Using the Quantifying Bilingual Language Experience Questionnaire to Identify Risk for Language Impairment in Bilingual Children

Accepted for publication in the Journal of Speech, Language, and Hearing Research

Using Q-BEx to Identify Risk for Language Impairment in Bilingual Children

Cécile De Cat[1], Laurice Tuller[2], Arief Gusnanto[1], Draško Kašćelan[3], Philippe Prévost[2], Ludovica Serratrice[4], Sharon Unsworth[5]

[1] University of Leeds, UK

[2] University of Tours, FR

[3] University of Essex, UK

[4]University of Reading, UK

[5]Radboud University, NL

Author Note

Author Contributions

Cécile De Cat: Co-first author, Conceptualization, data curation, formal analysis, funding acquisition, project administration, writing–original draft

Laurice Tuller: Co-first author, Conceptualization, data curation, formal analysis, writing–original draft

Arief Gusnanto: Formal analysis, writing-editing

Draško Kašćelan: Conceptualization, data collection, data curation, writing–review & editing

Philippe Prévost: Conceptualization, data curation, writing–review & editing

Ludovica Serratrice: Conceptualization, data curation, writing–review & editing

Sharon Unworth: Conceptualization, data curation, writing–review & editing

Conflict of Interest statement: There are no conflicts of interest.

Corresponding author: Laurice Tuller, tuller@univ-tours.fr

Abstract

Purpose:
Clinical triage, but also inclusive population sampling in research, requires efficient detection of children likely to be at risk for language impairment (LI), a significant challenge in the case of bilingual children. Our goal was to derive automatic risk factor indices from a parental questionnaire and compare their usefulness in initial identification of children at risk for LI.

Method:

Alternative risk factor indices were derived from the Q-BEx online questionnaire backend calculator, differing in the weight and detail given to language proficiency. Each index was applied to two datasets comprised of five- to eight-year-old children: one with independent diagnosis of Developmental Language Disorder (DLD)/Typical Development (TD) (109 bilingual children tested in France) and one of largely all-comer children (278 bilingual and monolingual children tested in France, the Netherlands, and the United Kingdom). We compared how these alternative indices predicted (likely) DLD/TD status, and structural language outcomes, assessed with LITMUS tools, designed for bilingual children.

Results:

RF4-min, a risk factor index composed of four equally weighted components (age of first word, age of first sentence, early development concerns, and strongest speaking skills between the HL and the SL), satisfactorily separated bilingual children according to developmental status, and was a significant predictor of language outcomes. This composite score showed the best balance between separability, sensitivity, and predictive value compared to indices without a proficiency component or with a more heavily weighted/detailed proficiency component.

Conclusion:
RF4-min, a Q-BEx automatically derived composite risk factor index, yielded promising results, indicating usefulness as part of first-level triage for LI.

Key words: parental questionnaire; screening; bilingualism

Using Q-BEx to Identify Risk for Language Impairment in Bilingual Children

Introduction

Understanding bilingual language development entails studying children growing up in a variety of linguistic and cultural settings, reflecting the diversity that is the reality of so many populations across the world. Understanding how children acquire more than one language also means including the significant numbers of children who are growing up with a neurodevelopmental condition that affects language (such as Developmental Language Disorder, DLD, but also in many cases, autism, among others). Both endeavors, understanding typical bilingual language acquisition and understanding atypical bilingual language acquisition, involve identifying children who are at risk for language impairment (LI), the focus of this study. Identification of risk for LI in bilingual children, a by-now widely discussed issue, is vexing as traditional assessments are not only time-consuming, but they are typically not designed with bilingual children in mind, leading to misidentification of language disorders (Armon-Lotem et al., 2015; Paradis et al., 2011; Prévost & Tuller, 2022). Yet, pinpointing bilingual children likely to have LI is important in research, for identification and characterization of population samples, but also in educational settings, since understanding the nature of a child's language difficulties is the first step toward organizing useful support. Its importance in clinical settings is fundamental: a rich empirical base is crucial for optimal detection and treatment of language disorders in bilingual children, but also, more specifically, for the design and implementation of effective triage procedures.

Effective triage is a pressing issue in many clinical contexts, where health systems are over-run (e.g., a 24-month wait for initial assessment for LI in children is not uncommon in some countries: in England in April 2024, over 24,000 children and young people were on a waiting list for over a year for speech and language therapy services – National Health Service, 2024). Furthermore, bilingual children are generally over-represented in speech-language pathology (SLP) screening; for example, while 30% of children in Sweden are bilingual, waiting lists for SLP services can include 70% bilingual children (Wehbe, 2023). Optimal triage must clearly be geared to handle bilingualism.

The goal of triage is to ensure that children who are the most likely to need services are the ones prioritized for screening. By contrast, screening seeks to identify children who may benefit from deeper evaluation and intervention. Identification of LI in the contexts of triage and/or screening is based notably on the following factors: (a) language performance below age-expectations in each of the child's languages, (b) late early language milestones, (c) caregiver concerns about the child's early language development, and (d) history of language impairment/neurodevelopmental disorders (NDD) in the child's immediate family.

Language performance below age expectations is part of the very definition of (Developmental) Language Disorder: "Language abilities are substantially and quantifiably below those expected for age, resulting in functional limitations in effective communication, social participations, academic achievement, or occupational performance, individually or in any combination" (DSM-5, APA, 2013). Language screening for a child growing up in a bilingual context requires, in addition, ascertaining whether the child is competent in any one of their languages, since a child who is competent in at least one language is not language impaired. Ideally, testing a child in each of their languages is therefore the key to understanding whether that child's difficulties with the societal language (SL) are the result of relatively low amounts of exposure and or the result of LI. This method is generally recommended in guidelines for speech-language pathologists (e.g., Kohnert, 2010).

Reliance on dual language testing to identify LI may result in the exclusion of many children from the screening process (see Boerma & Blom, 2017; Paradis et al., 2010, 2013). This is because many children growing up with more than one language speak a language which has been the object of very little, if any, relevant scientific scrutiny (psycholinguistic, or even also just plain descriptive), for which there are no language assessment tools, normed or not, and/or for which there are no trained personnel to administer and interpret assessment. Yet, these children represent a significant portion of bilingual children in many countries in Europe and elsewhere. For example, Sub-Saharan African languages, languages for which tests and bilingual speech-language pathologists (SLPs) are generally not available, are not only highly represented among bilingual children growing up in France, but they are also extremely diverse. More than 100 different languages just from the Niger-Congo family have been identified for schoolchildren in the Paris region—Bambara, Lingala, Malinké, Mandingue, Peul, Sango, Senoufo, Soninké, Wolof, and many, many more (Condon & Régnard, 2010). It is simply not feasible to test all of these home languages, in clinical or research settings. This issue may be further exacerbated when dialectal variants need to be identified and assessed (see Pert, 2023).

Reliance on direct assessment of the home/heritage language (HL) to identify risk for LI means that the nature of many children's SL difficulties cannot be elucidated, resulting in a largely inadequate screening process. Paradoxically, even though bilingual children may be overrepresented in SLP screening, as reported above, unavailability of adequate screening also means that many such children are excluded from research; inclusive research on bilingual language acquisition cannot be based exclusively on bilingual children speaking a limited number of pairs of languages such as English-French, Spanish-English, Turkish-German, or Arabic-Swedish, to name a few. The need for diverse population sampling in work on child LI is part of growing diversity awareness in the fields of psycholinguistics and developmental psychology (Kidd & Garcia, 2022; Nielsen et al., 2017). In addition to concerns about inclusion due to lack of assessment tools and qualified personnel/funds for test administration and interpretation in languages other than the SL, severe time constraints do not allow for direct assessment of more than one language in many clinical settings.

One answer to the challenges regarding assessing children in all of their languages is indirect assessment of the HL. Parental estimates about bilingual children's proficiency in the HL(s) have been found to be generally reliable and useful. These measures have been shown to align with direct testing of the HL and to be useful for discriminating between typical and atypical language development (e.g., Abutbul-Oz & Armon-Lotem, 2022; Bedore et al., 2011; Boerma & Blom, 2017; Paradis et al., 2010; Pratt et al., 2020).

The predictive value of late early language milestones and caregiver early language development concerns is well established, and indeed has widely been integrated into public health guidelines (e.g., the French government guide for detecting atypical development in children under age seven – Ministère du travail, de la santé et des solidarités, 2024). The usefulness of information regarding these factors has been shown to be relevant for monolingual children regardless of socio-economic status (e.g., Callu et al., 2003), and for bilingual children across a variety of cultural contexts (Abutbul-Oz & Armon-Lotem, 2022; Boerma & Blom, 2017; Grimm & Schulz, 2014; Paradis et al., 2010, 2011, 2013; Restrepo, 1998; Tuller et al., 2018).

Familial history for language impairment/NDD is another well-studied risk factor for LI in children. Its relevance has been demonstrated for monolingual children. However, for bilingual children, familial history of LI as a predictor of language performance has a mixed track record. While it has been shown to be meaningful for multilingual children in some cultural contexts (Grimm & Schulz, 2017; Restrepo, 1998), for multilingual children in many other cultural contexts, its usefulness has been found to be minimal or lacking. As pointed out by Paradis et al. (2010), gathering information on family language and other developmental difficulties tends to be problematic in many bilingual contexts due to issues in willingness and ability to report this history. They cite three factors they encountered in clinics and in their pilot study: lack of school success due to external factors (conflicts/war, displacement, economic hardships), non-recognition of difficulties due to availability of services and differences in how they are viewed, and cultural barriers about disclosing information about relatives, including shame. In other words, and this has been our experience as well, not only do some parents view these questions as intrusive (particularly for cultures in which such difficulties may not be well known/accepted), but also many bilingual children grow up in families whose parents (and older siblings) may have had difficulties that were not detected in their home country, due to lack of relevant infrastructure, making their answers to such questions irrelevant. Given these considerations, it is not surprising that information about this risk factor has been shown to be less predictive of language performance in bilingual children than information about other risk factors (e.g., Abutbul-Oz & Armon-Lotem, 2022; Paradis et al., 2010), and not to increase discriminant accuracy in identifying LI in bilingual children (Boerma & Blom, 2017; Tuller et al., 2018). Boerma and Blom (2017), for example, found that family history distinguished typically developing children from children with DLD among monolingual children, but not among bilingual children: parents of monolingual children with LI reported family language problems more often than parents of bilingual children with LI.

Several questionnaires targeting bilingual children have incorporated *composite* risk factor indices, putting risk factors together in various combinations. Proposed indices vary widely in terms of which factors are included, the number of questions devoted to each factor, and the relative weight of the different factors (i.e., the proportion they represent in the overall score) (see Table 1). Regarding questions about the child's current language skills, the Alberta Language and Development Questionnaire (ALDeQ; Paradis et al., 2010) includes HL skills, but not SL skills;

however, there is a separate, additional question about HL attrition to modulate a low HL proficiency score, in order to take into account children for whom the SL has become dominant (at least in some linguistic domains). The Bilingual Parent Questionnaire (BIPAQ; Abutbul-Oz & Armon-Lotem, 2022) composite score includes both HL and SL skills scores, and, together, they constitute 71.4% of the total index. This requires high reliability of parental estimates in both languages, and indeed the authors noted that parents in their study rated their own SL skills on the higher end of the scale and that this is typical of heritage speakers in Israel. Finally, the Parents of Bilingual Children Questionnaire (PaBiQ) No Risk Index (Tuller, 2015) did not include questions about the child's current skills, in either the HL or the SL.[1]

Table 1

Risk Factor Composite Indices in Three Questionnaires: Number of Questions and Respective Weight (%) of Each Component

These indices have each been examined for predictive value. They have shown moderate to good diagnostic accuracy (as per reported ROC analyses) when used to distinguish groups of bilingual children with and without a diagnosis for language impairment, though sensitivity has generally been found to be lower than specificity. In other words, while they do well in correctly sorting children without a diagnosis, they fail to identify some children with a diagnosis (e.g., in Paradis et al., 2010, 34% were missed by the ALDeQ and in Boerma & Blom, 2017, 30% were missed by the PaBiQ). This is unsurprising, as indices are based on risk factors, meaning that they just indicate increased *likelihood* for language disorder. Moreover, discriminate accuracy obviously depends on how (well) children have been diagnosed in the first place. Abutbul-Oz and Armon-Lotem (2022) found the BIPAQ index, which included both HL and SL skills, to have good accuracy (85% sensitivity and 84% specificity) in distinguishing between clinically referred children who were later ascertained to have DLD and those later ascertained to be typically developing. Such indices have also been examined for how well they predict direct language measures on instruments designed for language assessment of bilingual children. For example, Tuller et al. (2018) found that the PaBiQ Positive Early Language Development index (consisting solely of the Early Milestones & Parental Concerns component) was the major predictor of performance on Language Impairment Testing in Multilingual Settings (LITMUS) Quasi-Universal Nonword Repetition (QU-NWR) and Sentences Repetition (SR) tasks administered in the SL.

We report here on how we derived composite risk factor scores designed for use as part of a triage procedure to be used by SLPs and teachers to optimize the screening process for language services, as well as by researchers seeking to characterize population samples. Mindful

[1] Many questionnaires and the literature surrounding them use the term "parental questionnaire". This term in fact clearly refers to a primary caregiver, whether that person be a parent, a foster parent, a grandparent, or any other adult who takes care of the child on a regular, sustained basis. The Q-BEx (see below) explicitly asks for the two primary caregivers of the child.

that triage needs to be expeditious, we sought to limit the number of questions needed to derive the index to those proven to be the most useful in most bilingual settings. We did not include questions about family history for developmental difficulties, because they can be difficult for some parents and studies of their predictive value have yielded mixed results (see above). On the other hand, in order to take into account potential HL-SL differentials such as that resulting from HL attrition (Paradis et al., 2010) or from recent exposure to the SL (Paradis et al., 2022), our proficiency measures included both HL and SL skills, to allow for calculation of the child's strongest language skills. This way of incorporating proficiency into a risk factor index could be particularly useful for inclusion of specific kinds of bilingual populations, such as refugee populations. Franck & Delage (2022) found that refugee status, along with Length of Exposure, was a significant predictor of a PaBiQ risk factor index which incorporated difficulties in the immediate family, but, crucially, did not incorporate proficiency. They suggested that late exposure to the SL, since it corresponds to late age of migration, could mean that caregivers' responses are biased by their perception of how vulnerable they feel their child is regarding language, with children migrating at an older age perceived as being more vulnerable. Refugee children have moreover been shown to test low on SL measures, compared to other bilingual children (Franck & Delage, 2022; Paradis et al., 20222). We reasoned that incorporating these children's much stronger HL skills in a composite risk factor index could provide a way of neutralizing parental response biases stemming from parental language worries related to migration.

Our study draws on a new parental questionnaire for gathering information on language experience, proficiency, and risk factors: the Quantifying Bilingual Language Experience questionnaire (Q-BEx – De Cat et al., 2025; https://www.q-bex.org/). Q-BEx is a modular, online, freely available parental questionnaire that presents many advantages compared to existing questionnaires used to quantify bilingual language experience (Q-BEx, 2024). It was designed following a comprehensive review of questionnaires that measure bilingual language experience (Kašćelan et al., 2021) and a Delphi Consensus survey of researchers, educators, and speech-language pathologists from 29 different countries (De Cat et al., 2023). It is built on best practice, incorporating questions and questionnaire techniques from previous questionnaires such as the PaBiQ and the ALDeQ, a critical review of current practice and reflections on practitioner needs and on how questions are worded and exemplified, to ensure maximal intelligibility. The risk factor questions include all and only questions that have demonstrated widespread usefulness. Q-BEx is filled in via an online platform (directly by the parent, or by the clinician/researcher in an interview), currently available in 33 languages, has a backend calculator that provides key information to the researcher, clinician, or teacher, and generates an individual language profile that summarizes the child's language experience, qualitatively and quantitatively.

The Q-BEx individual child report is designed to include a risk factor index, automatically generated from answers to specific Q-BEx questions and based on the considerations outlined above. This study reports on comparison of alternative risk factor composite scores. We first of all compared these composite scores to identify the one that best predicted developmental status

(TD/DLD) among bilingual children whose status had been independently ascertained. Then, we identified the optimal cut-off points for each index to effectively triage children into those needing specialist referral for LI assessment versus those not in need of such referral, and compared their over-all separability and their sensitivity. In a second set of analyses, we estimated the informativeness of the alternative risk factor indices as predictors of language proficiency (in the SL), as measured by LITMUS repetition tasks specifically designed for LI screening in bilingual children (LITMUS, 2024), in both the initial dataset and in a second dataset including both children in speech-language therapy (SLT) and all-comer children (i.e., children for whom language developmental status had not been assessed independently and is therefore uncertain). Using the optimal cut-off points identified in the first dataset, we then examined the distribution of at-risk children in monolingual versus bilingual children in the second dataset. Our expectations were the following: 1) given the demonstrated usefulness of parental questionnaire generated risk factor information, the optimal composite index would significantly predict children's known language development profiles and children's structural language measures, in bilingual and monolingual children, and 2) given that LI rates are not affected by bilingualism and that parents of bilingual children have been shown to be as able as parents of monolingual children to answer core risk factor questions, the proportion of all-comer children determined to be at risk for language impairment by our composite index should not be different for bilingual and monolingual children.

Method

Our study is informed by data from two questionnaires, PaBiQ and Q-BEx, which elicit largely similar risk factor questions. The first dataset is based on the PaBiQ and consists of bilingual participants whose clinical status is independently known (the Tours-PaBiQ dataset). The second dataset (the Q-BEx Validation Study dataset) is based on the Q-BEx and consists largely of all-comer children, regardless of any clinical status, known or unknown. Both datasets include a diverse group of families, in terms of HL and socioeconomic status (SES), reflecting bilingual populations in the relevant countries (France, the Netherlands, and the UK) and beyond.

Participants

The Tours-PaBiQ dataset consists of 109 bilingual families living in France. Children, all aged five to eight, were recruited in SLP private practices and in ordinary schools and were assessed with an in-depth evaluation of language skills in both their HL (Arabic, Portuguese, Turkish, or English) and the SL (French). Cut-offs for DLD status according to test norms were adapted to take into account language dominance via a procedure laid out in Tuller et al., 2018 (with a cut-off of -1.5 SD for child’s dominant language, -1.75 for each of equally dominant languages, and -

2.25 for weaker language—see Supplemental Material for complete procedure). As presented in Table 2, many of the children recruited in SLP practices and undergoing treatment for DLD tested as having typically developing (TD) language according to this procedure (24/52), generally because, although they scored below adapted norms when tested in French, their HL performance fell within adapted norms. One child recruited in schools and not undergoing SLT (1/57) was identified as DLD following this procedure. All children were included in the study, with their evidence-based profile determining their Bi-DLD/Bi-TD "final" status, resulting in 29 children labeled as likely to have DLD ("Bi-DLD") and 80 children likely to be TD ("Bi-TD"). Families varied in SES, as measured by the highest parental education level; more than half (58%) the families did not include a parent with post-secondary education.

Table 2

Participants in Tours-PaBiQ Dataset: Status at recruitment and status after testing in HL and SL ("final status")

The Q-BEx Validation Study dataset (see Table 3) included 278 monolingual and multilingual 5- to 8-year-old children and their caregivers taking part in the Q-BEx project validation study, in France (FR; 77 children), in the Netherlands (NL; 106 children), and in the United Kingdom (UK; 95 children). Reflecting the diversity observed in multilingual children in these and in many other countries, HLs varied widely: there were 15 different HLs among the children in France, with a predominance of Arabic (16 families), 24 different HLs in the Netherlands, with a predominance of Turkish, English, and German (11, 10, and 9 families, respectively), and 25 different HLs in the UK, with Urdu being the most frequent (9 families). Socio-economic status, measured by the highest level of education attained by either parent, varied within and between countries. At least one parent had a university degree in 43% of the families in France, 74% in the Netherlands, and 67% in the UK, while 26% in France, 13% in the Netherlands, and 15% in the UK had no postsecondary training.

Recruitment procedures differed across the three countries. Children in France were recruited in both schools and in SLP private practices; caregivers reported in the consent form if their child had previous or current SLT, and why (if known), resulting in 21 children with a current diagnosis of oral language impairment (and receiving SLT) and 16 children with some SLT (recruited in schools). In the NL, recruitment took place in schools and via social media advertisement; language disorder, as reported by caregivers or (remedial) teachers, was an exclusionary criterion. Finally, in the UK, recruitment was exclusively through schools, and children were genuine all-comers (no exclusionary criteria applied and no SLT information collected). We return to these differences in our analyses.

Table 3

Participants in Q-BEx Validation Study Dataset: Sample Characteristics by Country

Table S.1 provides a summary of known language development status in the two datasets.

This study was carried out in accordance with the recommendations of ethics committees governing research on human subjects in the respective universities: the University of Tours Comité d'éthique de la recherche Tours-Poitiers (n° CER-TP 2021-09-04), the Radburg University Ethics Assessment Committee of the Faculty of Arts and the Faculty of Philosophy, Theology and Religious Studies (no 2021-9263), and the Leeds University Faculty of Arts, Humanities & Communications Research Ethics Committee (FAHC 21-006 Amd2).Written informed consent was given by the caregivers of each participant.

Materials and procedure

The PaBiQ and Q-BEx questions that gathered information about early language milestones and eventual concerns about early language development are very similar (see Table S.2). These ask caregivers to estimate two language milestone ages, first word and first sentence, and to respond yes or no as to whether there had been any concerns about the child's early language development.

The child's proficiency in the HL and in the SL was based on four similarly-worded questions in the two questionnaires providing the following information for each of the child's languages (a) overall proficiency, (b) proficiency compared to other children, (c) difficulties with sentence formation, and (d) satisfaction with the child's language ability (see Table S.3). Whereas the PaBiQ, following the ALDeQ, is based on speaking skills only, the Q-BEx also asked about understanding. Furthermore, in the Q-BEx only, comparison children were specified as bilingual. In both questionnaires, parents responded by selecting a response on a four-point Likert scale. For example, the question "How well can your child speak [Language]?" had the following choices: hardly at all (0), not very well (1), pretty well (2), and very well (3).

As a first step in the creation of risk factor composite scores, we converted questionnaire responses to the questions about early language development and current proficiency to a common degree-of-risk scale, with 0 corresponding to a *zero risk for LI*, -1 a *moderate risk for LI*, and -2 a *high risk for LI*. Early developmental milestone answers were categorized following Paradis et al. 2010: for Age of First Words, 0 risk if the caregiver responded 15 months or younger, -1 if between 16 and 24 months, and -2 if 25 months or older. For Age of First Sentence, 0 if younger than 25 months, -1 if between 25 and 30 months, and -2 if 31 months or later. Early language concerns were assigned 0 for a "no" response and -1 for a "yes" response. The 4-point Likert scale responses to the HL and SL skills questions were converted such that any positive response (i.e., "very well" or "pretty well") was assigned 0, "not very well" was assigned -1 (moderate risk), and "hardly at all" was assigned -2 (high risk). For each of the four proficiency scores listed above (see also Table S.3), the score for the child's strongest language was retained. So, for example, a child with -2 for SL-speaking and -1 in HL-speaking would have a risk factor score for speaking of -1. For Q-BEx scores, we had calculated the average between speaking and understanding scores for each proficiency measure (see Table S.3).

The scores from these components were then combined to create several composite scores, presented in Table 4. Composite scores varied for whether a proficiency component was included, and for the number of questions considered in this component. The Risk Factor Index with three components, *RF3*, corresponds to the sum of the three Early Language Development scores (three questions total). *RF4* is based on four components: it combines information about the child's proficiency (strongest skills over the HL and the SL) into a single component (as a mean of the strongest proficiency scores), which was then added to the three Early Language Development scores, and thus the weight of the proficiency component was ¼ of the composite score. *RF7*, was obtained by adding up seven components, one for each of the three Early Language Development scores and each of the four proficiency scores. This effectively assigned more weight to proficiency (which is represented by 4/7 components). In addition, we compared three versions of RF4, according to the number of questions considered for the proficiency component—*RF4-min* includes only the strongest language skills for speaking (and thus a total of 2 proficiency questions, 1 for each language), *RF4-mid* the strongest language skills for both speaking and for understanding (4 questions, 2 for each language), [2] and *RF4-max* includes all four proficiency questions, each one declined for both speaking and understanding (16 proficiency questions in all, 4 questions, each one declined for speaking and understanding, and thus 8 for each language). In other words, we looked at both the quantity of information to be obtained from parents and the weight given to proficiency with respect to that given to Early Language Development.

Table 4

Alternative Risk Factor indices, RF3, RF4 (min, mid, max), and RF7: Weight of components and number of questions

The other questionnaire-based information considered in this study was related to known predictors of language proficiency in bilingual children, i.e., the cumulative amount of language exposure (i.e., the amount of month-equivalent exposure to each language over the child's lifetime, based on parental estimates for each language-defined period in the child's life—see De Cat et al., 2025), and the richness of language experience (based on the frequency of activities, the number of interlocutors, and the proficiency of these interlocutors – see Unsworth et al., 2025). We used these as control variables in our models, as explained below. (See the Supplemental Material for more information about these variables.)

---

[2] Since proficiency questions in the PaBiQ ask about speaking (See Table S3), but not about understanding, the PaBiQ has an RF4-min score, but no RF4-mid score.

Direct assessment of all child participants in both datasets included a number of SL measures, as well as measures of non-linguistic cognition. The SL outcome measures came from two tasks from the Language Impairment Testing in Multilingual Settings tool kit (LITMUS, 2024), specifically designed to assess aspects of structural language abilities in bilingual children: the LITMUS-Sentence Repetition task in the SL (either Dutch, English, or French) for evaluation of morphosyntax, and the LITMUS-Quasi-Universal Nonword Repetition task for evaluation of phonology. We also tested vocabulary breadth and working memory, with tests selected because of their availability in the three societal language and their widespread use for these constructs. For vocabulary breadth, we used relevant version of the PPVT-III: the PBVS (Dunn et al., 2009) for English, the PPVT-III-NL (Dunn et al., 2005) for Dutch, and the EVIP (Dunn et al., 1993) for French). Working memory was assessed in the SL via the Backward Digit Span task (using the WPPSI/WISC protocol (Wechsler, 2002, 2003) in the PaBiQ dataset and with Q-BEx participants who had been recruited via SLP clinics and using the protocol from Hill et al (2021) with all other Q-BEx participants).

The PaBiQ was administered to a caregiver in the form of an in-person interview, and all children were assessed in person, either in their clinical setting or at school. The Q-BEx was filled in online by all or nearly all participants in the UK and in the NL, but about half of the families in France received in-person assistance in filling in the questionnaire. Children were assessed in person, either at school, at home, in their clinical setting, or at a university lab; a few children in the NL and the UK were assessed online (see Kašćelan, submitted).

All analyses were done in R (Core Team, 2024; version 4.5.1). In the first set of analyses, we compared the informativeness of the alternative RF composite indices as predictors of TD versus DLD status in the Tours PaBiQ dataset. This was done, using R package *lme4* (version 1.1-37), by fitting generalized linear models predicting Diagnostic Status by each of the RF indices in turn, and comparing the goodness of fit of these models. We then carried out a Receiver Operating Characteristic (ROC) analysis, performed through a GLM procedure (Peppe, 2000), using the *pROC* package, version 1.19.0.1 (Robin et al, 2011). The Area Under the Curve (AUC) needed to be at least 0.70 to reach the range of values typically considered to be acceptable (de Hond et al., 2022). We used the Youden Index (YI) (Youden, 1950) to determine the optimal threshold (of cut-off score) for each index. YI is the point on the ROC curve that is farthest from the random chance line along the sensitivity axis (Sensitivity + (Specificity – 1)). It identifies the widest part of the curve, or the point which gives maximum true positive rate (or sensitivity) while maintaining the lowest false positive rate (FPR), i.e. maximum separation between TPR and FPR. Higher AUC and YI correspond to greater separability. In order to identify the optimal RF and its cut-off score, we compared the cut-off scores for their usefulness for purposes of triage—i.e., aiming to minimize false negatives, so as not to miss children who could possibly benefit from further assessment.

Thereafter, in a second set of analyses, we applied the optimal cut-off score to the Q-BEx dataset and compared the indices against known developmental profiles. Note that ROC analyses could not be performed on the Q-BEx data set. Empirically based developmental status equivalent to what was determined for PaBiQ participants (who were tested in both their HL and their SL, see above) was not available for Q-BEx participants. In keeping with our goal of maximum inclusion and representability of bilingual participants, we did not limit participation to specific HLs, and thus, as presented above, these were high in number (15 in France, 24 in the Netherlands, and 25 in the UK) and included a number of languages for which no standardized tests are available.

Finally, in a third set of analyses, we fitted logistic regression models to determine the significance of the alternative RF composite scores as predictors of LITMUS-QU-NWR and LITMUS-SR scores in all children in both datasets, and compared the models obtained for each RF using AIC weights (Burnham & Anderson, 2003). Akaike information criterion (AIC) weights are interpreted as the strength of evidence in favor of a model in a set of alternative models: they estimate the probability that a given model is the optimal one in the set. The amount of support for the other models in the set is based on their delta value, which corresponds to the difference in AIC between that model and the optimal one. The model fitting procedures for each dataset are explained in the Supplementary Materials.

Results

Risk Indices Applied to the PaBiQ Dataset

Each of the alternative Risk Factor indices (RF3, the relevant RF4 composite scores, and RF7, summarized above in Table 4) distinguished the PaBiQ dataset participants ascertained to have DLD from those having TD, at the group level, as seen in the violin plots in Figures 1, 2, 3, and 4. In these figures, we see that the 95% confidence intervals (boxes) around the means (horizontal bars) are distinct, in spite of the overlap in RF score range. That overlap is expected, given the nature of risk factor indices: the various components are all known to increase the likelihood, other things being equal, of a child having LI, but are incapable, individually or taken together, of diagnosing LI.

Figure 1

Applying the RF3 Index to the PaBiQ Dataset: Data spread for Bi-TD and Bi-DLD participants (horizontal bar = mean and box = 95% confidence interval)

Figure 2

Applying the RF4-min Index to the PaBiQ Dataset: Data spread for Bi-TD and Bi-DLD participants (horizontal bar = mean and box = 95% confidence interval)

Figure 3

Applying the RF4-max Index to the *Tours-PaBiQ Dataset: Data spread for Bi-TD and Bi-DLD participants (horizontal bar = mean and box = 95% confidence interval)*

Figure 4

Applying the RF7 Index to the PaBiQ Dataset: Data spread for Bi-TD and Bi-DLD participants (horizontal bar = mean and box = 95% confidence interval)

The alternative RF indices each significantly predicted Bi-DLD/Bi-TD status. A comparison of the models with each of these indices using AIC weights (Burnham & Anderson,

2003) showed that RF3, RF4-min, and RF4-max were equally informative as predictors of language development status (RF3 AIC=119.75, RF4-min AIC=119, RF4-max AIC=119.28). They were very marginally better than RF7 (AIC=120.52), with no statistically significant difference in inter-model goodness of fit.

ROC analyses were carried out for each RF index, Table 5 presents both the GLM-modeled threshold and AUC, and the corresponding raw score thresholds and sensitivity/specificity rates.

The AUC for RF3 was below 0.70, and thus below the range of values typically considered to be acceptable, while the others were acceptable. None of the inter-AUC differences were statistically significant (see Supplemental Material). The RF3 modelled score threshold corresponded to a raw score of 0, rendering this index entirely uninformative, as it amounts to flagging all children (100% sensitivity, but 0% specificity). The optimal threshold returned for RF4-min corresponded to a raw score threshold of -1, yielding 76% sensitivity and 55% specificity. The raw score threshold for RF4-max was -0.75 and this yielded a sensitivity of 76% and a specificity of 56%. For RF7, the ROC analysis yielded a raw score threshold of -2 (i.e., a score of -2 or below), which yielded sensitivity of 62% and specificity of 61%. RF4-max emerged as the index with the best sensitivity, while maintaining acceptable separability (how well the index, over-all, discriminates between participants ascertained to have DLD and those ascertained to have typical language development, as measured by the AUC, where a value higher than .70 is typically considered to be acceptable). Since our goal was not diagnosis but development of a first-level triage tool, we considered relatively low specificity less problematic. In other words, our priority was to minimize the risk of not flagging children likely to be in need of further testing. These considerations suggest a very slight advantage for RF4-max over the other indices, as it yielded the best balance between separability and sensitivity in the PaBiQ dataset in which children's developmental status was independently ascertained. We note that RF4-min was a very close contender.

Table 5

PaBiQ Dataset: GLM-modeled ROC analyses of alternative RF indices using Youden's Index and Corresponding RF raw score thresholds, sensitivity, and specificity

RF4-max Threshold Applied to the Q-BEx Validation Study Data

Applying now the RF4-max cutoff to the Q-BEx Validation Study dataset, including bilinguals and monolinguals, the results reveal several expected trends (Figure 5). The highest proportion of children identified as at-risk (68%) was found among the (French) participants with a current diagnosis for DLD or with some SLT. This proportion was significantly higher than the percentage of RF4-max-flagged FR participants with no reported SLT (20%; $X^2(1)$ = 17.759, p

<.001, V = .48). It was also significantly more than the percentage of RF4-max-flagged NL participants (22%; $X^2(1)$ = 25.878, p <.001), to which a language disorder exclusionary recruitment criterion had been applied (. It was also more than the percentage of children flagged by RF4-max among the UK all-comer participants (33%; $X^2(1)$ = 13.306, p <.001, V =.43).

Among the children without reported SLT (FR and NL), the proportion of flagged children was higher in bilinguals (29% and 30%) compared to monolinguals (11% and 8%), though this difference was significant only in the NL group (NL: $X^2(1)$ = 7.124, p = .008, V =.26; FR: $X^2(1)$ = 2.030 p = .154, V =.23). Among the children with a current diagnosis or some reported SLT, the proportion of flagged children was similar (FR, 67% and 69%). In the all-comer group (UK), children for whom no information about developmental status was available and thus at least some of them presumably were in SLT alongside many who were not, rates were similar in bilinguals and monolinguals (35% and 30%, UK). When the French children recruited in schools with reported SLT are added to the French children recruited in schools with no reported SLT, forming the FR all-comer group (n = 56; see Table 3), there was a higher rate of flagged children in the bilinguals (40%) compared to the monolinguals (19%), though this was not significant (FR, $X^2(1)$= 2.842, p = .092, V =.23).

Figure 5

Proportion of participants flagged by the RF4-max index: Q-BEx dataset, by country, by language development profile (FR only), including bilinguals and monolinguals

Plotting RF4-max indices against the known developmental profiles in the entire Q-BEx population sample also revealed expected trends. Illustrating with RF4-max (Figure 6), children with a current diagnosis tested in SLP private practice had the lowest RF4 scores (M = -1.81, SD 1.77), followed by children tested in school whose caregivers reported them having (had) SLT (M = -1.29, SD=1.74), and then by children tested in schools with no reported SLT (M=-0.44, SD=0.93). Inter-group differences were significant (H(2)=22.21, p < .001), and Dunn's post hoc comparisons revealed that RF4 scores for the Current.diagnosis group did not differ from RF4 scores for the Some.SLT group (p = .225), whereas the scores of children with no reported SLT were significantly higher than both Current.Diag scores (p < .001) and Some.SLT scores (p = .006).

Figure 6

Applying the RF4-max Threshold to the Q-BEx Dataset: Data spread for presumed-Bi-TD and presumed-Bi-DLD participants (vertical bar = median and box = interquartile range)

Summarizing, RF4-max flagged more of the children likely to have DLD (because of a current diagnosis or because of parental report of SLT for language) than children with no such presumption; moreover, children without such presumption had higher RF4-max scores. Bilingual children were flagged more frequently than monolingual children only in the groups in which children with language problems had been excluded.

Informativeness of each RF index as a predictor of SL language performance

We have seen that each of the Q-BEx-generated composite scores distinguished children according to their independently ascertained or presumed developmental status in our two datasets (DLD/TD or Current Diagnosis/Some SLT/Presumed TD). ROC analyses performed on the PaBiQ dataset, in which developmental status was independently ascertained, showed that RF4-min and RF4-max were superior to RF3 and RF7, in terms of combined separability and sensitivity, with RF4-max very marginally better (see Table 4 regarding the weight and detail of the proficiency component in each index). Application of RF4-max to the Q-BEx dataset showed expected trends when comparing children for likely developmental status.

We turn now to the relationship between each of these risk factor indices and the two measures of SL structural language proficiency designed to identify bilingual children at risk for DLD (LITMUS-QU-NWR for phonology and LITMUS-SR for morphosyntax), beginning with the Tours-PaBiQ dataset. We fitted two sets of regression models: one predicting the accuracy of identical repetitions in the SR task, and one predicting the accuracy score in the NWR task. In each set, we included either RF3, RF4-min, RF4-max, or RF7 as a predictor, alongside control variables for SL experience, SL vocabulary outcomes, and working memory for the models within each set, using the procedure described in the Supplemental Material. The results are summarized in Table 6: RF3, RF4-min, and RF4-max were almost equally informative as predictors of NWR and of SR, and more so than RF7. When predicting NWR outcomes, the models with RF3 and the two RF4 variants had a similar probability of being the best models in the set (i.e., 29%-36% chance, as shown in the *weight* column). When predicting SR outcomes, the model with RF3 was marginally better, although the difference in AIC with the two RF4 variants was smaller than 1. RF7 had considerably less support than RF3 and the two RF4 variants (dAIC is 4 or higher).

Table 6

PaBiQ dataset: Comparison of models containing alternative RF indices for predictive value for LITMUS-QU-NWR and for LITMUS-SR

The same procedure was employed for the Q-BEx data from all three countries (including this time RF4-mid). The results are summarized in Table 7. In predicting NWR outcome, RF4-min is clearly superior to RF3, RF4-max, RF7, and marginally better than even RF4-mid. In predicting SR outcome, RF4-mid is clearly superior to RF7, but only very marginally better than RF3, RF4-min, and RF4-max, which each fare better than RF7. RF4-mid had a statistically significant effect on sentence repetition accuracy (AME = 0.035, 95% CI [0.009, 0.061], $p = .009$). Each one-point increase in RF4-mid —indicating greater risk of atypical development— was associated with a 3.5 percentage point decrease in sentence repetition accuracy. Children with the highest risk profile (RF4-mid = -6) had an average sentence repetition accuracy that was approximately 21 percentage points lower than children with typical development (RF4-mid = 0). RF4-mid also had a statistically significant effect on non-word repetition accuracy (AME = 0.053, 95% CI [0.031, 0.076], $p < .001$). Each one-point increase in RF4-mid was associated with a 5.3 percentage point decrease in non-word repetition accuracy. Children with the highest risk profile (RF4-mid = -6) had an average non-word repetition accuracy that was approximately 32 percentage points lower than children with typical development (RF4-mid = 0).

Table 7

Q-BEx dataset: Comparison of models containing alternative RF indices for predictive value for NWR and SR scores

In sum, the predictive value of RF7 for both NWR and SR was inferior to the other indices, in both datasets. RF3 and the RF4 variants fared equally well for SL outcome predictive value in the PaBiQ dataset. In the Q-BEx dataset, RF4-min was clearly the best predictor of NWR, but RF3 and the three RF4 variants were basically equally informative for SR, with a very slight advantage for RF4-mid.

Discussion

We set out to create and compare several composite risk factor indices based on caregiver responses to a set of risk factor questions that are automatically derivable from Q-BEx, a free, customizable, online questionnaire designed to measure children's language experience . The aim was to determine the most useful and economical (in terms of amount of information required of caregivers) index for use to flag bilingual children likely to be at risk for LI. As such, this index could be used in first-level clinical triage and in characterizing research population samples, thereby contributing to more adequate inclusion, notably of bilingual children whose HL cannot be directly assessed. We compared indices for their ability to flag children according to known developmental profiles in two independent datasets, including 387 children tested in France, the

Netherlands, and the UK, whose HL and SES diversity reflect commonly observed population diversity in these and many other countries, to predict SL outcomes, and we furthermore compared flagging rates in bilingual participants to those in monolingual participants. Composite indices including a proficiency component covering skills in the child's strongest language weighted equally with each of three early language development questions (age of first word, age of first sentence, and parental concerns about early language development) had better diagnostic accuracy and was a better predictor of direct language measures compared to an alternative without a proficiency component and an alternative with a heavier weighted proficiency component.

The composite indices we considered differ from each other and from previously proposed composite risk factor indices in several ways. The critical difference between the indices we created was in whether and how parental estimates of the child's language proficiency were incorporated. RF3, like the composite index proposed for the PaBiQ (Tuller, 2015), did not include language proficiency at all. In RF7, as in the BIPAC (Abutbul-Oz & Armon-Lotem, 2022), language proficiency was the heaviest component, at 57% (4 out of 7 components). In the RF4 variants, the language proficiency component weighed 25%. The indices including parental estimates of the child's current language skills, instead of relying uniquely on measures regarding the HL, or on both the HL and the SL, were based on a component stemming from calculation of the child's strongest skills between the HL and the SL. In this way, any HL attrition, for example, can be counter-balanced by stronger skills in the SL.

Another inter-index difference was in the number of questions required to compute the index and the computation mode. Aware of the need for expeditious first-level triage and the need for a risk factor index to be maximally useful across cultural settings, we did not include in any of our indices any questions asking caregivers about family history for developmental difficulties, as these can be unreliable (Paradis et al., 2010). More generally, in keeping with one of the major goals of the Q-BEx project, we sought to minimize the number of questions so as to keep questionnaire administration time low, a fundamental consideration for clinical settings, and potentially for response reliability. RF4-max and RF7 are each based on 19 questions in all: 4 questions about current language skills, speaking and understanding, in the HL, and in the SL (= 16 questions in all, for bilinguals) and 3 early language development questions (age of first word, age of first sentence, and parental concerns about early language development). RF4-mid is based on seven questions, and RF3 (which has no proficiency component) on just three questions. In comparison, the composite indices in the ALDeQ (25 different questions) and the BIPAQ (23 different questions) are based on a greater number of questions asked of caregivers.

Each of the generated indices was a significant predictor of known language developmental status in both datasets, and each was also a significant predictor of SL structural language abilities, as measured by LITMUS-NWR and LITMUS-SR, tasks designed measure structure language abilities, while not penalizing for bilingualism. Inter-index comparisons based on Bi-DLD/Bi-TD separability and sensitivity for developmental status and then informativeness of SL outcome led to some clear and some less clear conclusions. RF7, the index built in which proficiency had the heaviest weight, was clearly the least useful index, based on both analyses, over both datasets—its sensitivity was unacceptably low, and models including this index were significantly less successful in predicting SL outcomes. Turning to the most minimalist of the indices, RF3, which includes no proficiency component at all, it did not fare so well either: the ROC-analysis-generated threshold yielded no separability at all, and its predictive value for SL outcome had different results in the two datasets (one of the equally most informative indices in the PaBiQ dataset, while in the Q-BEx dataset, it was not the best index for predicting either NWR or SR). Turning to the RF4 variants, in which the proficiency component weighed 25%, it is hard to distinguish between the contenders. While RF4-max had very slightly better sensitivity, the RF4 variants did not differ in informativeness for either SL outcome measure in the PaBiQ dataset, or for LITMUS-SR in the Q-BEx dataset, though RF4-min was the best predictor of LITMUS-NWR in that dataset (marginally better than RF4-mid).

Comparison of bilingual participants to monolingual participants, which we illustrated with RF4-max, the index that fared best in the ROC analyses, revealed a couple of interesting trends, which would need confirmation, as some numbers were small. In all-comer children (i.e., all children tested in schools, with no exclusionary criterion for language disorder), no significant differences in rates of flagging by RF4-max were found between bilinguals and monolinguals, suggesting that RF4-max works equally well for both groups of children, as risk factor rates are not expected to be influenced by bilingualism. No significant differences were found either between bilinguals and monolinguals in the French children with a current diagnosis or some SLT, though numbers were small. However, among children with no reported SLT or language disorder (in the French and NL population samples), the proportion of red-flagged children was significantly higher among the bilinguals compared to the monolinguals. This difference might at least in part be attributed to some under-diagnosis of LI (with bilinguals being more likely to be over-looked because of their bilingualism), meaning there could be more bilingual children likely to be at risk in these groups. Grimm & Schulz (2014), in a study conducted among German school children in the same age range as our study, found frequent under-diagnosis of LI, at similar rates in bilinguals and monolinguals. Under- and over-diagnosis of LI in bilingual children are known to vary considerably from country to country, and at different points in time (Armon-Lotem et al., 2015).

One of the major goals of the Q-BEx project was development of a parental questionnaire that is as efficient as possible, and that includes only questions that are useful. We therefore wanted to see if an efficient RF index really needed to include information on the child's proficiency, and, if so, how much and at what weight in the composite risk factor index. Our results suggest that while heavily weighting strongest language proficiency, such as RF7 does, may not be useful, inclusion of the child's strongest proficiency appears to provide helpful information. We observed, at the individual level, that there were at least some children in the PaBiQ dataset for whom caregivers indicated weak skills in their strongest language, even though they had not indicated any worrying early development factors. There were also a number of children (nearly a third) who had both low scores for early developmental measures and for proficiency, raising the question of whether such children are at greater risk for LI than children having low scores for one of these, a question we were unable to pursue due to the size of our population sample of children whose developmental status could be independently determined. Inclusion of at least a minimal strongest language proficiency component might also be more informative than a risk factor index not including any such information for children with strong HL skills, but for whom caregivers indicated high risk answers regarding early language development. This constellation has notably been observed for children from refugee populations. We might expect that children with this profile would be more accurately screened by one of the RF4 indices (compared to RF3), since these include the strongest proficiency component, which could mitigate potential parental perception of language vulnerability related to migration.

If these lines of reasoning are on the right track, and if the results from our two population samples can be extended, then the most widely useful risk factor index should include a strongest current language skills component, though the weight of this component should not exceed that of the other components, as shown by the low predictive power of the RF7 index. Our results furthermore suggest that this component could include just a single question (per language), about current speaking skills (as in RF4-min). Although the comparative results we have presented are based on two different datasets, they are still likely to be dependent on the proportion of children in these population samples with the different profiles mentioned above. For example, in the PaBiQ sample only 28% of all of the children had strong HL proficiency (while 50% had strong SL proficiency), and there were no more than a couple of refugee children. We note that the Q-BEx online platform is designed for automatic calculation of a risk factor index that is included in the downloadable individual child reports. That risk factor index is currently set for RF4-mid, as it corresponded to a midway point in terms of the number of needed questions; it is based on 7 questions, rather than 19 questions (RF4-max) or 3 questions (RF4-min). This setting can be modified as results of studies such as this one can be replicated such that the empirical base for a specific RF index becomes more robust.

Extending empirical coverage regarding these alternative risk factor indices therefore is not merely a question of increasing statistical power by inclusion of more participants. Rather, inclusion of participants residing in additional countries, with participants in different bilingual contexts, is needed. Countries vary considerably in terms of educational policy regarding both heritage language support and societal language support for minority-language-speaking students, as well as in societal integration of minority language speakers. Another limit of our study is its entirely cross-sectional nature. Pursuing these indices in a longitudinal study of both clinically referred and all-comer populations would provide additional ways of testing its usefulness. Abutbol-Oz & Armon-Lotem (2022) were able to confirm the clinical status of referred children three months after screening with the BIPAQ via expert SLP appraisal; however, the status of non-referred children was not confirmed. We believe that a complete longitudinal study design in which all children are directly assessed at least 12 months after initial assessment and all caregivers fill in the questionnaire again would allow for new ways of assessing how an optimal risk factor index might be improved.

Another limitation we acknowledge is related to the manner in which we converted questionnaire responses to a common degree-of-risk scale, with a 0 to -2 scale. Although based in part by previous practice and results (e.g., Paradis et al. 2010's ternary categorization of early developmental milestone answers), the degree-of-risk scale conversions were based on hypothesis, yielding either binary or ternary scales, and the proficiency component questions conflated to 0-risk both "very well" and "pretty well" parental responses. These conversions were not subject to empirical scrutiny in this study, although they have an automatic impact on risk factor index weighting. It would be fruitful for future research on broader datasets to employ an inductive, data-driven approach to the RF component scales, similar to what we have done in this study regarding risk factor component inclusion and relative weight.

In this study, we explored risk factor indices designed to be used in first-level triage for language impairment screening. Looking beyond first-level triage, based on a growing body of research results (see Prévost et al., submitted), second-level triage might usefully include LITMUS tools. While we believe that that avenue needs to be pursued, we also feel that work devoted to first-level triage, to meet the specific needs of research and clinical settings, can and should be pursued, in parallel.

Acknowledgements

The Q-BEx project was funded by the UK's Economic and Social Research Council (from October 2019 to September 2023 - ES/S010998/1). The Tours-PaBiQ study was funded by the French National Research Agency (ANR-12-FRAL-0014-01). We gratefully acknowledge the research assistants, the children, their families, and their teachers/clinicians for their participation in these studies.

Data Availability Statement

The data and scripts are available via the OSF:
https://osf.io/m7sbw/?view_only=094699182a974b218205b564428ba01c

References

Abutbul-Oz, H., & Armon-Lotem, S. (2022). Parent questionnaires in screening for developmental language disorder among bilingual children in speech and language clinics. *Frontiers in Education 7*. https://doi.org/10.3389/feduc.2022.846111

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.). (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment* (Vol. 13). Bristol, UK: Multilingual matters.

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism*, *14*(5), 489-511. https://doi.org/10.1080/13670050.2010.529102

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

Burnham, K., & Anderson, D. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Callu, D., Jacquier-Roux, M., Cusin, F., Giannopulu, I., & Dellatolas, G. (2003). Pertinence du repérage par les parents des retards de langage chez l'enfant entre quatre et six ans. *Archives de pédiatrie*, *10*(12), 1061-1067. https://doi.org/10.1016/j.arcped.2003.09.034

Condon, S. & Régnard, C. (2013, May 29) Héritages et pratiques linguistiques des descendants d'immigrés en France. *Hommes et migration, 1288.* http://hommesmigrations.revues.org.

De Cat, C., Gusnanto, A., Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., & Unsworth, S. (2025). How detailed do measures of bilingual language experience need to be? A cost-benefit analysis using the Q-BEx questionnaire. *Bilingualism: Language and Cognition.* Published online 2025:1-12. doi:10.1017/S1366728925100497.

De Cat, C., Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & Q-BEx Consortium. (2023). How to quantify bilingual experience? Findings from a Delphi consensus survey. *Bilingualism: Language and Cognition*, *26*(1), 112-124. https://doi.org/10.1017/S1366728922000359

De Cat, C., Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., & Unsworth, S. (2022). *Quantifying Bilingual EXperience (Q-BEx): questionnaire manual and documentation.* DOI 10.17605/OSF.IO/V7EC8

De Hond, A. A., Steyerberg, E. W., & Van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, *4*(12), e853-e855.

Dunn, L. M., Dunn, L. M, & Schlichting, L. (2005). Peabody Picture Vocabulary Test-III-NL. Lisse: Hartcourt Test Publishers.

Dunn, L. M., Dunn, D. M., Sewell, J., Styles, B., Brzyska, B., Shamsan, Y., et al. (2009). The British Picture Vocabulary Scale, 3rd Edition. London: GL Assessment.

Dunn, L. M., Thoriault-Whalen, C. M., & Dunn, L. M. (1993). Échelle de Vocabulaire en Images Peabody (EVIP). Toronto (Canada): Psychan.

Franck, J., & Delage, H. (2022). The interplay of emotions, Executive functions, memory and Language: challenges for refugee children. *Languages*, *7*(4), 309. doi.org/10.3390/languages7040309

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

Hill, L. J., Shire, K. A., Allen, R. J., Crossley, K., Wood, M. L., Mason, D., & Waterman, A. H. (2021). Large-scale assessment of 7-11-year-olds' cognitive and sensorimotor function within the Born in Bradford longitudinal birth cohort study. *Wellcome Open Res*, *6*, 53. https://doi.org/10.12688/wellcomeopenres.16429.2

Kašćelan, D., Unsworth, S., Gusnanto, A., Prévost, P., Serratrice, L., Tuller, L., & De Cat, C. (submitted). Putting the Q-BEx questionnaire to a quality test: Questionnaire design and quality checks.

Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & De Cat, C. (2022). A review of questionnaires quantifying bilingual experience in children: Do they document the same constructs? *Bilingualism: Language and Cognition*, *25*(1), 29-41.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703-735. https://doi.org/10.1177/0142723721106640

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

LITMUS. (2024, June 21). *LITMUS Tools*. https://www.bi-sli.org/litmus-tools

Ministère du travail, de la santé et des solidarités (2024). *Détecter les signes d'un développement inhabituel chez les enfants de moins de 7 ans Deuxième Edition*. (January 2024) Handicap.gouv.fr https://handicap.gouv.fr/nouvelle-version-du-guide-de-reperage-des-signes-inhabituels-de-developpement-chez-les-enfants-de, March 21.

National Health Service. (2024).

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of experimental child psychology*, *162*, 31-38. https://doi.org/10.1016/j.jecp.2017.04.017

Paradis, J., Emmerzael, K., & Duncan, T. S. (2010). Assessment of English language learners: Using parent report on first language development. *Journal of communication disorders*, *43*(6), 474-497. https://doi.org/10.1016/j.jcomdis.2010.01.002

Paradis, J., Genesee, F., & Crago, M. B. (2011). *Dual language development and disorders. A handbook on bilingualism and second language learning*, *2nd Edition*. Baltimore, ML: Paul H. Brookes Publishing.

Paradis, J., Schneider, P., & Duncan, T. S. (2013). Discriminating children with language impairment among English-language learners from diverse first-language backgrounds. *Journal of Speech, Language, and Hearing Research 56*, 971-981. https://doi.org/10.1044/1092-4388(2012/12-0050)

Paradis, J., Soto-Corominas, A., Vitoroulis, I., Al Janaideh, R., Chen, X., Gottardo, A., ... & Georgiades, K. (2022). The role of socioemotional wellbeing difficulties and adversity in the L2 acquisition of first-generation refugee children. *Bilingualism: Language and cognition*, *25*(5), 921-933. doi:10.1017/S136672892200030X

Pepe, M. S. (2000). An Interpretation for the ROC Curve and Inference Using GLM Procedures. *Biometrics*, *56*(2), 352–359.

Prévost, P., & Tuller, L. (2022). Bilingual language development in autism. *Linguistic Approaches to Bilingualism*, *12*(1), 1-32.

Pert, S. (2023). *Working with Children Experiencing Speech and Language Disorders in a Bilingual Context: A Home Language Approach*. Abingdon, UK: Routledge.

Pratt, A. S., Adams, A. M., Peña, E. D., & Bedore, L. M. (2022). Exploring the use of parent and teacher questionnaires to screen for language disorders in bilingual children. *Topics in Early Childhood Special Education*, *42*(1), 77-90. https://doi.org/10.1177/0271121420942308

Prévost, P., De Cat, C., Gusnanto, A., Kašćelan, D., Serratrice, L., Tuller, L., & Unsworth, S. (submitted). Predicting Accuracy on the LITMUS Sentence Repetition task.

Q-BEx (2024). Retrieved from https://www.q-bex.org/.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Raven, J., (2019) *Raven's 2 - Matrices Progressives - 2nde édition*. French adaptation ECPA. Pearson.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of speech, language, and hearing research*, *41*(6), 1398-1411. https://doi.org/10.1044/jslhr.4106.1398

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "ROCR: visualizing classifier performance in R." _Bioinformatics_, *21*(20), 7881. http://rocr.bioinf.mpi-sb.mpg.de.

Tuller, L. (2015). Clinical use of parental questionnaires in multilingual contexts. In S. Armon-Lotem, J. de Jong, J. & N. Meir, (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 299-328). Bristol, UK: Multilingual Matters.

Tuller, L., Abboud, L., Ferré, S., Fleckstein, A., Prévost, P., dos Santos, C., Scheidnes, & M., Zebib, R. (2015). Specific language impairment and bilingualism: Assembling the pieces. In Hamann, C. & Ruigendijk, E. (Eds.) *Language Acquisition and Development: Proceedings of GALA 2013* (pp. 533-567). Newcastle: Cambridge Scholars Publishing.

Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., ... & Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders*, *53*(4), 888-904. https://doi.org/10.1111/1460-6984.12397

Unsworth, S, Gusnanto, A., Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., & De Cat, C. (2025). Unpacking language richness as a predictor of bilingual children's language proficiency. *Journal of Child Language*. (Accepted for publication in August 2025.)

Wehbe, P. (2023, September 11-12). Behind the scenes of clinical language assessment, [Workshop Presentation], *Child Bilingualism Workshop*, Uppsala University, Uppsala, Sweden, 11-12 September.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV)*. Pearson.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence--Third Edition (WPPSI – III).* Pearson.

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical evaluation of language fundamentals–fifth edition (CELF-5)*. Pearson.

Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer* 3: 32–35. doi:10.1002/1097-0142(1950)3:1

Description of Supplemental Material

A Methods Supplement note details the "Procedure for estimation of likely Bi-TD/Bi-DLD status in PaBiQ dataset, as laid out in Tuller et al., 2018". **Table S.1** provides a summary of language developmental profiles in the two datasets. Early language development questions and Current skills/proficiency/satisfaction questions, as they appear in the PaBiQ and in the Q-BEx questionnaire, are provided in **Table S.2** and **Table S.3**. Confidence Intervals from De Long's test for correlated ROC curves, comparing RFs in the PaBiQ dataset are provided in **Table S.4**. The Models estimating the informativeness of each RF index as a predictor of SL language performance are provided for the PaBiQ dataset in **Table S.5**, and for the Q-BEx data set in **Table S.6** and **Table S.7**. **Figures S.1, S.2**, and **S.3** present the distribution of control variables included in the regression models.

Figure1_P_RF3_final_status

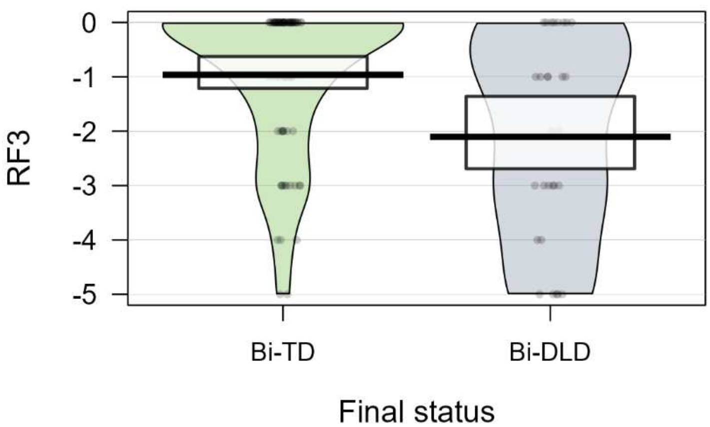


Figure2_P_RF4min_final_status

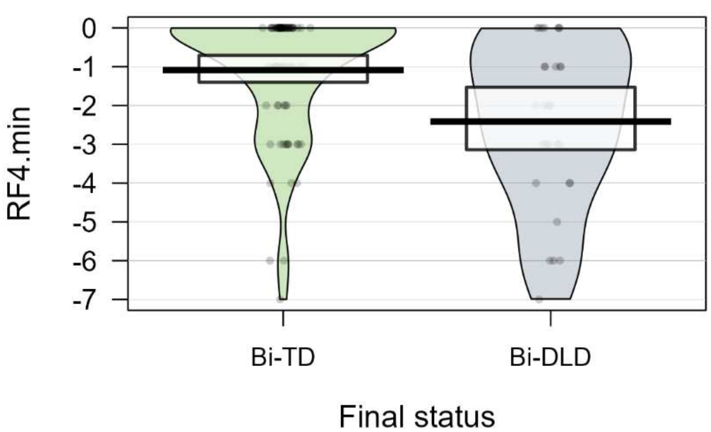


Figure3_P_RF4max_final_status

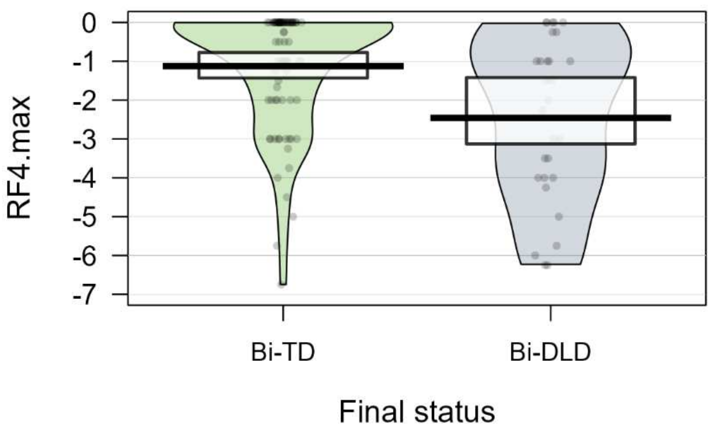


Figure4_P_RF7_final_status

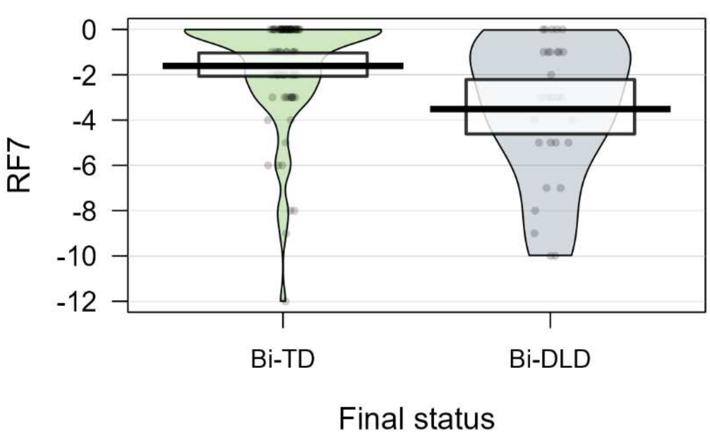


Figure5_RF4-Max-Flagging_Q-BEx

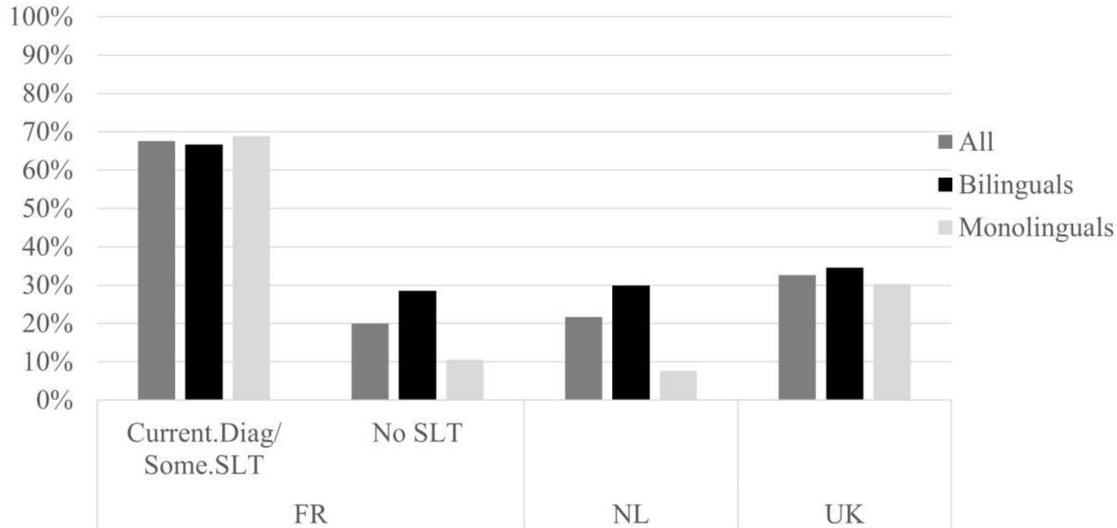


Figure6_Q_RF4_dev_profile

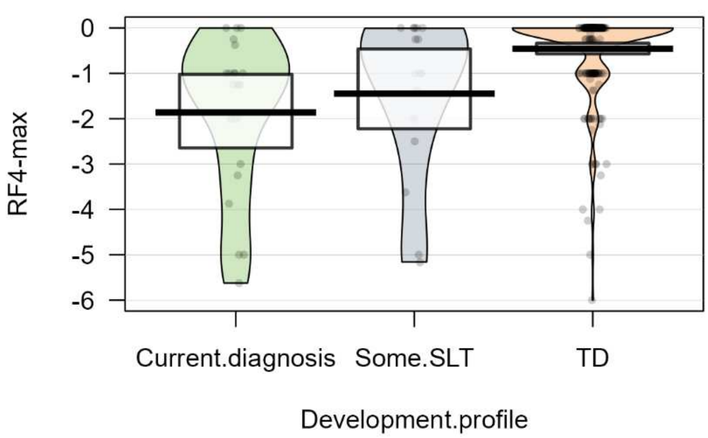


Table 1

Risk Factor Composite Indices in Three Questionnaires: Number of Questions and Respective Weight (%) of Each Component

	HL Skills	SL Skills	Early Milestones & Parental Concerns	Family History	Behavior Patterns & Activity Preferences	Total *n* Questions
ALDeQ	6 / 29%	0	4 / 29%	9 / 14%	6 / 29%	25
BIPAQ	7 / 33%	8 / 38%	7 / 25%	1 / 3%	0	23
PaBiQ	0	0	3 / 61%	3 / 39%	0	6

Note. HL = Home Language; SL = Societal Language.

Table 2

Participants in Tours-PaBiQ Dataset: Status at recruitment and status after testing in HL and SL ("final status")

	Final Status: Bi-TD (*n* = 80)	**Final Status: Bi-DLD** (*n* = 29)
Age (*M*, *SD*)	6;11 (1;1)	6 ;11 (1;0)
F / M (*n*)	34 F / 46 M	9 F / 20 M
Bilingual	73	29
Trilingual	7	0
Home Languages (*n*)	Arabic (27), Portuguese (18), Turkish (24), English (11)	Arabic (9), Portuguese (8), Turkish (9), English (3)
In SLT for DLD at recruitment	24	28

Note. Bi-TD = typically developing bilingual; Bi-DLD = bilinguals with Developmental Language Disorder; SLT = Speech-Language Therapy

Table 3

Participants in Q-BEx Validation Study Dataset: Sample Characteristics by Country

	FR (*n* = 77)	**NL** (*n* = 106)	**UK** (*n* = 95)
Age (*M*, *SD*)	6;10 (1;1)	6;10 (1;1)	6;11 (1;0)
F / M / Non-Binary (*n*)	30 / 47 / 0	50 / 55 / 1	47 / 48 / 0
Bilingual (*n*)	37	48	38
Trilingual (*n*)	5	19	14
Monolingual (*n*)	35	39	43
Home Languages (*n*)	15	24	25
Recruitment	21 SLP private practices 56 Schools	106 Schools, Social media	95 Schools
Known Lg. Develop. Profile	21 Current DLD Diagnosis, 16 Some SLT,[a] 40 No SLT	106 No Lg. Disorder (exclusionary criterion)	95 All-comers (no SLT information available, no exclusionary criterion applied)

Note. FR = France; NL = Netherlands; UK = United Kingdom.

[a]The reason for SLT, when provided by parents, varied considerably (8 expressive oral language, 2 reading, 2 unclear, 4 no reason given).

Table 4

Alternative Risk Factor Indices, RF3, RF4 (min, mid, max), and RF7: Weight of components and number of questions asked of caregivers

	RF3 (sum of 3 components)	RF4 (sum of 4 components)			RF7 (sum of 7 components)
		RF4-min	RF4-mid	RF4-max	
Age 1st Word	1/3	1/4	1/4	1/4	1/7
Age 1st Sentence	1/3	1/4	1/4	1/4	1/7
Parental concerns about early lg. development	1/3	1/4	1/4	1/4	1/7
Strongest lg. skills		1/4 Speaking only	1/4 Speaking + understanding		1/7
Strongest lg. skills with comparison				1/4[a]	1/7
Strongest sentence formation skills					1/7
Strongest parental satisfaction with lg. abilities					1/7
***N* Q-BEx questions for monolinguals**	3		5	11	11
***N* Q-BEx questions for bilinguals**	3	5	7	19	19

[a]The RF4-max proficiency component consists of a composite over the four strongest language scores (each one for speaking and understanding, and that weighs ¼ of the RF index.

Table 5

PaBiQ Dataset: GLM-modeled ROC analyses of alternative RF indices using Youden's Index and Corresponding RF raw score thresholds, sensitivity, and specificity

		RF3	RF4-min	RF4-max	RF7
GLM-modeled ROC analyses	Threshold (fitted value)	0.19	0.26	0.20	0.27
	AUC	0.68	0.70	0.71	0.71
Corresponding analysis over raw scores	RF Threshold (raw score)	0	-1	-0.75	-2
	RF Sensitivity	100%	76%	76%	62%
	RF Specificity	0%	55%	56%	61%

Table 6

PaBiQ dataset: Comparison of models containing alternative RF indices for predictive value for LITMUS-QU-NWR and for LITMUS-SR

Model	LITMUS-QU-NWR as outcome measure				LITMUS-SR as outcome measure			
	Resid. Df	**Resid. Dev**	**dAIC**	**weight**	**Resid. Df**	**Resid. Dev**	**dAIC**	**weight**
RF3	101	2.51	0.0	0.36	100	45159.73	0	0.42
RF4-min	101	2.52	0.4	0.29	100	45710.11	1.3	0.22
RF4-max	101	2.52	0.4	0.30	100	45529.36	0.9	0.27
RF7	101	2.60	4.0	0.05	100	46537.43	3.2	0.09

Table 7

Q-BEx dataset: Comparison of models containing alternative RF indices for predictive value for NWR and SR scores

Model	LITMUS-QU-NWR as outcome measure				LITMUS-SR as outcome measure			
	Resid. Df	logLik	dAIC	weight	Resid.Df	logLik	dAIC	weight
RF3	262	75.1247	4.0	0.07	240	299.6943	0.5	0.21
RF4-min	262	76.3189	0.0	0.55	240	299.8434	0.2	0.25
RF4-mid	262	76.3189	1.6	0.24	240	299.9492	0.0	0.28
RF4-max	262	75.4281	3.4	0.10	240	299.6014	0.7	0.19
RF7	262	74.3918	5.5	0.04	240	298.5494	2.8	.07

Supplementary materials

Control variables included in the models

Figure S.1: Questionnaire-based variables from the PaBiQ dataset

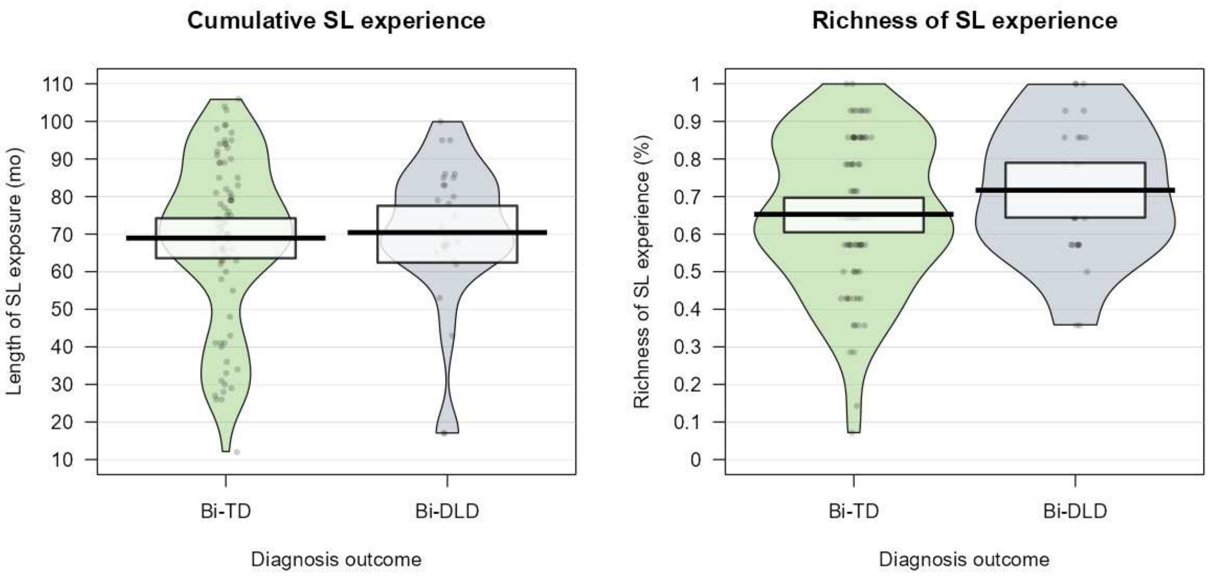


Figure S.2: Questionnaire-based variables from the Q-BEx dataset

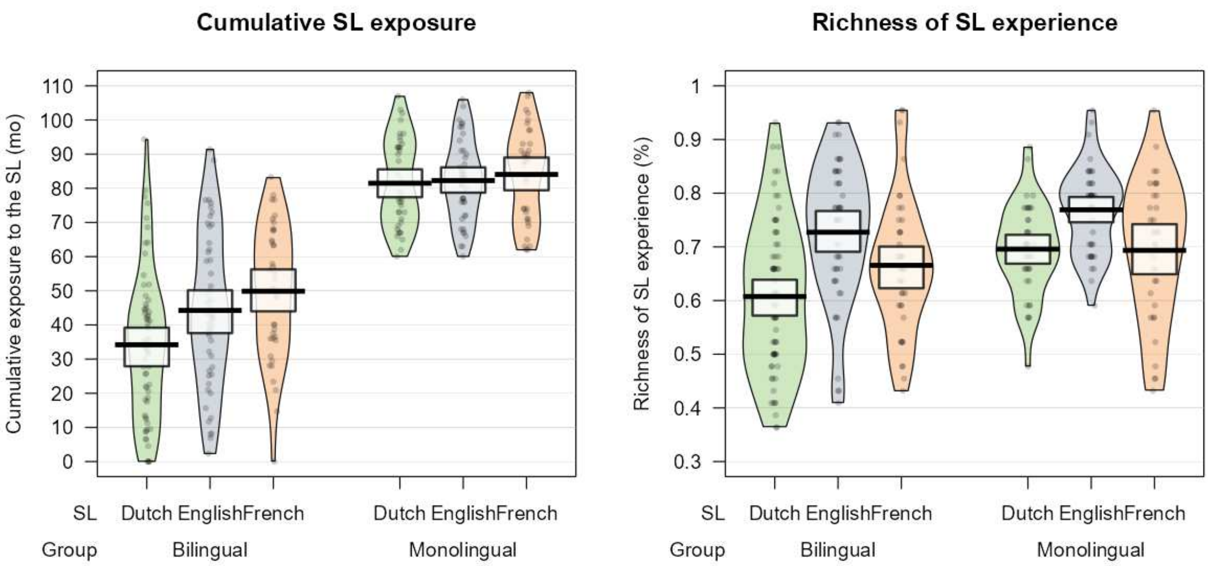
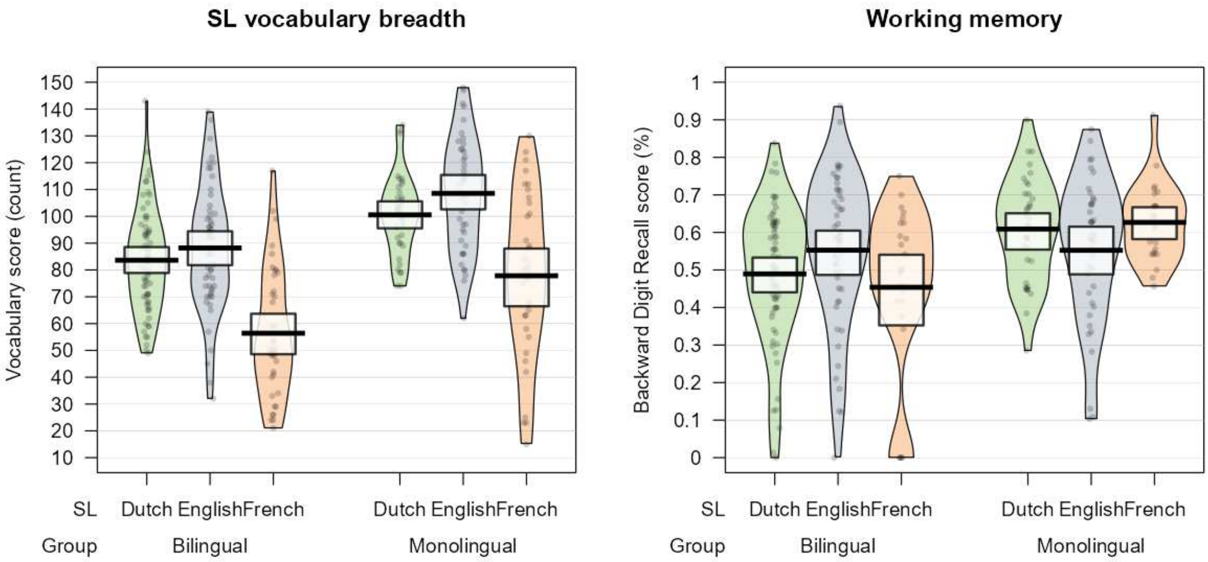


Figure S.3: Direct measures from the Q-BEx dataset



Comparison of alternative RFs in the ROC analysis

Table S.1: *Confidence Intervals from De Long's test for correlated ROC curves, comparing RFs in the PaBiQ dataset*

	CI low	**CI high**
RF3 vs RF4-min		
RF3 vs RF4-max	-0.05668131	0.01314683
RF3 vs RF7	-0.08492142	0.03880074
RF4-min vs RF4-max	-0.03779891	0.01754029
RF4-min vs RF7		
RF4-max vs RF7	-0.03254989	0.02996368

Models estimating the informativity of each RF index as a predictor of SL language performance

In the PaBiQ dataset

Starting with NWR accuracy as an outcome measure, we fitted models including (in turn) each of the alternative RFs and considering the following control variables: Length of SL Exposure, and SL Richness. The models were compared using likelihood ratio tests. The control variables were not included in the final model, as they did not improve model fit. The same procedure was followed for SR accuracy as an outcome measure. The optimal models are shown in Table S.2 for NWR and SR outcomes.

Table S.2: Regression models predicting SL outcomes in the PaBiQ dataset (illustrated here with RF4-min)

	NWR as outcome measure				**SR as outcome measure**			
	Estimate	**Std. Error**	***z* Value**	***Pr(>|z|)***	**Estimate**	**Std. Error**	***z* Value**	***Pr(>|z|)***
(Intercept)	0.7830	0.070	11.22	<.001	-45.816	14.17	-3.23	0.002
RF4-min	0.0345	0.008	3.92	<.001	2.666	1.282	2.08	0.040
SL Richness	0.0371	0.080	0.47	0.642	18.545	10.82	1.71	0.090
SL Exposure	0.0006	0.001	0.91	0.367	0.124	0.097	1.27	0.208
NWR					115.331	13.48	8.558	<.001

In the Q-BEx dataset

The same procedure was adopted to identify the optimal model for the Q-BEx dataset, considering the following control variables: Cumulative SL exposure, SL Richness, Vocabulary raw score (centred), and Backward Digit Recall. In the models predicting SR accuracy, we also included NWR accuracy as a control variable. For each SL outcome, we first identified the optimal model for the French part of the dataset as our training set, as it is the only one containing a sizeable proportion of children with an SLP diagnosis. That model was then fitted to the datasets from the UK and the Netherlands.

The optimal models predicting NWR accuracy and SR accuracy (using RF4-min) in the French dataset are shown in Table S.3. The optimal models predicting NWR accuracy and SR accuracy (using RF4-min) in the datasets from the UK and the Netherlands are shown in Table S.4. In the latter dataset, the effect of RF4-min becomes a significant predictor of NWR outcomes if BDS is removed from the model (beta = 0.14, p=.03).

Table S.3 *Q-BEx-FR dataset: Regression models for LITMUS-QU-NWR and LITMUS-SR, including RF4*

	NWR as outcome measure				SR as outcome measure			
	Estimate	**Std. Error**	***z* Value**	***Pr(>|z|)***	**Estimate**	**Std. Error**	***z* Value**	***Pr(>|z|)***
(Intercept)	-1.77	.777	-1.52	.130	-2.52	.896	-2.81	.005
RF4-min	.263	.093	2.85	.004	.278	.113	2.46	.014
NWR					2.781	.819	3.40	<.001
BDS	.254	.108	2.34	.019	.330	.125	2.64	.008
SL Vocab.	.015	.005	2.97	.003	.020	.007	3.06	.002
SL Richness	3.316	1.020	3.25	.001				
SL Exposure					.017	.007	2.46	.014

Table S.4 *Q-BEx-NL+UK dataset: regression models for NWR and for SR, including RF4*

	NWR as outcome measure				SR as outcome measure			
	Estimate	**Std. Error**	***z* Value**	***Pr(>|z|)***	**Estimate**	**Std. Error**	***z* Value**	***Pr(>|z|)***
(Intercept)	1.51	.397	3.80	<.001	-4.664	.591	-7.89	<.001
RF4-min	.116	.397	1.82	.069	.208	.091	2.27	.023
NWR					1.026	.377	2.72	.007
BDS	.156	.050	3.14	.002	.155	.064	2.40	.016
SL Vocab	.008	.004	2.21	.027	.039	.005	7.95	<.001
SL Richness	-1.97	.542	-3.64	<.001				
SL Exposure					.017	.003	5.16	<.001