How detailed do measures of bilingual language experience need to be? A cost-benefit analysis using the Q-BEx questionnaire*

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

[1]University of Leeds, UK
[2]University of Essex, UK
[3]University of Tours, FR
[4]University of Reading, UK
[5]Radboud University, NL

August 6, 2025

Address for correspondence:
Cécile De Cat
School of Languages, Cultures and Societies
University of Leeds
Woodhouse lane
Leeds LS2 9JT
UK
c.decat@leeds.ac.uk

*This project was supported by a grant from the Economic and Social Research Council (ES/S010998/1). We also gratefully acknowledge funding from the Harmonious Bilingualism Network. A big thank you to the many schools who enabled us to test children on their premises, to the parents and caregivers who patiently completed a very long questionnaire, and to the children to took part in the testing. We are grateful to the research assistants who contributed to the data collection and/or coding: Marieke van den Akker, Lotte Andersen, Zina Tamiatto, Laurent Lombard, and Laure Mallevays.

Abstract

What is the optimal level of questionnaire detail required to measure bilingual language experience? This empirical evaluation compares alternative measures of language exposure of varying cost (i.e., questionnaire detail) in terms of their performance as predictors of oral language outcomes. The alternative measures were derived from Q-BEx questionnaire data collected from a diverse sample of 121 heritage bilinguals (5- to 9-years of age) growing up in France, the Netherlands and the UK. Outcome data consisted of morphosyntax and vocabulary measures (in the societal language) and parental estimates of oral proficiency (in the heritage language). Statistical modelling exploited information theoretic and cross-validation approaches to identify the optimal language exposure measure. Optimal cost-benefit was achieved with cumulative exposure (for the societal language) and current exposure in the home (for the heritage language). The greatest level of questionnaire detail did not yield more reliable predictors of language outcomes.

Keywords: language exposure, background questionnaire, cross-validation, information theory

How detailed do measures of bilingual language experience need to be? A cost-benefit analysis using the Q-BEx questionnaire

1 Introduction

Bilingualism is a multi-faceted phenomenon, manifested through a myriad of individual differences in terms of age of onset, contexts of language exposure and use, quantity of exposure and use, language outcomes, attitudes, and changes over the lifetime. Bilingualism research spans many scientific disciplines, but even within each discipline, the tools used to document and quantify bilingualism vary widely (e.g., Kašćelan et al., 2021). In recent years, there have been many calls for greater comparability of methods to measure and document bilingualism (e.g., Byers-Heinlein et al., 2019; Rocha-Hidalgo and Barr, 2022). This is essential not only to improve research replicability, but also to facilitate exchange of information across sectors (notably with education professionals and speech & language therapists). The first step towards achieving comparability and replicability is to use the same tools (e.g., questionnaires) when gathering information which will subsequently be used to derive relevant variables (e.g., amount of exposure). Journals have started to publish methodological reviews to inform that debate (see e.g., Luk and Esposito, 2020). However, the focus tends to be on the qualitative evaluation of existing questionnaires (e.g., in terms of content overlap — Dass et al., 2024).

This paper addresses a related, but hitherto ignored methodological question: What is the optimal level of questionnaire detail required to derive empirically adequate measures of bilingualism? This question arises from two distinct problems: the quantity problem and the quality problem. It is notoriously difficult to get participants to fill in questionnaires, and completion time is often invoked as a major hurdle. This *quantity problem* not only results in missing data, but also potential bias in which participants actually complete the questionnaire. It may be, for example, that those who have the time and resources for such a task tend to come from more privileged backgrounds, resulting in sampling bias. The quality problem rests on the assumption that a greater level of questionnaire detail leads to greater precision of the estimates it can generate. We believe that if we ask more questions, or more precise questions, we will be able to obtain more informative and reliable predictors for the outcomes of interest in our study. However, that assumption remains untested.

To the extent that more questionnaire detail increases the quality of the information gathered, the quantity and quality problems overlap. Let's consider the concrete example of estimating a bilingual child's current exposure to their two languages. This can be informed by questions asking for global estimates (1), estimates by context (2-a), estimates for typical weeks vs. during holidays (2-b), or estimates by (group of) interlocutor(s) (2-c).

(1) How often do people talk to the child in each language overall?

(2) How often do people talk to the child in each language

a. at home / at school / in the local community?
b. during a typical week / during the weekend / during holidays?
c. depending on whether they are caregivers / siblings / extended family / peers / adults in the community?

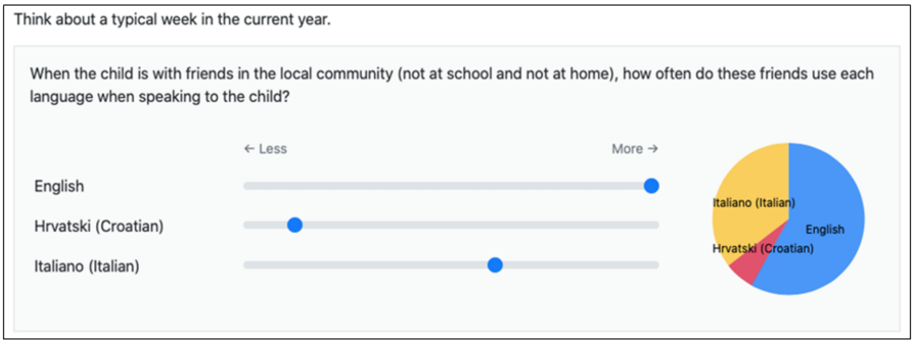
The estimates obtained from (2) could also be fine-tuned, by weighing them according to the amount of time the child spends in each context and/or with each (group of) interlocutor(s). And this could be estimated both for typical weeks and holidays.

In all cases, and even if this is not required explicitly, the questionnaire respondent is asked to recall what happens over a period of time (e.g., a typical week, the current year) and estimate the relevant average (whether global or by context). This is a complex cognitive task in terms of memory and numeracy. If recall or quantification are not accurate, there may be a large error margin in the data.

One approach to the quality problem could be to perform a psychometric evaluation of the trade-off between accuracy and error at different levels of question precision. This would require asking the same respondents to answer a set of questions about the same aspect(s) of the child's language experience (e.g., current language exposure), but with different levels of precision, as illustrated above. This could be repeated to compare responses within participant and estimate variations across sessions. By comparing the consistency of responses across questions and across sessions, we could obtain estimates of the level of error at each level of precision. Ideally this would be compared to data collected objectively by recording all the language interactions with the child (similar to the approach adopted by Verhoeven, Witteloostuijn, Oudgenoeg-Paz, and Blom 2024).

Here, we adopt an information-theoretic approach. Our aim is to evaluate the impact of different levels of questionnaire detail when documenting bilingual children's language experience. We therefore perform a 'cost-benefit analysis' to address the quantity and quality problems, by comparing the informativity of alternative measures of language exposure when used as predictors of language outcomes (within the same participants). These alternative measures vary in the amount of questionnaire detail and in the type of questionnaire information used to derive them. The trade-off to be considered is between the cost of obtaining these measures (based on the amount of respondent time required) and the benefit of these measures (based on how informative they are as predictors of language outcomes). Our aim is to make research-informed recommendations regarding the optimal level of questionnaire detail required to obtain informative and reliable measures of language experience for bilingual children.

This study focuses on Q-BEx (De Cat, Kašćelan, Prévost, Serratrice, Tuller, and Unsworth, 2022), a customisable online questionnaire to document the language experience of bilingual and trilingual children, currently available in 28 languages (with another 6 forthcoming - www.q-bex.org). Its design was informed by an international, cross-sector, Delphi consensus survey (De Cat, Kašćelan, Prévost, Serratrice, Tuller, Unsworth, and the Q-BEx consortium, 2023) to determine the aspects of bilingual experience it should document. The goal of Q-BEx is to enable maximal comparability across languages, children, and countries, in order to gain a better understanding of the role of language experience in the language development of bilingual children across these varying contexts. The design of Q-BEx is informed by the state of the art not only in terms of coverage, but also in its design features. The formulation of the questions and the formatting of the interface were based on recommendations from the psychometric literature (e.g., DeVellis 2017). This aimed to minimise the cognitive effort required from respondents, and to maximise the clarity and accessibility of the questions. For example, to answer questions probing the proportion of exposure to each language, respondents used sliders to adjust the sections of a pie chart, in which each language was represented in a different colour (see Figure 1). This avoided asking for explicit calculations, and it allowed the respondent to record their estimates precisely, while ensuring that the sum of proportions did not exceed 100% (representing the child's total language exposure). For a detailed description of the Q-BEx design (including the quality-control procedures implemented during the design process), see De Cat, Kašćelan, Prévost, Serratrice, Tuller, and Unsworth (2022).



Think about a typical week in the current year.

When the child is with friends in the local community (not at school and not at home), how often do these friends use each language when speaking to the child?

Figure 1: Intuitive estimates of language exposure

A key feature of Q-BEx is the possibility to implement the questionnaire components selectively, which results in substantial differences in completion time (from the respondents' point of view). The shortest version can take between 5 and 10 minutes to complete; the longest version can take up to an hour (or more, if the child's language experience varies substantially across contexts or over time).

The choice of language outcome measures and language exposure measures are presented and justified in the next section. As explained below, our information-theoretic investigations consist of two sets of analyses. First, we adopt the method outlined in Burnham and Anderson (2003) to identify the most informative and parsimonious model out of a set of candidates. Second, we adopt a cross-validation approach (Desmarais and Harden, 2014) to better take sampling effects into account and inform recommendations for other datasets. The implications regarding the optimal level of questionnaire detail are considered in the discussion, including specific recommendations for the customisation of Q-BEx.

2 Methods

The data informing this investigation were collected as part of a large-scale study aiming to validate the Q-BEx questionnaire. The design features of the questionnaire are presented after the Participants section. The language outcome measures are presented subsequently.

The study was approved by the ethics committee of each participating university : Tours (Comité d'éthique de la recherche Tours-Poitiers: 2021-09-04), Leeds (the Faculty of Arts, Humanities and Cultures Research Ethics Committee: 21-006 Amd2), and Nijmegen (Ethics Assessment Committee of the Faculty of Arts and the Faculty of Philosophy, Theology and Religious Studies: 2021-9263). Written informed consent was obtained from the parents of each child participant.

2.1 Participants

Participants were bilingual (n=88) and trilingual (n=29) children growing up in France (n=15), the Netherlands (n = 56), or the United Kingdom (n=46), aged between 5 and 9 years old (mean=83 months, sd=12). This dataset excludes children who did not have a full set of measures as required by our analyses. Recruitment targeted schools in deprived vs privileged areas and relied on existing participant databases and contacts from outreach activities. In France, assistants helped the parents fill in the questionnaire if they had no computer equipment or limited literacy. The original sample included children recruited via speech and language therapy clinics in France. These children were excluded from the present study. In the Netherlands, children with a known history of speech and language therapy (SLT) were excluded from recruitment. In the UK, no exclusion criterion was applied (i.e., the SLT status of children was unknown).

Children's language experience background was documented using the full version of the Q-BEx questionnaire. The only exception was that in France, the language mixing module was not implemented. Just under half of the participants (n=54) had been exposed to the Societal Language (SL: either French, Dutch, or English) from birth. For the rest (n=63), SL exposure started between 1 and 84 months of age. As shown the left pane of Figure 2, the amount

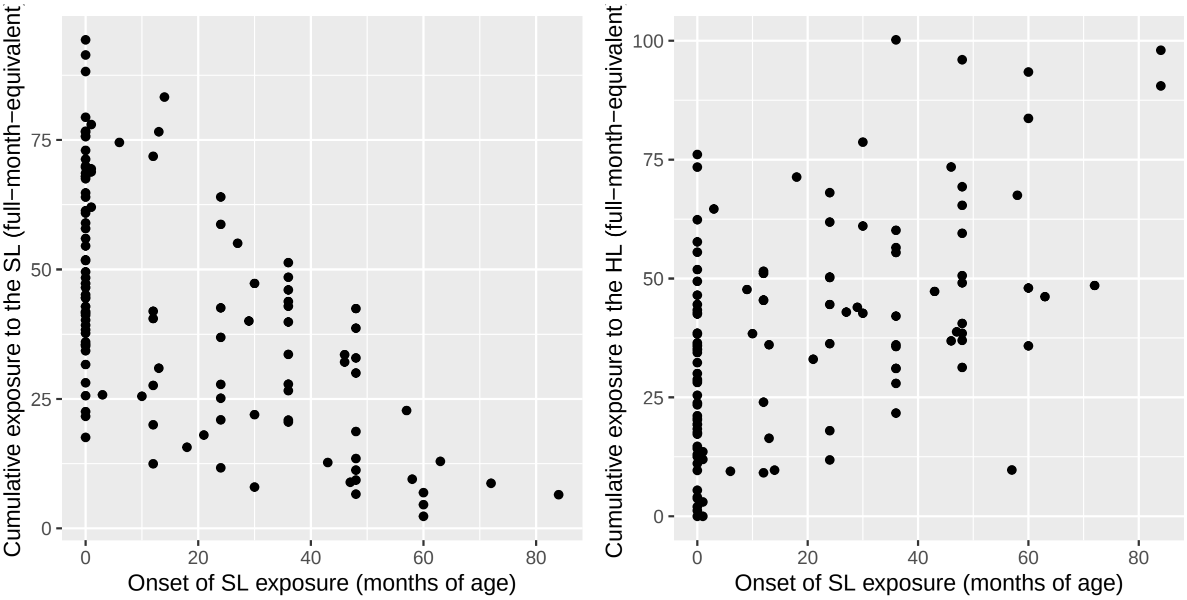


Figure 2: Correlation between onset of exposure and cumulative exposure in each language (left: SL; right: HL)

of cumulative experience to the SL varies substantially, both within the sequential and the simultaneous bilinguals. Relatedly, in the simultaneous bilinguals, the correlation between (biological) age and cumulative SL exposure is relatively weak (r=0.36, p=0.01).

All children were exposed to a Heritage Language (HL) from birth. In this study, we only consider children's strongest HL in the case of trilinguals. We use the term ”bilingual” to refer to both bilingual and trilingual children, and the term HL to designate the child's strongest or unique HL. Across our sample, these HLs were: Afrikaans, Arabic, Bangla, Bulgarian, Chinese, Creole, Czech, English, French, Frisian, German, Greek, Gujarati, Hindi, Hindko, Italian, Japanese, Kurdish, Latvian, Lithuanian, Lingala, Manouche, Mirpuri, Polish, Portuguese, Punjabi, Romanian, Russian, Spanish, Telugu, Turkish, Ukrainian, and Urdu. The number of participants per HL per country is given in Table 6 in the Supplementary Materials. Another aspect of the diversity of our sample can be seen in the right pane of Figure 2, which shows the correlation between onset of SL exposure and cumulative experience in the (main) HL.

Two thirds (66%) of participants grew up in homes where the highest level of caregiver education was university level. A sixth (15%) did not have post-secondary school education.

In sum, our sample features substantial variability in terms of age of onset of bilingualism, languages experienced in the home and in the society, and the amount of experience in the

HL and the SL. There is also variability in the highest education level achieved by caregivers across their languages.

2.2 The Q-BEx questionnaire

We asked children's caregivers to complete the full version of Q-BEx, including all components from all modules (i.e., background information, risk factors, language exposure and use, language proficiency, richness, attitudes, language mixing).[1] It took between 15 and 130 minutes to complete (35 minutes on average).[2]

In this study, we focus on the Q-BEx module with the highest impact on questionnaire completion time, i.e. language exposure and use.[3] However, we also consider information from other Q-BEx modules, when they are relevant covariates alongside language exposure, in models predicting language outcomes (as explained in the section about language outcome measures). Table 7 in the Supplementary Materials lists these covariates and identifies the Q-BEx module from which the relevant information was elicited.

2.3 Alternative estimates of language exposure

Table 1 lists the alternative estimates of language exposure considered in this study. For each, it specifies how the estimate is calculated, and on the basis of what information from the questionnaire. Whenever they apply to both languages, the estimates are calculated in the same way for both the SL and the HL. However, some estimates are only considered with respect to the HL: we assume that exposure in the home and the number of interlocutors are likely to be reliable predictors for the HL but not for the SL. Indeed, for many heritage bilinguals, the home is where most of their HL experience takes place, and the number of HL interlocutors can be very small. HL exposure during the holidays seems to be particularly important for

[1] In the French arm of the study, the language mixing module was not included.

[2] This excludes two outliers, who took 183 and 213 minutes respectively, and may therefore have had interruptions while completing the questionnaire.

[3] Here, we limit ourselves to language exposure. It is strongly correlated with language use: In our population sample, the correlation between SL exposure and use is r=.93 for current estimates and r=.79 for cumulative estimates. Consequently, we expect that the results will be similar for the two dimensions. For an investigation of the impact of "passivity of language experience" (which we define as the extent to which the child uses the language less than they are exposed to it) on heritage language outcomes, see De Cat, Gusnanto, et al., in preparation.

HL acquisition and maintenance (as shown for example by Kubota and Rothman 2024). It is currently unclear whether the same is true of the SL.

Both for current exposure estimates and for cumulative estimates, the order of presentation in the table (from top to bottom) broadly aligns with a decrease in the level of questionnaire detail required. This corresponds to the “cost” of collecting that information in terms of respondent time. This cost can be interpreted as the “complexity” of the measure, in terms of information quantity. Although we administered the longest version of the questionnaire, the alternative measures we compare are proxies for what would be obtained with a shorter version of the questionnaire.

In summary, we compared six estimates of SL exposure (as alternative predictors of morphosyntax and vocabulary outcomes), and eight estimates of HL exposure (as alternative predictors of HL outcomes). Based on questionnaire completion time, the most costly estimate for each language was current exposure adjusted (i.e., *weighted*) for the amount of time the child spends with each (group of) interlocutor(s) in each context. We considered four alternative estimates of current exposure: (i) the weighted estimates (defined above), (ii) their non-weighted counterpart, (iii) a global estimate (calculated as an average by context),[4] and (iv) a global estimate of exposure during holidays. In the case of the HL, we additionally considered (v) a global estimate of current exposure in the home. Furthermore, for both languages, we included (vi) an estimate of cumulative exposure (consisting in the sum of global estimates for each respondent-defined period in the child’s life). We also considered (vii) the age of SL onset as a coarse estimate of cumulative exposure to each language (indicating the length of SL exposure, or the length of any monolingual exposure to the HL). Finally, (viii) the number of HL interlocutors was considered as a coarse approximation for HL exposure.

[4] This global estimate can now be obtained from one question per context, if the short version of the language exposure and use module is implemented. This option wasn’t available at the time of our data collection. We calculated this estimate manually, by averaging by context.

Measure	How the measure is calculated	Information required from the Q-BEx questionnaire	Calculated for
Current exposure			
weighted measures	Proportion of exposure to each language from each (group of) interlocutor(s), adjusted for the amount of time spent with these interlocutors during a typical year, estimated from patterns in typical week days, weekends, and holidays. Automatically calculated by Q-BEx.	(1) For each interlocutor in the home and group of interlocutor outside the home: proportion of exposure to each language. (2) Timetable data: amount of time spent daily with each (group of) interlocutor(s), during typical week days, a typical weekend, and during holidays.	SL and HL
unweighted measures	Average proportion of exposure to each language calculated across the following contexts: home, school, community. Automatically calculated by Q-BEx.	For each interlocutor in the home and group of interlocutor outside the home: proportion of exposure to each language. (Alternative: the short version of this questionnaire module now allows for this information to be asked globally for each context.)	SL and HL
global estimate from the latest language period	Proportion of exposure to each language in the current language period, as defined by the respondent. This period varies in length, depending on the circumstances of the child. If the child experienced no substantial change in patterns of language experience, the latest period of language exposure is calculated from birth. No calcuation required.	Language period data: (1) ages at which the child's language experience changed substantially, (2) for each period defined by these ages: proportion of exposure to each language.	SL and HL
exposure during holidays	Proportion of exposure to each language during the holidays. No calcuation required.	Question included in either the long version (which asks about interlocutors in the home individually) or the short version (which asks about exposure globally by context).	SL and HL
average across interlocutors in the home	Proportion of exposure to each language in the home, averaged across interlocutors (irrespective of the amount of time spent with each). Calculated via R script.	Question included in either the long version (which asks about interlocutors in the home individually) or the short version (which asks about exposure globally by context).	HL

Measure	How the measure is calculated	Information required from the Q-BEx questionnaire	Calculated for
number of HL interlocutors	Approximate number of interlocutors in the HL	5-point Likert scale estimate	HL
Cumulative exposure			
full-months equivalent of exposure to each language	The respondent defines language periods (based on relevant changes in the child's life). For each period, they are asked to specify the proportion of exposure of each language. The duration of the period in month is multiplied by that proportion, resulting in full-month-equivalent measures for each language. The cumulative exposure to each language is calculated by adding the values obtained for each period.	Language period data: (1) ages at which the child's language experience changed substantially, (2) for each period defined by these ages: proportion of exposure to each language.	SL and HL
age of onset of SL exposure	Age of onset of substantial exposure to the SL, expressed in months.	Age in months (when people started addressing the child in the SL)	SL
length of HL monolingual period	Age of onset of substantial exposure to the SL, expressed in months. (Interpreted as the length of monolingualism in the HL, if any.)	Age in months (when people started addressing the child in the SL)	HL

Table 1: Questionnaire detail and calculation for each of the alternative measures of language exposure

2.4 Language outcome measures

Objective measures of proficiency in the SL (French, Dutch, or English) were obtained for morphosyntax and vocabulary[5], as there is ample evidence that they are affected by language exposure (e.g., Hoff 2003; Thordardottir 2011; Verhagen et al. 2024).

Morphosyntactic abilities in the SL were measured using the LITMUS Sentence Repetition Task (SRT; Marinis and Armon-Lotem, 2015). The reliability of this methodology is well established (see, e.g., the scoping review by Rujas et al. 2021). Repeating a sentence taps into underlying linguistic knowledge in order to parse the stimulus sentence and then reconstruct its structure and meaning. If the morphosyntactic properties featured in the sentence have not been acquired or are too complex to compute, performance is negatively affected (Marinis and Armon-Lotem, 2015; Polišenská, Chiat, and Roy, 2015).

In the French version, the SRT consisted of 16 sentences varying in complexity, from less (short sentences in present simple) to more (object relative clauses) complex. In the English and the Dutch versions, the SRT consisted of 30 sentences. They were ordered in two blocks (of 16 and 14 items, respectively) of identical complexity, so that the first block would be as close and equivalent to the French version as possible. No significant difference in accuracy was observed between the two blocks (see Prévost et al. under review). In all three versions, the sentences were presented auditorily using headphones in a fixed order. Response accuracy was scored in three different ways (eliciting 3 scores), as defined in (3):

(3) a. **Identical repetition score**: each item was given 1 point if the child had produced a verbatim repetition of the target sentence and 0 points otherwise.

b. **Target repetition score**: each item was given 1 point if the target structure was repeated correctly (even if other parts of the sentence were not repeated verbatim), and 0 points otherwise.

c. **Grammatical attempt score**: each item was given 1 point if the sentence pro-

[5] We also assessed phonology, with the LITMUS Quasi-Universal Non-Word Repetition task (Marinis and Armon-Lotem, 2015). We do not include this as an outcome measure in this study, as typically-developing children in this age range are expected to perform at or close to ceiling (given that the task is quasi-universal, i.e., designed not to be dependent on knowledge of a particular language), and the children suspected of atypical development were excluded from the analyses below. For an investigation of NWR outcomes in our entire sample, see De Cat, Tuller, et al. (under review).

duced by the child was grammatical (irrespective of whether they repeated the target structure), and 0 points otherwise.

This yielded three outcome measures, which we transformed into percentages (i.e., the proportion of accurate repetitions out of the total number of items).

Vocabulary breadth, which corresponds to a receptive measure of vocabulary size, was assessed using the receptive Peabody Picture Vocabulary Task: BPVS-3 for English (Lloyd Dunn, Leota Dunn, Sewell, et al., 2009), EVIP for French (Lloyd Dunn, Thoriault-Whalen, and Leota Dunn, 1993), and PPVT-III-NL for Dutch (Lloyd Dunn, Leota Dunn, and Schlichting, 2005). In this task, children are presented with four pictures, they hear a single word and are asked to point to the corresponding picture. Administration followed the instructions in the manual, starting at the age-appropriate starting set and moving up (and if necessary, down) until the required number of errors was met. The outcome measure is the total number of correct responses (i.e., the raw score).[6]

Receptive vocabulary depth, which corresponds to how well words are known, was assessed using the Word Classes sub-test of the CELF-5 in English (Semel, Wiig, and Secord, 2017) and French (Wiig, Semel, and Wayne, 2019), and CELF-4 for Dutch (Kort, Schittekatte, and Compaan, 2008). In this task, children hear words and are asked to indicate which words belong together. As the task progresses, the number of words from which children need to make a selection increases from three to four and visual support in the form of pictures is removed. Administration followed the instructions in the manual until children reached the end or failed to provide a correct response to the required number of consecutive items (four in English, five in Dutch and French). The proportion of correct responses out of the total number of items answered was included as dependent variable in the analyses.

Subjective measures of proficiency outcomes in the HL were derived from parental estimates (based on questions from the Q-BEX module on language proficiency). The questions asked how well the child speaks and how well they understand the HL for their age. We did not include reading and writing abilities, as many children hadn't learned to read yet. Answer

[6] Standard scores are inaccurate for bilingual children given that they are not adjusted for reduced experience in the SL.

options included: hardly at all / not very well / pretty well / very well.

2.5 Cognitive measures

To include as control variables in the models predicting language outcomes, we collected measures of non-verbal intelligence, short-term memory, and working memory. Non-verbal intelligence was assessed with the Matrices task from the WISC–V (Wechsler, 2014) or the WPPSI (Wechsler, 2013) (the latter for children under 6 years of age).

Short-term memory was assessed through Forward Digit Recall (FDR); working memory was assessed through Backward Digit Recall (BDR). Both tasks were administered through Psychopy, according to the protocol described in Hill et al. (2021). Children were presented auditorily with sequences of numbers (through headphones) and asked to repeat these numbers either in the same order (in the FDR task) or in reverse order (in the BDR task). The length of the sequence increased by one digit after 4 trials, starting with 3 digits in the first block of the FDR task, and 2 digits in the first block of the BDR task. The maximum sequence length was 6 digits in the FDR task, and 5 digits in the BDR task.

2.6 Modelling procedures

We adopted a predictive modelling approach throughout this study. This determines the baseline assumptions for the two sets of analyses reported below. Predictive modelling aims to identify the smallest set of predictor variables required for a model to achieve the highest level of generalisability in relation to an outcome variable. It seeks to maximise the variance explained by the model, while avoiding over-fitting the data. It does not aim to interpret the individual effect of a particular variable. Our quest for the optimal estimate of language exposure (as predictor of language proficiency) is therefore situated in the context of a previously-identified optimal set of predictor variables. In other words, the control variables included in each model were determined by the optimal predictive model we identified in previous work for each of the outcome variables (in the same group of children): We refer the reader to Prévost et al. (under review) for SL morphosyntax outcomes, Serratrice et al. (under review) for SL vocabulary outcomes, and De Cat, Gusnanto, et al. (in preparation) for outcomes of oral proficiency in the HL. The optimal set of control variables retained in each model is listed in the Supplementary Materials. Across models, this included the following aspects: SL (coinciding with country of residence), parental education, cognitive measures, phonological competence, attitudes towards the SL, early language development milestones and concerns, diversity and richness of the language environment and experience. The type of regression depended on the distribution of the outcome variable. We used beta regressions for percent scores (SRT and vocabulary depth), linear regression for numeric scores (vocabulary breadth), and cumulative link models for ordinal scores (parental estimates of HL proficiency).

For the first set of analyses, we use the information-theoretic approach to model selection proposed by Burnham and Anderson (2003). Rather than selecting a single ”best” model, this approach considers a set of models, and makes inferences to determine the relative support for each model in the set. It thereby acknowledges the unavoidable uncertainty in model selection. Akaike weights, derived from the AIC, are calculated for each model, balancing model fit against model complexity to identify the most parsimonious model. The weights add up to 1 (i.e., 100%) across the models in the set. A model's weight can be interpreted as the probability (expressed as a percentage) that a given model is the best in the set, as will be explained in detail in the Results section. This method provides a rigorous framework to compare alternative predictors within an otherwise identical model (including the same set of covariates), by focusing on the strength of the evidence rather than arbitrary significance thresholds. In that framework, differences in model parsimony can be interpreted as differences in the parsimony of the alternative predictors (as everything else is held constant).

For each language outcome measure, a series of models are fitted treating that measure as outcome variable, and considering in each model one of the alternative measures of language exposure as predictor, together with a set of control variables. The control variables included in each model were determined by the optimal predictive model we identified in previous work for the outcome variable in question (in the same group of children), as explained under Modelling procedures.

For the second set of analyses, we use cross-validation (Desmarais and Harden, 2014) to assesses the true predictive performance of each alternative measure of language exposure (in relation to language outcomes) and estimate the risk of over-fitting. The data is split (by participants) into an estimation set and a validation set (with the same response variable and predictor variables in both sets). We employ k=5-fold cross-validation where observations are randomly split into 5 different groups and 4 of them (80%) are assigned as estimation set and the other (20%) as validation set. Model fitting is then performed (5 times) on the estimation set to obtain parameter estimates for each of the language experience measure. The parameter estimates obtained from the estimation set are then used to make 'new sample' prediction in the validation set using the same predictors.

In the validation set, therefore, we have a 'new sample' prediction on the outcome variable, in addition to the observed outcome variable itself. We calculate the average of squared differences between the two as the root mean square error (RMSE) of prediction. RMSE is interpreted as the predictive ability of the model. The lower the RMSE, the better the predictive ability of the model. This allows us to assess the extent to which the analysis of the validation set could generalise to an independent dataset including different children.

For each language outcome measure, cross-validations are performed on models using, in turn, each of the alternative measures of language experience as well as the same control variables as in the first set of analyses. In addition to RMSE, we calculate the range of RMSE across the five different folds to assess whether the calculation of RMSE is stable across different folds. Furthermore, to measure the uncertainty in the calculation of RMSE, we calculate the confidence interval of RMSE using the *bootstrap* method (Efron and Tibshirani, 1986; Efron and Tibshirani, 1994). The bootstrap method is implemented by resampling with replacement observations in the estimation set to obtain bootstrapped RMSEs in the validation set. We can therefore calculate the 95% confidence interval from the 2.5 and 97.5 percentiles of the distribution of bootstrapped RMSE's.

According to the cross-validation analysis, the optimal model (i.e., the one containing the optimal estimate of language exposure) is the one that has the lowest RMSE, and that shows the greatest stability of RMSE across folds. Furthermore, the confidence intervals indicate RMSE precision and allow comparison between models.[7] Overlapping CIs across models

[7] The comparison of CIs can only be done across models predicting the same outcome variable.

would suggest that improvements are modest.

When evaluating cost/benefit overall (across the two sets of analyses), the optimal measure of language experience (as predictor of language outcome) should be the one that requires the least questionnaire detail while yielding the best predictive performance, i.e., the smallest error size across folds.

3 Results

As a preliminary observation, we note that the alternative estimates of language experience are all correlated, for both the SL (Figure 5 in the Supplementary Materials) and the HL (Figure 6 in the Supplementary Materials). Most of the correlations are strong (ranging from $r = .52$ to $r = .81$), apart from those with the number of interlocutors in the HL matrix. All correlations are significant in both matrices ($p < .001$).

3.1 Information-theoretic approach

3.1.1 Predicting outcomes in the SL

The information-theoretic approach aims to guide the choice between alternative language exposure measures as predictors of sentence repetition scores in the SL. The summary of results is presented in Table 2 (for target repetitions). Similar findings were obtained for identical repetitions and grammatical attempts (see Table 9 and Table 10 in the Supplementary Materials). In each table, the ”Model” column specifies the language exposure variable included in the model; the ”weight” column specifies the probability that the model in question is the best in the set; and the ”dAIC” column specifies the difference in AIC between the model in question and the best one in the set.

For all three outcome measures from the sentence repetition task, Age of SL Onset is uncontroversially more informative as a predictor: in each table, the model including it has the highest possible probability of being the best one in the set.

The model comparisons focusing on vocabulary as outcomes measures (involving each time a different language exposure predictor, as above), are summarised in the Supplementary

Table 2: Comparison of models with different language exposure predictors of Target Repetitions (SRT)

dAIC	weight	Model
17.9	0.00	Current.period.exposure
12.3	0.00	Cumulative.exposure
11.5	0.00	Current.exposure.non.weighted
19.9	0.00	Holiday.exposure
18.4	0.00	Current.exposure.weighted
0.0	0.99	Age.SL.onset

Materials in Table 11 for vocabulary breadth and Table 12 for vocabulary depth.

Here, the models including Age of SL Onset as predictor no longer have the highest probability of being the best in the set: rather, cumulative exposure to the SL is superior in both models, albeit only very marginally so in the models where vocabulary depth is the outcome measure.

3.1.2 Predicting outcomes in the HL

Table 3 summarises the model comparison for alternative measures of exposure to the HL, as predictors of speaking proficiency in the HL.

Table 3: Comparison of models predicting oral proficiency in the HL

dAIC	weight	Model
30.6	0.00	Number.of.HL.interlocutors
20.4	0.00	Age.SL.onset
17.2	0.00	Holiday.exposure
17.2	0.00	Current.period.exposure
0.0	0.78	Current.home.exposure
12.7	0.00	Current.exposure.non.weighted
2.6	0.21	Current.exposure.weighted
23.6	0.00	Cumulative.exposure

Here, the model with the highest probability of being the best one in the set is the one with Current HL Exposure in the Home as predictor.

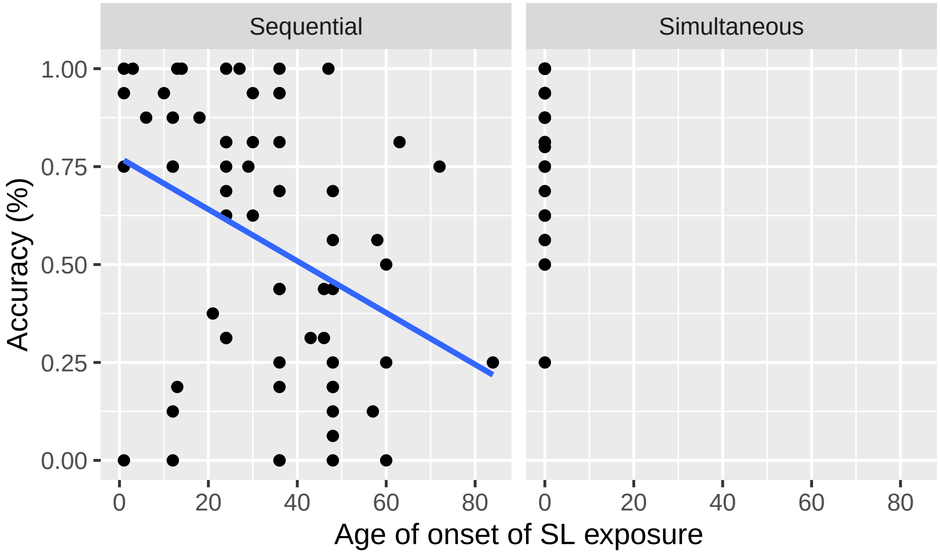


Figure 3: Correlation between Age of SL onset and Sentence Repetition outcomes (target structures)

3.1.3 Sensitivity to sampling effects

The strong preference for Age of SL Onset as predictor of morphosyntax outcomes in the SL is surprising, as it only correlates with SL outcomes in the sequential bilinguals and cannot account for the variability of outcomes in the simultaneous bilinguals (illustrated in Figure 3). The impact of Age of SL Onset might be due to the presence of children with very late onset of SL exposure in our sample.

As a first exploration of how variations in length of SL exposure might have affected the results, we refitted the models predicting the identical repetition score, excluding children with a late onset of exposure to the SL. We repeated the procedure three times, excluding first the children with onset of SL exposure above 4 years of age (48 months), then at or above 3;6 years of age (42 months), then at or above 3 years of age (36 months), as shown in Figure 4. These three thresholds have been proposed in the literature as cutoff points to distinguish late bilinguals from early bilinguals (see Schulz and Grimm 2019 for a review). The model comparisons at each iteration are shown in Tables 13-15 in the Supplementary Materials. We found that the probability of the model with Age of SL Onset being the best in the set decreased as the sample was restricted to children with lower ages of onset. If children with an Age of SL Onset above 36 months are excluded, Cumulative Exposure has a greater likelihood of being the most informative predictor than Age of SL Onset.

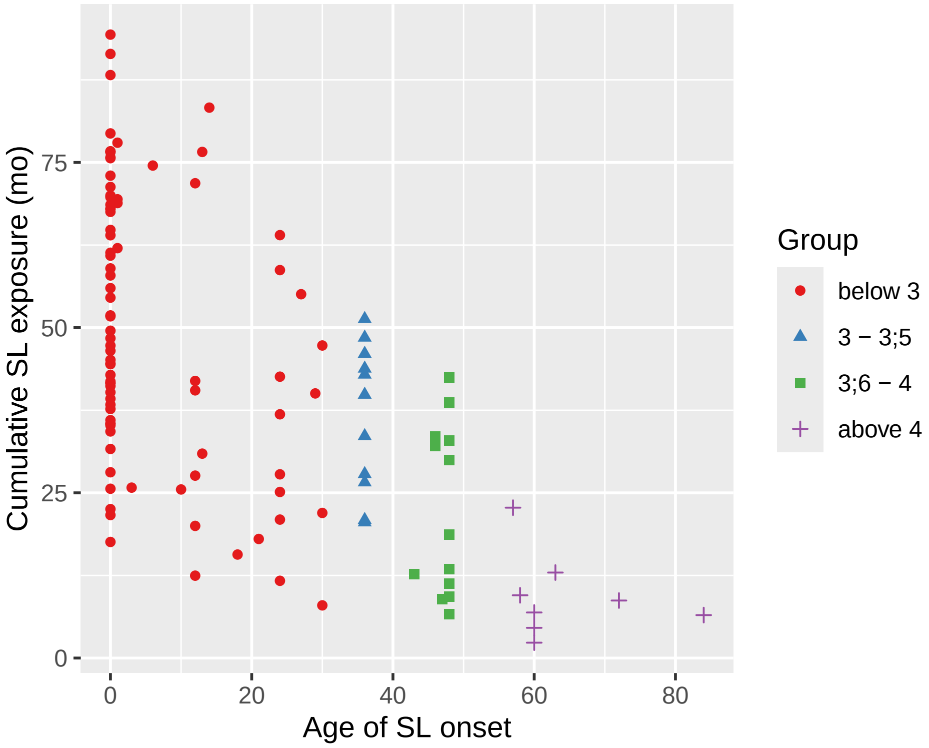


Figure 4: Exploratory grouping of children according to bands of SL onset

This suggests that model selection (from among variants each including a different language exposure predictor) is prone to sampling effects. We therefore turn to cross-validation: a method of informing model selection that is more robust to overfitting.

3.2 Cross-validation approach

We start with the cross-validation analyses comparing models predicting SL outcomes. The results for morphosyntax (based on the accurate repetitions of target structures in the sentence repetition task) are summarised in Table 4. The results for the other measures of SL outcomes can be found in in the Supplementary Materials: see Tables 16 and 17 for the other morphosyntax outcomes and Tables 18 and 19 for vocabulary outcomes.

For each alternative measure of language exposure (used in turn as predictor in the model), listed in the first column, we report the Root Mean Square of the Error between predicted and actual values, averaged across the five folds of data (in column ‘Mean error’). We also report the range of RMSE across folds, to estimate the stability of the model across subsets of data,[8] as well as the Confidence Interval boundaries (at 2.5% and 97.5% respectively). In each table, the 'Baseline model' does not include any of the alternative measures of language exposure. Note that the model includes all the other control variables, and that some of these reflect other aspects of language exposure (a point we come back to in the Discussion).

Table 4: Cross-validation comparisons with Target Repetitions as outcome measure

	Mean error	Range	CI 2.5%	CI 97.5%
Baseline.model	0.2304	0.1164	0.1807	0.3152
Current.period.exposure	0.2261	0.1223	0.1757	0.3157
Current.exposure.non.weighted	0.2244	0.0969	0.1819	0.3146
Current.exposure.weighted	0.2282	0.1158	0.1793	0.3153
Holiday.exposure	0.2340	0.1118	0.1837	0.3171
Age.SL.onset	0.2185	0.0735	0.1718	0.2949
Cumulative.exposure	0.2280	0.1016	0.1761	0.3041

The patterns of results observed in models predicting SL proficiency are as follows:[9] (1) Age of SL Onset yields the lowest mean error for most outcome measures (with the exception of Vocabulary breadth). (2) Cumulative Exposure tends to show smaller ranges across outcomes, suggesting more stable predictions. (3) The CIs overlap across models, suggesting that the improvement of better performing models is only modest. In particular, there is always CI overlap between models with Age of SL Onset and those with Cumulative Exposure. Note that in each analysis, the CIs of the baseline model overlap with those of the other models. This is because the baseline model includes control variables that are associated with the amount of language exposure, such as estimates of the quality of language exposure. Here, we adopt the conservative approach of focusing on the impact of language exposure while controlling for the impact of these other predictors.

We now turn to the cross-validation analyses comparing models predicting HL outcomes.[10]

[8] The range of errors was calculated by subtracting the smallest RMSE value from the largest RMSE obtained across folds for a given model. Please note that RMSEP values can only be interpreted relative to each other, and not in absolute sense. However, if the model prediction error (i.e., RMSE) is smaller than the natural spread of our data (i.e., the SD of the outcome measure), the model has good predictive value. This was the case in all our models.

[9] Note that the Mean error cannot be compared across tables, as they are based on data measured on different scales, analysed with different types of statistical models.

[10] Four of the control variables in the baseline models were too unevenly distributed to meet the requirements

As seen in Table 20 in the Supplementary Materials, Current Home Exposure is the strongest predictor, yielding the minimum error. The range for this variable is relatively low. It is closely followed by Exposure in the Current Period. Age of SL onset yields a much higher mean error. The weighted measure of current exposure does not fare better than its non-weighted alternative. As observed for SL outcomes, the CIs overlap, indicating modest differences between the alternative predictors.

4 Discussion

We performed a cost-benefit analysis to assess the optimal degree of questionnaire detail required to obtain informative predictors of language outcomes in bilingual children. Optimality was defined in terms of *quantity* as the minimum amount of information required to obtain the most informative predictor, and in terms of *quality* as the balance between precision and error at higher levels of questionnaire detail. We evaluated this through an empirical comparison of alternative predictors derived from language exposure estimates requiring a different amount of questionnaire-based information.

The alternative predictors were informed by Q-BEx, a customisable online questionnaire which can be implemented at various levels of detail (and hence questionnaire completion time). This study was based on data from 121 bilingual or trilingual children between the ages of 5 and 9, obtained with the longest version of the questionnaire. Using different subsets of measures (simulating implementation of the questionnaire at different levels of detail), we generated six alternative estimates of exposure to the Societal Language (SL) and eight alternative estimates of exposure to the Heritage Language (HL). These alternative estimates of language experience differed not only in levels of granularity but also in the type of information they are based on. For each language, all the estimates of language experience showed a consistent pattern of significant inter-correlations, corroborating their treatment as alternative

for cross-validation sampling. As all these control variables represented aspects of the "richness" of the language experience, we replaced them with the composite variable for the Richness of the language experience which is automatically calculated by Q-BEx - see Unsworth et al. (under review) for a validation of that composite Richness variable. We also performed an alternative cross-validation analysis on simplified alternatives of the original control variables (reducing them each from five levels to two or three) instead of using the Richness composite measure. The results for that analysis are provided in the Supplementary Materials. The two approaches yielded highly consistent results.

measures of the same latent variable.

The informativity of alternative predictors was assessed according to information theory, in two sets of analyses. The first one followed the approach of Burnham and Anderson (2003), which aims to determine the likelihood that a particular model is the best-fitting among a set of alternatives. It can also determine how closely competitor models perform, compared to the optimal model. Age of SL onset far outperformed the alternative (language exposure) predictors of sentence repetition accuracy in the SL. Cumulative exposure clearly outperformed the other predictors of SL vocabulary breadth. In the case of SL vocabulary depth, the performance of all models was quite close, with Cumulative exposure coming out as most informative (closely followed by Age of SL onset). Among language experience predictors of HL outcomes, current measures were the most informative. Current exposure in the home yielded the best fit, with the non-weighted estimate of current exposure a close contender. These results are summarised in Table 5.

Language outcome	Most informative measure
Morphosyntax in the SL	Age of SL onset
Vocabulary breadth in the SL	Cumulative SL exposure
Vocabulary depth in the SL	Cumulative SL exposure
Parental estimate of HL proficiency	Current HL exposure in the home

Table 5: Most informative predictors of language outcomes, based on AIC comparisons

We conclude from this first set of analyses that outcomes in the SL are best predicted by language exposure estimates that take into account the child's experience over their lifetime. By contrast, outcomes in the HL are best predicted by current estimates. Both for SL and for HL, the most informative estimate of language exposure is not costly to obtain (as it requires little questionnaire information): Age of SL onset (as predictor of SL morphosyntax), or Current HL exposure in the home (as predictor of HL outcomes).[11] However, Age of SL onset cannot predict variability among simultaneous bilinguals (given that it is zero for all of them). The informativity of Age of SL onset as predictor of sentence repetition accuracy does indeed diminish (in favour of Cumulative exposure), as "late SL starters" get progressively excluded

[11] In this study, we calculated the current HL exposure in the home as the average across all interlocutors in the home. It is now possible to elicit this information from a single question, using the short version of the Language Experience module in Q-BEx.

from the sample. In other words, the extent to which Age of SL onset is the best predictor depends on the extent to which it varies amongst participants.

To objectively assess the impact of sampling effects, we carried out a second set of analyses, using a cross-validation approach (Desmarais and Harden, 2014). This approach asks whether the predictions obtained from a randomly defined (80%) subset of the data generalise to the remainder (20%) of the data. For each of the alternative measures of language exposure, the procedure was repeated five times (in five *folds*), using different sub-samples. The magnitude of the error between the predicted parameter estimates (based on the training set) and the observed estimates (from the testing set) revealed the extent to which the predictor of interest is affected by sampling effects.

The cross-validation analysis revealed that, in this sample (characterised by a particular age range and variability in Age of SL exposure), Age of SL exposure was the most effective estimate of SL language exposure in terms of cost-benefit. This was the case even though the sample includes simultaneous bilinguals (for whom Age of SL onset is not an informative predictor, as it is identical for all these children), because late Age of SL onset has such a strong (negative) effect on SL outcomes on a (relatively small) group of children in the population sample. Cumulative SL exposure is less prone to variability in prediction error across the data subsets (i.e., folds). Cumulative SL exposure is informative for more children within the sample (compared with Age of SL Onset): it can vary in simultaneous bilinguals, and is strongly associated with Age of SL onset in sequential bilinguals.[12]

When it comes to HL outcomes, the cross-validation analysis identifies Current exposure in the home as the optimal predictor. This also happens to be the simplest direct estimate of current HL exposure (based on interactions in the home, to the exclusion of other contexts). Age of SL onset (which is equivalent to the length of any monolingual period in the HL) had middling performance, suggesting it was not among the strongest predictors. Contrary to Kubota and Rothman (2024), HL exposure during the holidays performed relatively

[12] Another advantage of the Cumulative Estimates sub-module of Q-BEx is that it provides data to cross-check the answers to the Onset of Exposure questions, which were sometimes misinterpreted by respondents in our study: some parents seemed to assume that language exposure doesn't count until the child starts producing language, and responded that exposure to any language (HL or SL) started when their child was 1 (or sometimes 2) years of age. See Kašćelan et al. (in preparation) for an explanation of how we corrected these implausible responses based on information provided via the Cumulative Exposure sub-module.

poorly. We interpret this as indicating that the reliability of this predictor of HL outcomes is strongly affected by the properties of the sample. For example, we speculate that factors affecting whether the family can holiday in a country where the HL is spoken may act as a moderator.

The answer to the quantity problem needs to be nuanced: the simplest measures can go a long way, depending on the properties of the population sample. The highly informative value of Age of SL onset in our sample is likely to be due to the magnitude of its effect in ”late” bilinguals in this age group (as they will not have accumulated much SL experience at the point of testing): if ”late-starters” are removed, Age of SL Onset is superseded by Cumulative Exposure. We conclude that Age of SL Onset is only informative to the extent that it is a reliable proxy for cumulative exposure (two highly correlated measures in our sample). A direct measure of cumulative exposure is therefore a “safer” option, because it is less susceptible to sampling effects.

Various studies have debated whether language proficiency is best predicted by current or by cumulative estimates of exposure (see, e.g., Cohen 2016; Unsworth 2013). Our results suggest that their effect is mediated by language status: in heritage bilinguals (who are exposed to their HL from birth but tend to become dominant in the SL, which is the school language), current exposure (especially in the home) is more informative as a predictor of HL outcomes, but cumulative is more informative as a predictor of SL outcomes. A limitation of our study is that it only included bilingual children growing up in countries where the education system was monolingual and delivered exclusively in the SL. Further research will be required to investigate whether home exposure to the HL remains the most informative predictor in children who are educated bilingually, or in the HL rather than the SL. Another question for further investigation is to determine what level of cumulative exposure in the SL is indicative of SL proficiency within the range found in monolingual children. We report on such an analysis in De Cat, Gusnanto, et al. (under review).

In answer to the quality problem, this investigation shows that greater questionnaire detail does not necessarily result in more precise and reliable measures. In particular, asking respondents to report how much time the child spends with each interlocutor (or group of interlocutors) in each context does not afford a more informative predictor of language outcomes, in spite of the greater precision it is designed to achieve in the estimates of current exposure to each language. This does not necessarily mean that the detailed information gathered in Q-BEx and other questionnaires may not be useful. In the context of a research study, this will depend on the specific research questions asked. However, when all that is needed is a more general bilingualism measure, or when the questionnaire is used in an educational or clinical setting, including these additional questions might not necessarily be worth the time investment. Further research will be required to ascertain whether the reliability of the weighted estimates increases if the questionnaire is administered via an interview. There is some preliminary evidence that this might be the case (Verhoeven et al., 2024). In our sample, where the Q-BEx questionnaire was self-administered in the Netherlands and the UK, we found a higher rate of unlikely answers on the timetabling questions (which are a part of the weighted exposure estimate) when compared to the French sample, to which Q-BEx was administered as an interview with some of the parents (see Kašćelan et al. in preparation for more details). Finally, we note that, across both sets of analyses, the differences in predictive ability among alternative measures of language exposure were relatively modest (as shown for example by the overlapping confidence intervals in the cross-validation). This is due to the fact that the baseline model already contains predictors that account for a substantial amount of variability in the data. Our comparisons were based on differences in a single additional predictor, using a relatively modest sample size.

5 Conclusion

The cost-benefit analysis presented here objectively assessed the informativity of language exposure estimates as predictors of language proficiency in bilingual and trilingual children.

In terms of recommendations for the customisation of Q-BEx when aiming to obtain reliable predictors of language proficiency, the results of this validation study suggest that the most effective implementation of the language exposure and use model, in terms of cost-benefit, is to include (i) the short version of current estimates submodule, which consists of

8 questions (one about the child's language exposure and one about their language use in each of the following contexts: home, school/daycare, community, and holidays), as well as (ii) the cumulative estimates submodule. We hope the protocol developed for this study will be used by others to extend the investigation to other age groups and bilinguals growing up in different contexts.

Beyond the validation of Q-BEx, this study shows that, in heritage bilinguals in their first few years of schooling, proficiency in the Societal Language is best predicted by cumulative estimates of language exposure, while proficiency in the Heritage Language is best predicted by current estimates, especially in the home.

References

Burnham, Kenneth and David Anderson (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer Science and Business Media.

Byers-Heinlein, Krista et al. (2019). “The case for measuring and reporting bilingualism in developmental research”. In: *Collabra: Psychology* 5.1, p. 37.

Cohen, Cathy (2016). “Relating input factors and dual language proficiency in French–English bilingual children”. In: *International Journal of Bilingual Education and Bilingualism* 19.3, pp. 296–313. DOI: 10.1080/13670050.2014.982506.

Dass, Ronessa et al. (2024). “A Content Overlap Analysis of bilingualism questionnaires: Considering diversity”. In: *Bilingualism: Language and Cognition* 27.4, pp. 744–750. DOI: 10.1017/S1366728923000767.

De Cat, Cécile, Arief Gusnanto, et al. (under review). “A data-driven approach to the issue of ”catching up with monolinguals””. In.

— (in preparation). “Individual differences in heritage language outcomes in 5- to 9-year-olds”. In.

De Cat, Cécile, Draško Kašćelan, Philippe Prévost, Ludovica Serratrice, Laurie Tuller, and Sharon Unsworth (2022). *Quantifying Bilingual EXperience (Q-BEx): questionnaire manual and documentation*. Manuscript. DOI: 10.17605/OSF.IO/V7EC8.

De Cat, Cécile, Draško Kašćelan, Philippe Prévost, Ludovica Serratrice, Laurie Tuller, Sharon Unsworth, and the Q-BEx consortium (2023). "How to quantify bilingual experience? Findings from a Delphi consensus survey". In: *Bilingualism: Language and Cognition* 26.1, pp. 112–124. DOI: 10.1017/S1366728922000359.

De Cat, Cécile, Laurie Tuller, et al. (under review). "Using Q-BEx to Identify Risk for Language Impairment in Bilingual Children". In.

Desmarais, Bruce and Jeffrey Harden (2014). "An unbiased model comparison test using cross-validation". In: *Quality and Quantity* 48.4, pp. 2155–2173.

DeVellis, R. (2017). *Scale Development: Theory and Applications*. 4th edition. Thousand Oaks, CA: Sage Publications.

Dunn, Lloyd, Leota Dunn, and Liesbeth Schlichting (2005). *Peabody Picture Vocabulary Test-III-NL*. Lisse: Harcourt Test Publishers.

Dunn, Lloyd, Leota Dunn, J. Sewell, et al. (2009). *The British picture vocabulary scale*. London: GL Assessment Limited. ISBN: 0708719554.

Dunn, Lloyd, Claudia Thoriault-Whalen, and Leota Dunn (1993). *Echelle de Vocabulaire en Images Peabody (EVIP)*. Toronto: Psycan.

Efron, B. and R. Tibshirani (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy". In: *Statistical Science* 1.1, pp. 54–75.

— (1994). *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9780429246593.

Hill, L. J. et al. (2021). "Large-scale assessment of 7-11-year-olds' cognitive and sensorimotor function within the Born in Bradford longitudinal birth cohort study". In: *Wellcome Open Res* 6, p. 53. DOI: 10.12688/wellcomeopenres.16429.2.

Hoff, Erika (2003). "The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech". In: *Child development* 74.5, pp. 1368–1378.

Kašćelan, Draško et al. (2021). "A review of questionnaires quantifying bilingual experience in children: Do they document the same constructs?" In: *Bilingualism: Language and Cognition* 25.1, pp. 29–41. ISSN: 1366-7289. DOI: 10.1017/S1366728921000390.

— (in preparation). "Putting the Q-BEx questionnaire to a quality test: Questionnaire design and quality checks". In.

Kort, W., M. Schittekatte, and E. Compaan (2008). *CELF-4-NL: Clinical evaluation of language fundamentals-vierde-editie*. Pearson Assessment and Information B.V.

Kubota, Maki and Jason Rothman (2024). "Modeling individual differences in vocabulary development: A large-scale study on Japanese heritage speakers". In: *Child Development*. DOI: 10.1111/cdev.14168.

Luk, Gigi and Alena G. Esposito (2020). "BLC mini-series: Tools to document bilingual experiences". In: *Bilingualism: Language and Cognition* 23.5, pp. 927–928. ISSN: 1366-7289. DOI: 10.1017/S1366728920000632.

Marinis, Theo and Sharon Armon-Lotem (2015). "Sentence Repetition". In: *Methods for assessing multilingual children: disentangling bilingualism from Language Impairment*. Ed. by Sharon Armon-Lotem, J. de Jong, and N. Meir. Multilingual Matters.

Polišenská, Kamila, Shula Chiat, and Penny Roy (2015). "Sentence repetition: what does the task measure?" In: *International Journal of Language Communication Disorders* 50.1, pp. 106–118. DOI: https://doi.org/10.1111/1460-6984.12126.

Prévost, Philippe et al. (under review). "Predicting accuracy on the LITMUS Sentence Repetition task". In.

Rocha-Hidalgo, Joscelin and Rachel Barr (2022). "Defining bilingualism in infancy and toddlerhood: A scoping review". In: *International Journal of Bilingualism* 27.3, pp. 253–274. DOI: 10.1177/13670069211069067.

Rujas, Irene et al. (2021). "Sentence Repetition Tasks to Detect and Prevent Language Difficulties: A Scoping Review". In: *Children* 8.7, p. 578.

Schulz, Petra and Angela Grimm (2019). “The Age Factor Revisited: Timing in Acquisition Interacts With Age of Onset in Bilingual Acquisition”. In: *Frontiers in Psychology* 9, p. 2732.

Semel, E., E. Wiig, and W. Secord (2017). *Clinical Evaluation of Language Fundamentals (Fifth edition)*. Fourth Edition (CELF-4). London: Pearson.

Serratrice, Ludovica et al. (under review). “Predictors of vocabulary breadth and depth in the societal language of multilingual children in three European countries”. In.

Thordardottir, Elin (2011). “The relationship between bilingual exposure and vocabulary development”. In: *International Journal of Bilingualism* 15.4, pp. 426–445. DOI: 10.1177/1367006911403202.

Unsworth, Sharon (2013). “Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: The case of Dutch gender”. In: *Bilingualism: Language and Cognition* 16, pp. 86–110.

Unsworth, Sharon et al. (under review). “Unpacking the richness of language experience as a predictor of bilingual children’s language proficiency”. In.

Verhagen, Josje et al. (2024). “Relationships between bilingual exposure at ECEC and vocabulary growth in a linguistically diverse sample of preschoolers”. In: *Journal of Applied Developmental Psychology* 93, p. 101657. DOI: 10.1016/j.appdev.2024.101657.

Verhoeven, Emma et al. (2024). *Comparing Different Methods That Measure Bilingual Children’s Language Environment: A Closer Look at Audio Recordings and Questionnaires*. DOI: 10.3390/languages9070231.

Wechsler, David (2013). *Wechsler Preschool & Primary Scale of Intelligence (WPPSI-IV)*. Bloomington, MN: Pearson.

— (2014). *Wechsler Intelligence Scale for Children (WISC-V): Technical and Interpretive Manual*. Bloomington, MN: Pearson.

Wiig, E., E. Semel, and A. Wayne (2019). *CELF 5—Batterie d’évaluation des fonctions langagières et de communication (Adaptation française ECPA)*. Pearson.

5.1 Data availability statement

The data and R scripts that support the findings of this study are available via the OSF at `https://osf.io/7v2yw` (for the data) and `https://osf.io/qajbd` (for the script)

Supplementary materials

Distribution of heritage languages per country

Table 6: Heritage language speakers by country of residence

Heritage Language	FR	NL	UK
Afrikaans	0	1	0
Arabic	7	4	2
Bangla	0	0	4
Bulgarian	0	4	0
Chinese	1	0	4
Creole	2	0	0
Czech	0	0	1
English	3	10	0
French	0	6	1
Frisian	0	1	0
German	0	9	1
Greek	0	1	5
Gujarati	0	0	3
Hindi	0	0	2
Hindko	0	0	1
Italian	1	2	1
Japanese	0	0	1
Kurdish	0	1	0
Latvian/Lithuanian	0	0	1
Lingala	2	0	0
Manouche	1	0	0
Mirpuri	0	0	1
Polish	0	3	2
Portuguese	2	4	1
Punjabi	0	1	2
Romanian	2	1	2
Russian	0	1	1
Spanish	0	3	2
Telugu	0	0	2
Turkish	0	11	0
Ukrainian	0	3	0
Urdu	0	0	9

Covariates based on Q-BEx information

Covariates included in the models

Table 8 shows which covariates were included in each model. The outcome measure is specified in the top two rows. General development was taken into account through the cognitive

Q-BEx module	**Covariates used in some models**
Background information	Older siblings
Risk factors	Age of first words
	Age of first word combinations
	Caregiver concern about language before the age of 4
Richness	Frequency of organised activities in each language
	Frequency of interactions with friends in each language
	Frequency of tech activities in each language
	Frequency of writing in each language
	Frequency of reading in each language
	Frequency of out-of-school lessons about each language
	Frequency of in-school lessons about each language
	Frequency of homework in each language
	Highest caregiver education
Attitudes	Willingness to speak each language

Table 7: Covariates derived from other Q-BEx modules

measures (short-term memory, working memory, and non-verbal intelligence) rather than age, which is less directly interpretable.

	Sentence repetition in the SL			SL vocabulary		HL
	identical	target	grammatical	breadth	depth	oral abilities
willingness to speak the SL	x	x	x		x	
backward digit recall		x	x		x	
forward digit recall	x	x	x	x	x	
non-verbal intelligence			x	x	x	
delay of first words					x	
early parental concerns					x	x
delay of first word combinations					x	
older siblings	x	x	x	x	x	
non-word repetition		x	x	x		
frequency of organised activities (SL)	x	x	x	x	x	
frequency of SL lessons at school		x		x	x	
frequency of SL lessons outside school	x	x		x	x	
SL homework frequency					x	
frequency of reading (SL)	x				x	
frequency of tech-related activities (SL)	x		x		x	
frequency of being with friends (SL)		x			x	
frequency of writing (SL)	x		x	x	x	
frequency of being with friends (HL)						x
frequency of reading (HL)						x
frequency of HL lessons at school						x
societal language	x	x		x	x	
highest caregiver education		x	x	x	x	
language entropy			x			
multilingualism (mono, bi, tri)						x

Table 8: Covariates included in each model

Correlations between the alternative estimates of language exposure

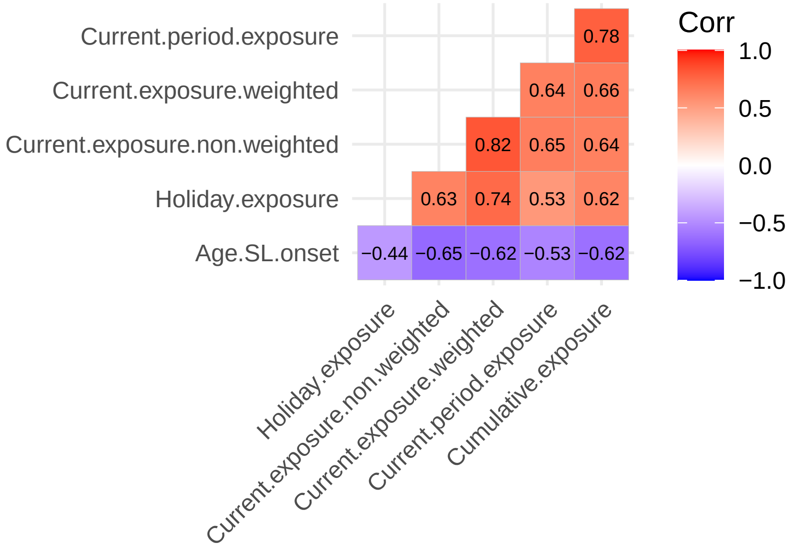


Figure 5: Correlation between the alternative measures of SL exposure

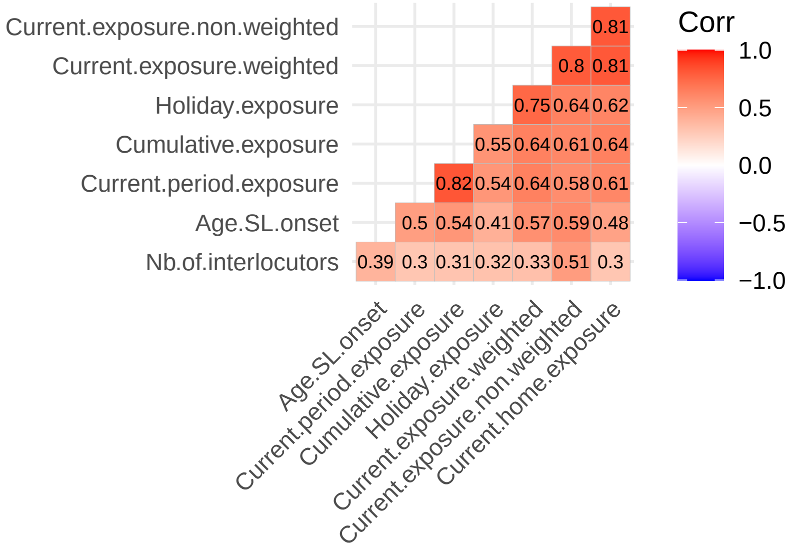


Figure 6: Correlation between the alternative measures of HL exposure

Model comparisons for SL outcome measures

Table 9: Comparison of models with different language exposure predictors of Identical Repetitions (SRT)

dAIC	weight	Model
15.1	0.00	res.Current_period_exposure_SL
8.5	0.01	res.SL.cumulative.exposure.mo
12.7	0.00	res.SL.current.exposure.Nw
13.4	0.00	res.Holiday.current.exposure.SL.unweighted
13.0	0.00	res.SL.current.exposure.w
0.0	0.98	res.SL.onset.age

Table 10: Comparison of models with different language exposure predictors of Grammatical Attempts (SRT)

dAIC	weight	Model
23.1	0	Current.period.exposure
16.4	0	Cumulative.exposure
17.3	0	Current.exposure.non.weighted
24.1	0	Holiday.exposure
20.8	0	Current.exposure.weighted
0.0	1	Age.SL.onset

Table 11: Comparison of models with different language exposure predictors of vocabulary breadth

dAIC	weight	Model
13.2	0.00	Current.period.exposure
0.0	0.99	Cumulative.exposure
11.5	0.00	Current.exposure.non.weighted
15.7	0.00	Holiday.exposure
15.0	0.00	Current.exposure.weighted
9.5	0.01	Age.SL.onset

Table 12: Comparison of models with different language exposure predictors of vocabulary depth

dAIC	weight	Model
2.1	0.14	Current.period.exposure
0.0	0.40	Cumulative.exposure
3.4	0.07	Current.exposure.non.weighted
2.9	0.10	Holiday.exposure
2.0	0.15	Current.exposure.weighted
2.1	0.14	Age.SL.onset

Exploring sampling effects

Table 13: Comparison of models predicting identical repetition scores, excluding children with age of SL onset above 48 months (111 observations)

dAIC	weight	Model
8.2	0.01	Current.period.exposure
4.5	0.09	Cumulative.exposure
6.8	0.03	Current.exposure.non.weighted
7.6	0.02	Holiday.exposure
7.6	0.02	Current.exposure.weighted
0.0	0.83	Age.SL.onset

Table 14: Comparison of models predicting identical repetition scores, excluding children with age of SL onset at or above 42 months (98 observations)

dAIC	weight	Model
3.7	0.06	Current.period.exposure
0.8	0.25	Cumulative.exposure
2.0	0.14	Current.exposure.non.weighted
2.7	0.10	Holiday.exposure
3.0	0.08	Current.exposure.weighted
0.0	0.38	Age.SL.onset

Table 15: Comparison of models predicting identical repetition scores, excluding children with age of SL onset at or below 36 months (87 observations)

dAIC	weight	Model
1.2	0.14	Current.period.exposure
0.6	0.19	Cumulative.exposure
0.0	0.25	Current.exposure.non.weighted
0.9	0.16	Holiday.exposure
1.4	0.12	Current.exposure.weighted
1.4	0.13	Age.SL.onset

Cross-validation results

Table 16: Cross-validation comparisons with Identical Repetitions as outcome measure

	Mean error	Range	CI 2.5%	CI 97.5%
Baseline.model	0.2372	0.0765	0.1807	0.3152
Current.period.exposure	0.2337	0.0766	0.1757	0.3157
Current.exposure.non.weighted	0.2345	0.0671	0.1819	0.3146
Current.exposure.weighted	0.2340	0.0715	0.1793	0.3153
Holiday.exposure	0.2359	0.0714	0.1837	0.3171
Age.SL.onset	0.2132	0.0745	0.1718	0.2949
Cumulative.exposure	0.2303	0.0721	0.1761	0.3041

Table 17: Cross-validation comparisons with Grammatical Attempts as outcome measure

	Mean error	Range	CI 2.5%	CI 97.5%
Baseline.model	0.2352	0.0745	0.1807	0.3152
Current.period.exposure	0.2338	0.0897	0.1757	0.3157
Current.exposure.non.weighted	0.2282	0.0911	0.1819	0.3146
Current.exposure.weighted	0.2335	0.0952	0.1793	0.3153
Holiday.exposure	0.2376	0.0775	0.1837	0.3171
Age.SL.onset	0.2156	0.0789	0.1718	0.2949
Cumulative.exposure	0.2296	0.0620	0.1761	0.3041

Table 18: Cross-validation comparisons with Vocabulary Breadth as outcome measure

	Mean error	Range	CI 2.5%	CI 97.5%
Baseline.model	16.1054	9.9289	0.1807	0.3152
Current.period.exposure	16.2024	8.4499	0.1757	0.3157
Current.exposure.non.weighted	16.0059	8.7840	0.1819	0.3146
Current.exposure.weighted	16.2715	9.1137	0.1793	0.3153
Holiday.exposure	16.2612	9.5322	0.1837	0.3171
Age.SL.onset	15.7773	10.1145	0.1718	0.2949
Cumulative.exposure	15.7483	6.2612	0.1761	0.3041

Table 19: Cross-validation comparisons with Vocabulary Depth as outcome measure

	Mean error	Range	CI 2.5%	CI 97.5%
Baseline.model	0.1510	0.0572	0.1807	0.3152
Current.period.exposure	0.1530	0.0476	0.1757	0.3157
Current.exposure.non.weighted	0.1532	0.0474	0.1819	0.3146
Current.exposure.weighted	0.1525	0.0600	0.1793	0.3153
Holiday.exposure	0.1548	0.0605	0.1837	0.3171
Age.SL.onset	0.1524	0.0579	0.1718	0.2949
Cumulative.exposure	0.1522	0.0532	0.1761	0.3041

Table 20: Cross-validation comparisons with oral proficiency in the Heritage Language as outcome measure

	Error	Range	CI_2.5%	CI_97.5%
Baseline.model	0.7500	0.1822	0.7344	0.8359
Current.exposure.weighted	0.7109	0.2049	0.6875	0.8594
Current.exposure.non.weighted	0.7422	0.2060	0.7031	0.8516
Current.period.exposure	0.7656	0.1595	0.7264	0.8438
Current.home.exposure	0.6797	0.1573	0.6484	0.8516
Holiday.exposure	0.7656	0.1827	0.7188	0.8672
Nb.of.interlocutors	0.7500	0.1589	0.7422	0.8594
Cumulative.exposure	0.7578	0.1827	0.7266	0.8594
Age.SL.onset	0.7656	0.1362	0.7344	0.8594