46$; $p<.05$), violation of tense form (β=0.07; $t(1111)z=15.78$; $p<.05$), and violation of case marking (β=0.15; $t(1111)=30.10$; $p<.05$). Children's response latency increased with increasing syntactic complexity and they needed more time to respond to sentences containing violations of tense form or case marking as compared to grammatically correct sentences. This effect was strongest for violations of case marking. There was no significant main effect for violation of word order.

Step 3. Based on the intercept, the specific feature characteristics of each item, and the regression coefficients for these features from the LLT-analogous model, predicted item difficulties for response latency were computed and correlated with empirical item difficulties. Fifty-four percent of the variance in the empirical item difficulties was explained by the predicted item difficulties (Figure 4b).

In sum, analyses of response accuracy and response latencies indicate that the systematic variation of theoretically-based item features resulted in items of varying difficulty. Thus, the items of the auditory grammaticality judgment task seem to be suitable to assess individual differences in syntactic skills. Finally, to test whether the test is sensitive to developmental changes and learning progress during primary school years, grade level was included into the LLTM for response accuracy and the LLT-analogous model for response latency in the form of three dummy-coded fixed effects for Grades 2, 3, and 4 with Grade 1 as the reference category. The regression coefficients for response accuracy are presented in Table 4, column two (Model 2) and for response

latency in column four (Model 2). There were no significant differences between Grade 2 and Grade 1. However, children responded faster (β=-0.04; t(1108)=-2.84; p<.05) and with overall higher accuracy (β=0.85; z=11.36; p<.05) to all sentence types in Grade 3 than in Grade 1. Moreover, they responded even faster (β=-0.10; t(1108)=-6.02; p<.05) and with even higher accuracy (β=1.39; z=17.10; p<.05) in Grade 4 as compared to Grade 1. These results suggest that the auditory grammaticality judgment task is suitable to detect developmental changes in syntactic skills during primary school years (Figures 5a and 5b).

---------- Figure 5 about here ----------

Discussion

The aim of the present study was to demonstrate the usefulness and feasibility of a process-oriented approach to assessing individual differences in cognitive component skills of reading comprehension selectively by well-defined process-specific tasks. To this end, we used the example of a grammaticality judgment task that selectively assesses syntactic skills in German primary school children. This test is characterized by a careful test design and stringent methodology: (1) It assesses individual differences in syntactic skills in a process-oriented fashion, (2) it does not involve reading-unrelated cognitive skills, (3) it assesses both aspects of efficient syntactic processes, i.e. their reliability and their degree of routinization, through response accuracy and response latency, and (4) individual differences in syntactic skills are assessed in reading as well as listening comprehension by means of parallel tasks and materials in both modalities. Most importantly, (5) the test is characterized by carefully constructed, theoretically-based test items with systematically varied item features that allow the meaningful interpretation of test scores in terms of individual differences in syntactic skills. The construct validity of this test was examined by way of explanatory item response models analyzing the impact of theoretically-based item features on response accuracy and latency. These models allow examining whether and to what

extent the test in fact captures individual differences in syntactic processing of written and spoken sentences.

For both the visual and the auditory grammaticality judgment task, items measured individual differences across the whole range of person abilities and differentiated person abilities best for children with average to high syntactic abilities. Furthermore, the high correlations between empirical and predicted item difficulties demonstrated that empirical item difficulties varied as a function of experimentally varied item features that are known to facilitate or hinder syntactic processing. These findings suggest that the carefully designed and theoretically-based items of both tasks indeed assess individual differences in syntactic skills. In the following, we summarize and discuss how empirical item difficulties varied as a function of experimentally varied item features (syntactic complexity, grammaticality, and type of syntactic violation).

As expected, children's response accuracy decreased and latencies increased with increasing syntactic complexity in both versions of the task. The only exception was found in the accuracy data of the auditory grammaticality judgment task. A possible explanation for the finding that response accuracy did not decrease with syntactic complexity in the auditory version of the task might be that even the most complex syntactic sentences in the test are not complex enough to show individual differences in response accuracy in the auditory processing modality. If we assume that spoken language processing requires less cognitive processing resources as compared to written language processing (see e.g. the verbal efficiency hypothesis, Perfetti, 1985), the assumption that the available resources in spoken sentence processing might be used to deal sufficiently with even highly complex sentences seems fairly reasonable. Yet, the impact of syntactic complexity on spoken sentence processing is reflected in the response latencies, indicating that the test is still suitable to measure individual differences in the degree of routinization of syntactic processes.

As to grammaticality, children needed more time and made more errors when responding to ungrammatical sentences – independent of violation type – as compared to grammatically correct sentences in both the visual and the auditory grammaticality judgment task. As expected, children responded least accurately to sentences containing a case-marking violation and most accurately to sentences containing a violation of tense form, with sentences containing a word-order violation in between them. Overall, analogous results were found for the response latencies. However, there were two unexpected findings: First, children did not need more time to respond to sentences containing a violation of word order as compared to grammatically correct sentences in the auditory grammaticality judgment task. A likely explanation might be that German children at the entry of primary school are already highly sensitive to word-order information as a cue to spoken sentence comprehension (Dittmar et al., 2008; Schipke et al., 2012). Therefore, they might have little problems detecting word-order violations when sentences are presented auditorily. Comparatively high response accuracy for these sentences (>85%) suggests that the missing significant main effect is not simply due to a speed-accuracy trade-off. We assume that processing of spoken word-order violations is comparatively easy, because word-order violations are likely to occur more often in spoken than in written language comprehension.

The second unexpected finding was that, in the visual grammaticality judgment task, children responded to sentences containing a violation of tense form as slowly as to sentences containing a violation of case marking. Possibly, children encountering the incorrect verb-tense form at the end of the written sentence regressed to the finite auxiliary that appeared earlier in the sentence to recheck whether the auxiliary and the sentence final verb form matched or not. Such a regression to the auxiliary might lead to longer response latencies and is possible only in the visual grammaticality judgment task.

Finally, when empirical and predicted item difficulties were correlated, there were a few items deviating considerably from the estimated regression line. Items (6) and (7) were actually somewhat easier and item (8) more difficult than would have been expected.

(6) *Ein Vogel ist zum Nest geflogen* / A bird has flown to the nest

(7) *Das Eis ist in der Sonne geschmolzen* / The ice has melted in the sun

(8) *Der Fuchs hat die Gans gestohlen* / The fox has stolen the goose

The most likely explanation is that there are some other item features, such as the verb's valence or lexical effects, which were not varied systematically but nevertheless determine item difficulties to some degree. However, despite these few exceptions, the predicted item difficulties correlated overall strongly and positively with the empirical item difficulties, suggesting that empirical item difficulties reflect the cognitive demands the items make on syntactic processing. Thus, the items of the visual and the auditory grammaticality judgment task indeed seem to assess individual differences in syntactic skills.

Last but not least, changes in response accuracy and response latency across grade levels suggest that both tasks are sensitive to developmental changes in cognitive syntactic processes. Whereas a speed-up across grade levels might also indicate an increase in more general cognitive maturity (rather than increased routinization in syntactic processes), the obtained increase in response accuracy strongly suggests increased sensitivity to the different types of grammatical cues. It might be argued that the observed developmental changes are a mere result of improved visual word recognition skills, which might have spared cognitive resources at the word level (see e.g. Perfetti, 1985) and, consequently, might have had a favorable impact on the outcome in the visual grammaticality judgment task. A related objection might be that the observed individual performance differences in the visual grammaticality judgment task could be, at least to a certain degree, a mere by-product of individual differences in word recognition skills. However, if

developmental changes or individual differences observed in the visual grammaticality judgment task were simply due to intra- or interindividual differences in visual word recognition skills, results should have differed substantially in both modalities (which was not the case), given that visual word recognition skills are not necessary in the auditory grammaticality judgment task. Moreover, results from the explanatory item response models strongly suggest that both tasks indeed assess syntactic skills. Thus, developmental changes and individual differences in both tasks are likely to reflect intra- and interindividual differences in syntactic skills, respectively.

Overall, the findings of the present study demonstrate that our grammaticality judgment task for German primary school children provides a valid instrument for assessing individual differences in syntactic skills, which can be expected to contribute to reading comprehension (see Schindler & Richter, in press). In an earlier study, Richter et al. (2012) demonstrated that response-accuracy and response-latency measures in the ProDi-L grammaticality judgment task are significantly related to reading comprehension assessed via the standardized text comprehension subtest of ELFE 1-6 (Lenhard & Schneider, 2005). Richter et al.'s findings together with the findings reported in the present study suggest that it should be possible to use the ProDi-L grammaticality judgment task to detect syntactic deficits that might contribute to reading difficulties. Consequently, it can serve as a basis for the construction of reading interventions that are tailored to the individual needs of poor readers with deficient syntactic processes.

Of course, diagnosed syntactic deficits do not necessarily imply that reading comprehension difficulties stem from poor syntactic abilities alone (Schindler & Richter, in press). A more comprehensive diagnosis of individual patterns of reading difficulties requires a test battery that assesses individual differences in all cognitive component processes of reading comprehension at the word, sentence, and text level selectively. Such a test battery is provided by ProDi-L, with the

grammaticality judgment task presented in this paper being part of this comprehensive process-oriented reading assessment.

The present study emphasized the necessity for process-oriented tests of assessing reading comprehension skills to identify individual deficits in cognitive component skills of reading comprehension, which can cause reading difficulties. A carefully constructed grammaticality judgment task that assesses individual differences in syntactic skills separately served as example. By means of explanatory item response models, we demonstrated that process-oriented tests involving well-defined reading tasks, theoretically-based test items with carefully varied item features, and accuracy as well as response latency as diagnostic measures, can be valid tools to achieve that goal.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press. [Online available from: http://dx.doi.org/10.1017/CBO9780511732935]

Angermaier, M. J. W. (1977). *Psycholinguistischer Entwicklungstest: PET* [Psycholinguistic test of development]. Göttingen, Germany: Hogrefe.

Bates, E., & MacWhinney, B. (1989). Functionalism and the Competition Model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3-76). New York: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Sigmann, H. (2014). *Lme4: Linear mixed-effects models using Eigen and S4* [Software]. R-package version 1.1-6. Retrieved in June 2014 from: http://cran.r-project.org/package=lme4

Bäuerlein, K., Lenhard, W., & Schneider, W. (2012). *Lesetestbatterie für die Klassenstufen 6-7*: *LESEN 6-7* [Reading-test battery for Grades 6-7]. Göttingen, Germany: Hogrefe.

Baum, S. R. (1991). Sensitivity to syntactic violations across the age-span: Evidence from a word-monitoring task. *Clinical Linguistics and Phonetics*, *5*(4), 317-328.

Byrne, B. (1981). Deficient syntactic control in poor readers: Is a weak phonetic memory code responsible? *Applied Psycholinguistics, 2*(3), 201-212.

Carpenter, P. A., Miyake, A., & Just, M. A. (1995). Language comprehension: Sentence and discourse processing. *Annual Review of Psychology*, *46*, 91-120.

Casalis, S., & Louis-Alexandre, M.-F. (2000). Morphological analysis, phonological analysis and learning to read French: A longitudinal study. *Reading and Writing, 12*(3), 303-335.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219-232.

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child Development*, *79*(4), 1152-1167.

Dixon, P. (2008). Models of accuracy in repeated measures designs. *Journal of Memory and Language*, *59*(4), 447-456.

Elben, C. E., & Lohaus, A. (2001). *Marburger Sprachverständnistest für Kinder: MSVK* [Marburg test of language comprehension for children]. Göttingen, Germany: Hogrefe.

Esser, G., Wyschkon, A., Ballaschk, K., & Hänsch, S. (2010). *Potsdam-Illinois Test für Psycholinguistische Fähigkeiten: P-ITPA* [Potsdam-Illinois test of psycholinguistic abilities]. Göttingen, Germany: Hogrefe.

Ferreira, F., Handerson, J. M., Anes, M. D., Weeks, P. A., & McFarlane, D. K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 324-335.

Fischer, G. H. (1974). *Einführung in die Theorie Psychologischer Tests* [Introduction to the theory of psychological tests]. Bern, Switzerland: Huber.

Flores d'Arcais, G. B. (1982). Automatic syntactic computation and use of semantic information during sentence comprehension. *Psychological Research*, *44*(3), 231-242.

Fox, A. V. (2011). *Test zur Überprüfung des Grammatikverständnisses: TROG-D* [Test for the comprehension of grammar]. Idstein, Germany: Schulz-Kirchner Verlag.

Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic Research*, *31*(1), 45-63.

Graesser, A. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior, 19*(2), 135-151.

Grimm, H., & Schöler, H. (2001). *Heidelberger Sprachentwicklungstest: HSET* [Heidelberg test of language development]. Göttingen, Germany: Hogrefe.

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten [Construct validity and scale interpretation in competence diagnostic by predicting task difficulty]. *Psychologische Rundschau*, *63*(1), 43-49.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665-686.

Hulstijn, H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. (Language Learning & Language Teaching, Vol. 41). Amsterdam: John Benjamins.

Hulstijn, J. H., van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555-582.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160-212.

Johnson-Laird, P. N. (1981). Comprehension as the construction of mental models. *Philosophical Transactions of the Royal Society, Series B, 295*(1077), 353-374.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random*

and fixed effects for linear mixed effect models (lmer objects of lme4 package). R-package version 2.06. Retrieved in June 2014 from: http://cran.r-project.org/web/packages/lmerTest/index.html

Lehmann, R. H., Peek, R., & Poerschke, J. (1997). *Hamburger Lesetest für 3. und 4. Klassen: HAMLET 3-4* [Hamburg reading test for Grades 3 and 4]. Weinheim und Basel: Beltz.

Lenhard, W., & Schneider, W. (2006). *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler* [ELFE 1-6: A reading comprehension test for grades one to six]. Göttingen, Germany: Hogrefe.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247-1282.

May, P., & Arntzen, H. (2000). *Hamburger Leseprobe: Testverfahren zur Beobachtung der Leselernentwicklung in der Grundschule* [Hamburg reading probe: Assessment instrument for oberserving reading development in primary school]. Hamburg: Selbstverlag.

Marton, K., & Schwartz, R. G. (2003). Working memory capacity and language processes in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *46*(5), 1138-1153.

Marton, K., Schwartz, R. G., & Braun, A. (2005). The effect of age and language structure on working memory performance. In B. G. B. L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the XXVII. Annual Meeting of the Cognitive Science Society* (pp. 1413-1418). Mahwah, NJ: Erlbaum.

Moll, K., & Landerl, K. (2010). *SLRT-II Lese- und Rechtschreibtest: Weiterentwicklung des Salzburger Lese- und Rechtschreibtests (SLRT)* [Test of reading and writing: Advancement of the Salzburg test of reading and writing]. Bern, Switzerland: Hans Huber.

Müller, B., & Richter, T. (2014). Lesekompetenz [Reading competence]. In J. Grabowski (Ed.), *Sinn und Unsinn von Kompetenzen: Fähigkeitskonzepte im Bereich von Sprache, Medien und Kultur* (pp. 29-49). Opladen, Germany: Budrich.

Nation, K., & Snowling, M. J. (2000). Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied Psycholinguistics, 21*(2). 229-241.

Oakhill, J. V., & Cain, K. (1997). Assessment of comprehension in reading. In J. R. Beech & C. Singleton (Eds.), *The psychological assessment of reading* (pp. 176-203). London, UK: Routledge.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. M. Brown & P. Hagoort (Eds.), *The neuroscience of language* (pp. 167-208). Oxford: University Press.

Perfetti, C., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulmes (Eds.), *The science of reading: A handbook* (pp. 227-247). Oxford, UK: Blackwell.

Petermann, F., Meth, D., & Fröhlich, L. P. (2010). *Sprachstandserhebungstest für Fünf- bis Zehnjährige: SETK 5-10* [Language assessment test for five to ten-year-olds]. Göttingen: Hogrefe.

Plaza, M., & Cohen, H. (2003). The interaction between phonological processing, syntactic awareness, and naming speed in the reading and spelling performance of first-grade children. *Brain and Cognition, 53*(2), 287-292.

Pylkkänen, L., & McElree, B. (2006). The syntax-semantic interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd ed., pp. 539-580). London: Academic Press.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 358-374.

Richter, T., Isberner, M.-B., Naumann, J., & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern. [Process-based assessment of reading skills in primary school children]. *Zeitschrift für Pädagogische Psychologie*, *26*(4), 313-331.

Richter, T., Isberner, M.-B., Naumann, J., & Neeb, Y. (2013). Lexical quality and reading comprehension in primary school children. *Scientific Studies of Reading, 17*(6), 415-434.

Richter, T., & Naumann, J. (2012). Schlussbericht der ersten Förderphase des BMBF geförderten Verbundprojekts „Prozessbezogene Diagnostik des Lese- und Hörverstehens im Grundschulalter“ [Report of the first three years of the BMBF funded project „Process-based assessment of reading and listening comprehension in primary school children“] (Research Report). Retrieved from Leibnitz Information Centre for Science and Technology University Library website: https://www.tib.eu/suchen/id/TIBKAT:875828442/

Richter, T., Naumann, J., Isberner, M.-B., Neeb, Y., & Knoepke, J. (2017). *Prozessbezogene Diagnostik des Leseverstehens bei Grundschulkindern: ProDi-L* [Process-based assessment of reading skills in primary school children] [Computerized test]. Göttingen, Germany: Hogrefe.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design* (ETS Research Report No. RR-13-30). Retrieved from Educational Testing Service website: http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02337.x/epdf

Schindler, J., & Richter, T. (in press). Reading comprehension: Individual differences, disorders, and underlying cognitive processes. In A. Bar On & D. Ravid (Eds.), *Handbook of communication disorders: Theoretical, empirical, and applied linguistic perspectives*. Berlin, Germany: De Gruyter Mouton.

Schipke, C. S., Knoll, L. J., Friederici, A. D., & Oberecker, R. (2012). Preschool children's interpretation of object-initial sentences: Neural correlates of their behavioral performance. *Developmental Science*, *15*(6), 762-774.

Schneider, W., Blanke, I., Faust, V., & Küspert, P. (2011). *WLLP-R – Würzburger Leise Leseprobe – Revision* [WLLP-R – Würzburg silent reading test - revision]. Göttingen, Germany: Hogrefe.

Souvignier, E., Trenk-Hinterberger, I., Adam-Schwebe, S., & Gold, A. (2008). *Frankfurter Leseverständnistest (FLVT 5-6)* [Frankfurt reading comprehension test (FLVT 5-6)]. Göttingen, Germany: Hogrefe.

Steinert, J. (2011). *Allgemeiner Deutscher Sprachtest: ADST* [General German language test]. Göttingen, Germany: Hogrefe.

Stothard, S. E., & Hulme, C. (1992). Reading comprehension difficulties. The role of language comprehension and working memory skills. *Reading and Writing: An Interdisciplinary Journal, 4*(3), 245-256.

Taraban, R., & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectation. *Journal of Memory and Language*, *27*(6), 597-632.

Trapman, M., van Gelderen, A., van Steensel, R., van Schooten, E., & Hulstijn, J. (2014). Linguistic knowledge, fluency, and meta-cognitive knowledge as components of reading

comprehension in adolescent low achievers: differences between monolinguals and bilinguals. *Journal of Research in Reading*, *37*(S1), S3-S21.

Tunmer, W. E., Nesdale, A. R., & Wright, A. D. (1987). Syntactic awareness and reading acquisition. *British Journal of Developmental Psychology*, *5*(1), 25-34.

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*(2), 359-376.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*(1), 2-40.

Walter, J. (2013). *VSL: Verlaufsdiagnostik sinnerfassenden Lesens* [VSL: Developmental diagnosis of reading comprehension]. Göttingen, Germany: Hogrefe.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43-74). New York: Springer.

Wimmer, H., & Mayringer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2-9: SLS 2-9* [Salzburg screening test of reading for Grades 2-9]. Bern, Switzerland: Huber.

Table 1:

Study Sample. Student's Age at Grades 1 to 4 (Auditory Grammaticality Judgment Task) and 3 to 4 (Visual Grammaticality Judgment Task)

	Auditory grammaticality judgment task				Visual grammaticality judgment task			
		Age (years; months)				Age (years; months)		
	n	*M*(*SD*)	*Min*	*Max*	*n*	*M*(*SD*)	*Min*	*Max*
Grade 1	439	7;6 (0;6)	5;5	9;10	-	-	-	-
Grade 2	237	8;5 (0;6)	5;9	10;4	-	-	-	-
Grade 3	232	9;7 (0;7)	6;9	11;10	332	9;5 (0;7)	6;9	11;10
Grade 4	209	10;6 (0;5)	9;2	12;4	359	10;5 (0;5)	8;8	12;4

Note. 429 children of Grades 3 and 4 participated in both versions of the task.

Table 2:

Descriptive Statistics for Response Accuracy and Response Latency (Log-transformed) as Dependent Variables in the Visual and Auditory Grammaticality Judgment Task

	Total		Grade 1		Grade 2		Grade 3		Grade 4	
	n	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)
Auditory grammaticality task										
Response accuracy[a]	39207	0.821 (0.38)	15006	0.752 (0.43)	8420	0.841 (0.37)	8286	0.855 (0.35)	7459	0.899 (0.30)
Response latency[b]	39207	8.038 (0.36)	15006	8.082 (0.41)	8420	8.048 (0.35)	8286	8.017 (0.31)	7459	7.964 (0.29)
Visual grammaticality task										
Response accuracy[a]	24764	0.883 (0.32)	-	-	-	-	11877	0.874 (0.33)	12887	0.892 (0.31)
Response latency[b]	24764	8.204 (0.57)	-	-	-	-	11877	8.309 (0.59)	12887	8.108 (0.53)

Note. *n* = number of data points (number of items X number of participants). [a]proportions, [b]log-transformed response latencies in ms. For one child (36 observations), grade level information was missing.

Table 3:

Fixed Effects and Variance Components in the LLTMs for Response Accuracy and LLT-analogous Models for Response Latency in the Visual Grammaticality Judgment Task

	Response Accuracy		Response Latency	
	Model 1	Model 2	Model 1	Model 2
Parameter	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
		Fixed Effects		
Intercept	2.479 (0.05)*	2.373 (0.06)*	8.162 (0.01)*	8.265 (0.02)*
Number of syntactic phrases	-0.044 (0.01)*	-0.044 (0.01)*	0.061 (0.00)*	0.061 (0.00)*
Violation of word order	-0.292 (0.05)*	-0.292 (0.05)*	0.049 (0.01)*	0.049 (0.01)*
Violation of tense form	-0.249 (0.06)*	-0.249 (0.06)*	0.094 (0.01)*	0.094 (0.01)*
Violation of case marking	-0.949 (0.06)*	-0.949 (0.06)*	0.071 (0.01)*	0.071 (0.01)*
Grade level		0.205 (0.07)*		-0.198 (0.03)*
		Variance Components		
Subjects	0.623 (0.79)	0.615 (0.78)	0.119 (0.34)	0.109 (0.33)

Note. Number of syntactic phrases (grand-mean centered). Violation of word order (dummy coded): Sentences containing a violation of word order (=1) vs. all other sentence types (=0). Violation of tense form (dummy coded): Sentences containing a violation of tense form (=1) vs. all other sentence types (=0). Violation of case marking (dummy coded): Sentences containing a violation of case marking (=1) vs. all other sentence types (=0). Grade level (dummy coded): Grade 4 (=1) vs. Grade 3 (=0).
*$p<0.05$ (two-tailed).

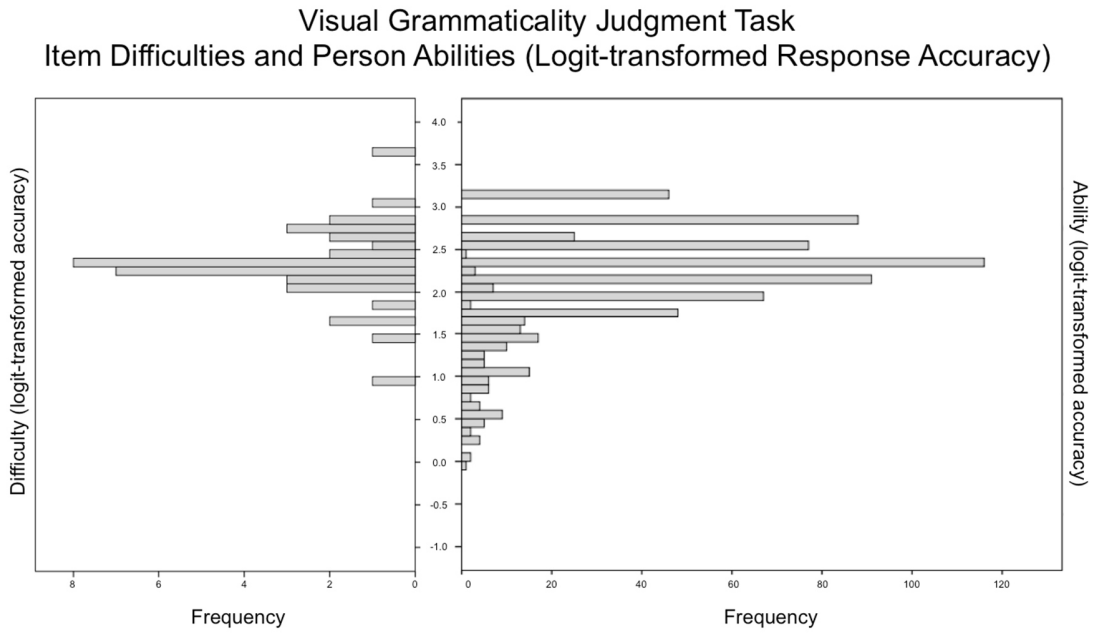
Table 4:

Fixed Effects and Variance Components in the LLTMs for Response Accuracy and LLT-analogous Models for Response Latency in the Auditory Grammaticality Judgment Task

	Response accuracy		Response latency	
	Model 1	Model 2	Model 1	Model 2
Parameter	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
		Fixed Effects		
Intercept	2.494 (0.04)*	1.902 (0.05)*	8.002 (0.01)*	8.033 (0.01)*
Number of syntactic phrases	-0.004 (0.01)	-0.003 (0.01)	0.107 (0.00)*	0.107 (0.00)*
Violation of word order	-0.734 (0.04)*	-0.732 (0.04)*	0.003 (0.00)	0.003 (0.00)
Violation of tense form	-0.430 (0.05)*	-0.431 (0.05)*	0.069 (0.00)*	0.069 (0.00)*
Violation of case marking	-3.061 (0.05)*	-3.063 (0.05)*	0.145 (0.00)*	0.145 (0.00)*
Grade 2		0.733 (0.07)*		-0.016 (0.02)
Grade 3		0.851 (0.07)*		-0.044 (0.02)*
Grade 4		1.391 (0.08)*		-0.097 (0.02)*
		Variance Components		
Subjects	0.834 (0.91)	0.571 (0.76)	0.036 (0.19)	0.034 (0.19)

Note. Number of syntactic phrases (grand-mean centered). Violation of word order (dummy coded): Sentences containing a violation of word order (=1) vs. all other sentence types (=0). Violation of tense form (dummy coded): Sentences containing a violation of tense form (=1) vs. all other sentence types (=0). Violation of case marking (dummy coded): Sentences containing a violation of case marking (=1) vs. all other sentence types (=0). Grade 2 (dummy coded): Grade 2 (=1) vs. Grade 1, 3, and 4 (=0). Grade 3 (dummy coded): Grade 3 (=1) vs. Grade 1, 2, and 4 (=0). Grade 4 (dummy coded): Grade 4 (=1) vs. Grade 1, 2, and 3 (=0).
*$p<0.05$ (two-tailed).

a)



b)

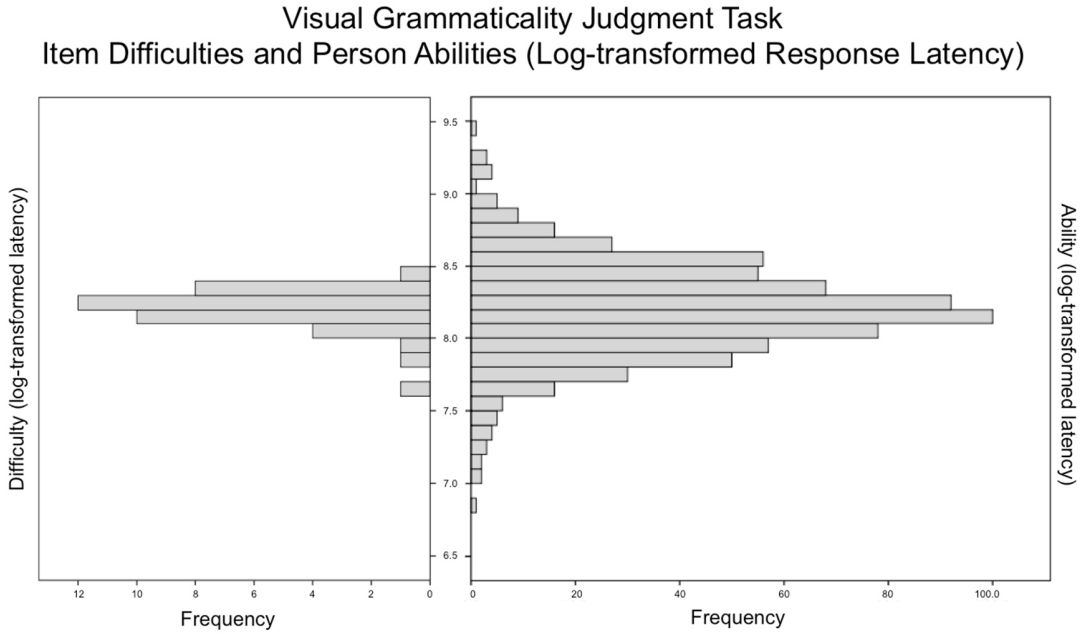
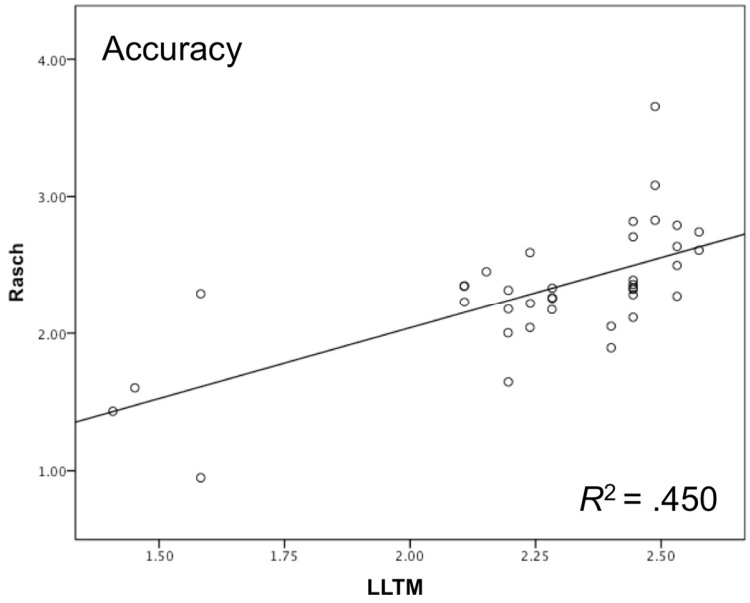


Figure 1. Item difficulties (left) and person abilities (right) for logit-transformed response accuracy (a) and log-transformed response latency (b) in the visual grammaticality judgment task in Grades 3 and 4.

a)



b)

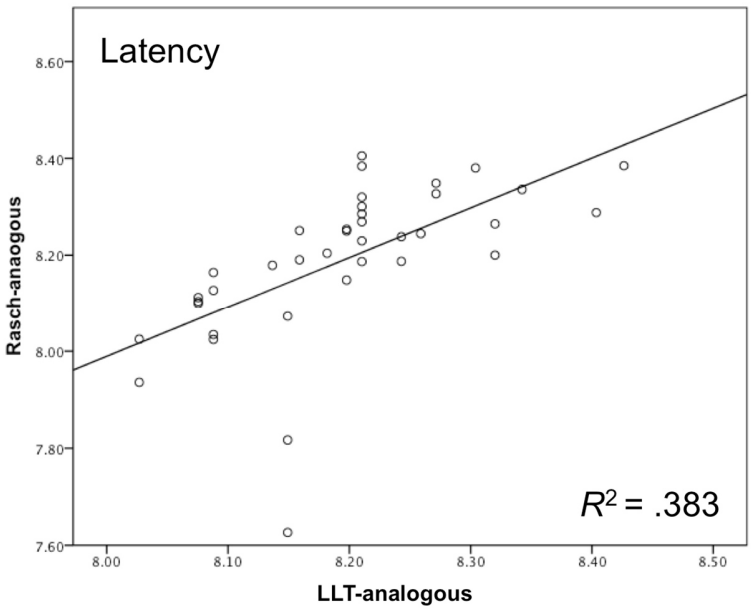
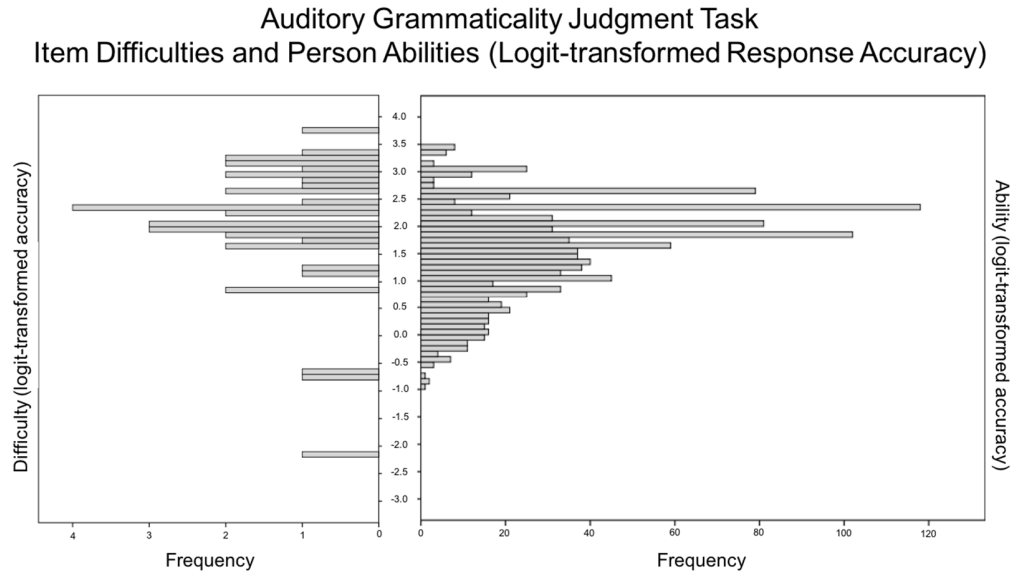


Figure 2. Variance in empirically observed item difficulties explained by item difficulties predicted from item characteristics for accuracy (a) and latency (b) in the visual grammaticality judgment task in Grades 3 and 4.

a)



b)

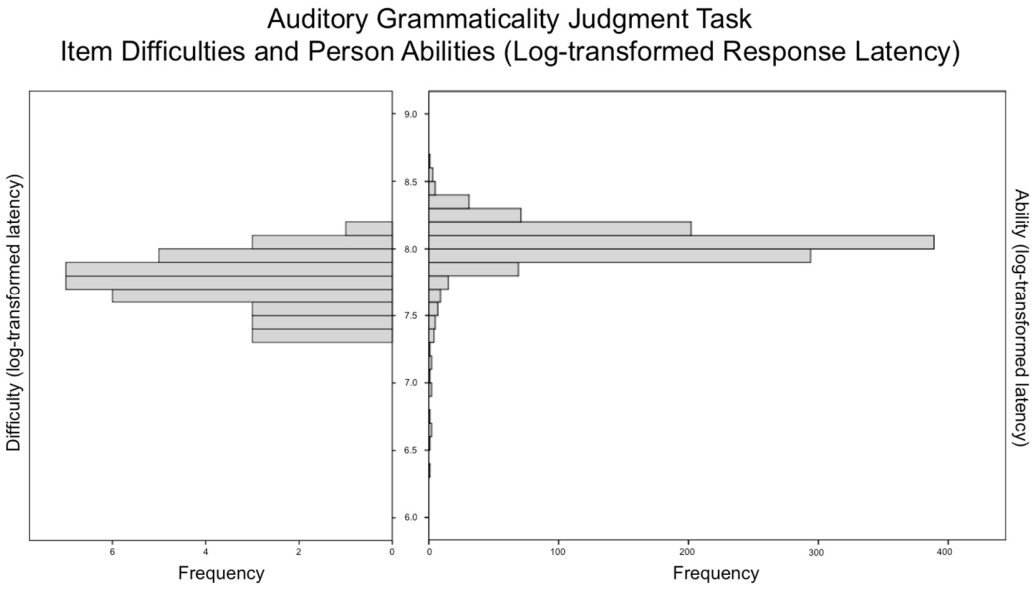
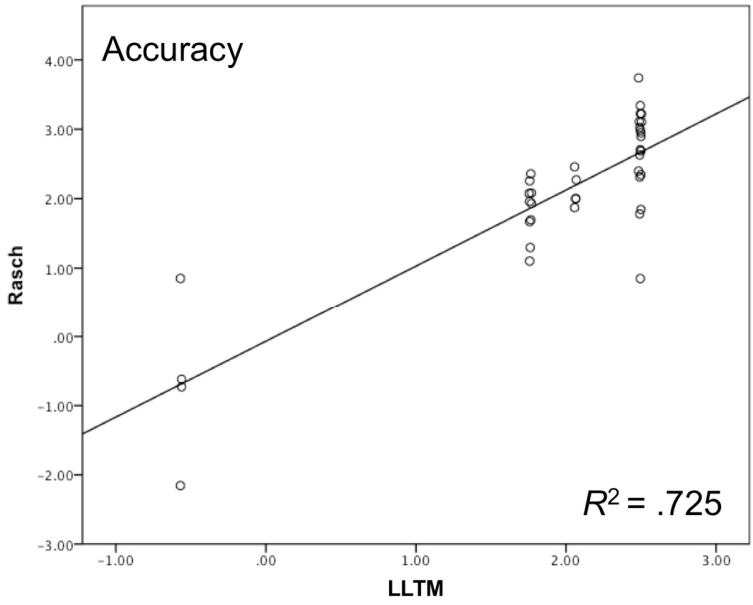


Figure 3. Item difficulties (left) and person abilities (right) for logit-transformed response accuracy (a) and log-transformed response latency (b) in the auditory grammaticality judgment task in Grades 1 to 4.

a)



b)

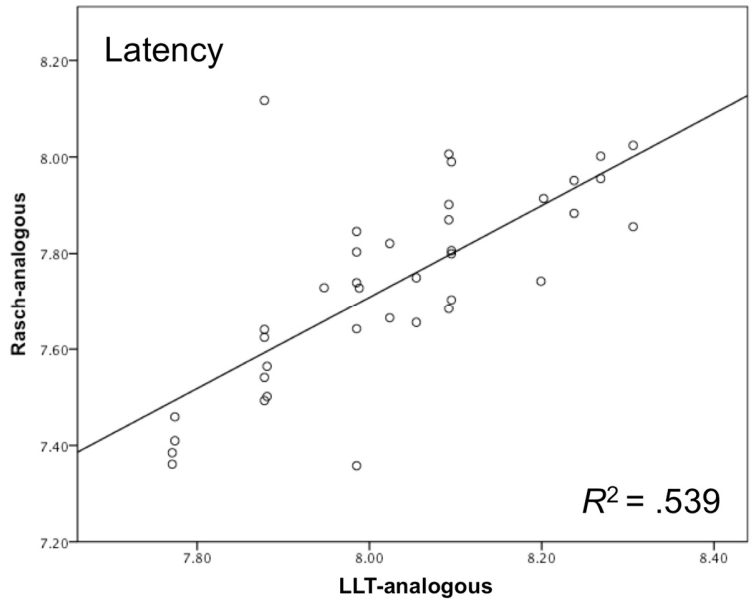
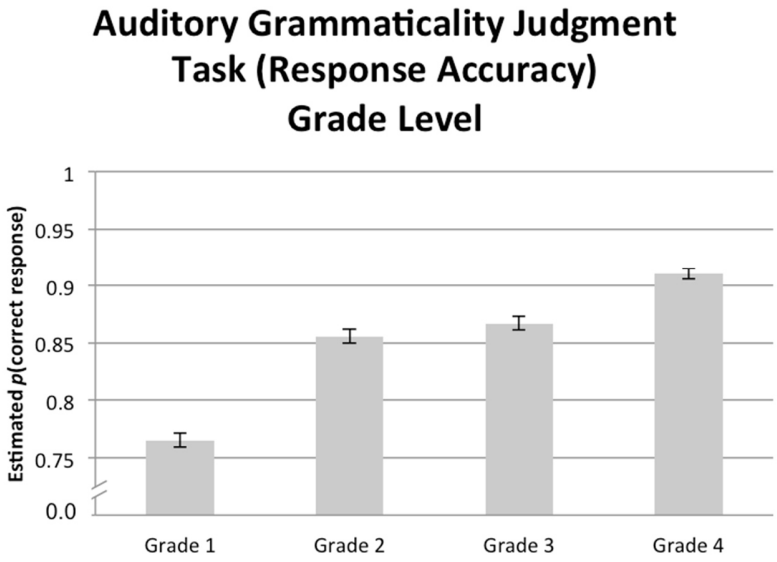


Figure 4. Variance in empirically observed item difficulties explained by item difficulties predicted from item characteristics for accuracy (a) and latency (b) in the auditory grammaticality judgment task in Grades 1 to 4.

a)



b)

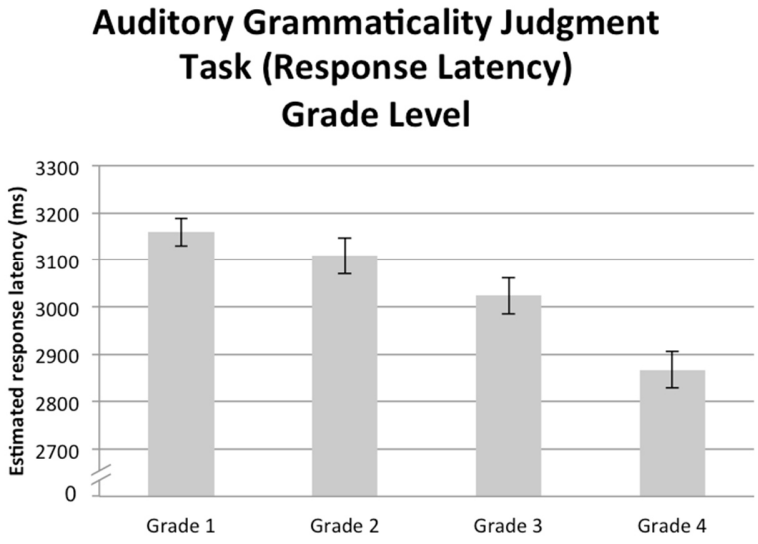


Figure 5. Model-based estimated probability of correct responses with standard error (a) and model-based estimated response latency with standard error (b) in the auditory grammaticality judgment task by grade level.