Q-BEx Design Documentation

Cécile De Cat, Draško Kašćelan, Philippe Prévost, Ludovica Serratrice, Laurie Tuller, Sharon Unsworth

Design: principles and processes

The challenge

Bilingualism is a multi-dimensional construct attempting to capture a complex and a dynamic process.[1] The term is sometimes even extended to speakers of more than two languages. Children who are exposed to two languages do not represent a homogeneous group: the amount of exposure to each language can vary widely, and this in turn results in considerable variation in language proficiency. Bilingual children also vary widely in their use of each language, and the specific effect that this has on language proficiency is currently poorly understood. The impact of cumulative experience (i.e., over lifetime) as opposed to current experience (e.g., in the current year) is also subject to debate. In addition to the amount of language experience, other aspects are also important, such as qualitative aspects of the language experience, attitudes towards each language and towards bilingualism, and language mixing.

Given this multifaceted reality, detailed documentation of language experience is essential in order to better understand and accurately assess the language development of bilingual children. A multitude of measures have emerged over the years in bilingualism research, with no consensus about what exactly they should index, or the level of precision required for different end-users (i.e. researchers, teachers, speech and language therapists) in their assessment of bilingual children. This multitude of measures results in a lack of comparability across studies and hinders cross-sector exchanges.

The Q-BEx project was set up to address these challenges. Our aims in that respect are:

1. To develop a customisable online tool (and associated back-end calculator) to document language experience in bilingual (and trilingual) children, informed by
    - a comprehensive review of existing questionnaires,
    - a consensus among a representative group of researchers and practitioners, and
    - best practice from psychometrics.

   Customisation is to allow
    - the selection of modules, each representing a key aspect of bilingual language experience;
    - different levels of detail within modules;
    - targeting adult or child respondents;
    - administration in different languages.
2. To determine the optimal level of detail required to predict various aspects of language proficiency in 5- to 9-year-old bilinguals.

The aim of this tool is to enable the use of identical measures across a wide range of studies, and thereby inform important questions in bilingualism research, beyond the lifetime

---

[1] References to the relevant literature have generally not been included in this manual to increase legibility.

of this project: Can a universal approach to the documentation of bilingualism be used across contexts and populations? Is it possible to reliably document all aspects of bilingual experience (e.g., language mixing, input quality) by means of a questionnaire? Will the creation of this tool truly facilitate communication across different fields and sectors?

The sections below report on the key stages in the development of the new questionnaire.

Review of existing questionnaires

At the outset of the project, it was necessary to review and understand the ways in which existing questionnaires capture bilingual experience. We identified and reviewed 48 questionnaires used to document bilingualism in children (broadly defined as 0-18-year-olds). Across these tools, our analysis identified 32 overarching constructs (e.g., language exposure, current skills in the societal language, etc.) and their 194 components (e.g., relative frequency of exposure, reading skills, etc.). Furthermore, for each construct and their components, we calculated the frequency with which they were documented across the questionnaires. A full list of constructs and their components, as well as the details about their frequency can be found in Kašćelan et al. (2021).

Apart from the general overview, we examined in detail the operationalisation of the following overarching constructs: language exposure and language use, current skills in the home language (HL) and in the societal language (SL), as well as activities in each language. These appear to be the most central to the documentation bilingual experience (e.g., Li et al., 2006; Unsworth, 2016).

Our analysis revealed that language exposure and language use were documented in the majority of questionnaires (in 96% and 73% of cases respectively). Current skills in both HL and SL were also quite commonly inquired about (each in 42% of questionnaires), while 44% of questionnaires asked about activities in specific languages. The operationalisation of these aspects varies considerably across questionnaires. For instance, even when questions were intended to tap into the same or closely similar constructs, differences in response scales could hinder the comparability of the obtained data (e.g., due to different number of points on the scale, different cut-off points, or different wordings).

To allow genuine comparability across studies, a minimum requirement is that of transparency in how each variable is operationalised, ideally by including the relevant questionnaire and calculations in a publication. The intent is that there should be a consensus among bilingualism experts to determine a common set of constructs to document, as well as a common set of tools for use across the sectors.

Delphi consensus survey

To inform the scope of the new questionnaire (i.e., identify the set of constructs to document), we conducted an international, cross-sector consensus survey, following the Delphi method (Iqbal & Pipon-Young, 2009). This approach can be used to explore diverse opinions in a group of experts in a particular field (also referred to as panellists or stakeholders) or to lead them towards consensus when there is none. We recruited bilingualism experts among practitioners (teachers and speech and language therapists) and researchers, aiming for as diverse a panel as possible.

In a Delphi study, the panellists are presented with a list of statements (e.g., “Language mixing should be documented in a questionnaire about bilingualism”) which they are asked to rate in terms of (dis)agreement. The original list of statements was informed by a scoping workshop in Leeds in January 2020, bringing together 22 researchers (from 11 countries) and 14 practitioners (from 3 countries).

In the online study, the panellists were asked to rate each statement on a 5-point scale (*strongly disagree, disagree, I don’t know, agree, strongly agree*). The consensus threshold and the number of survey rounds were pre-determined to limit the risk of bias. As the first set of statements had been informed by a diverse panel of experts, we opted for 2 rounds for the online survey. In the second round, the panellists were asked to re-rate the statements that had made it to the “grey zone” in the first round: below the consensus threshold (75% of panellists rating the statement as *agree* or *strongly agree*) but above 60% agreement. Crucially, in round 2, panellists were given for each statement (i) their original rating, (ii) the panel’s average rating and (iii) any comments provided by panellists at round 1. In that light, panellists were asked to amend or confirm their original rating. The two rounds of the online survey were completed by 132 panellists from 29 countries. At the end of the study, 98 statements had reached consensus (i.e., 79% of statements included over both rounds).

There was clear consensus regarding the need for “a set of common measures of children’s bilingual language experience, to allow comparability across studies and to facilitate communication across sectors (research, education, therapy)” (96% agreement). The statements reaching consensus allowed us to determine the extent to which each of the following aspects should be documented: language exposure and use, input quality, education and literacy, language proficiency, language difficulties, language mixing, attitudes towards languages and language mixing, and background information about the child. The consensus also included opinions on questionnaire administration and modularity. A complete list of the statements which reached the consensus, as well as the complete dataset from the online study can be accessed in De Cat et al. (2021).

Following the Delphi study, these statements were used to draft an initial matrix of modules to be included in the questionnaire. Each module consisted of a set of questions informed by the consensus statements. The question formulations, order, and presentation followed best practice from the psychometric literature and by additional consultations with experts in the field (for some of the modules). In the next section, we elaborate on these steps.

Questionnaire creation

The Delphi study indicated what should be included in the questionnaire, but it did not focus on methods. For each aspect of bilingual language experience reaching consensus for inclusion, we would need to work out its operationalisation (i.e., define how it should be measured). This was informed by a review of current practices, consultation with experts, and by learning from psychometrics (the branch of psychology that specialises in measurement and assessment) - as explained in the next couple of sections.

Consulting experts and integrating current best practice

The Delphi consensus called for the inclusion of questions probing the risk of language disorder and probing language proficiency (in case it could not be measured by other means). We derived the relevant questions from the PaBiQ questionnaire (Tuller, 2015). The PaBiQ itself is based on ALDeQ (Paradis et al., 2010) and ALEQ (Paradis, 2011). The choice of questions and their formulation was informed by the practice of two members of the Q-BEx team who were also involved in the design of PaBiQ: Philippe Prévost and Laurie Tuller. From the PaBiQ, we only included the most informative questions in relation to the risk of DLD: questions about the early language milestones (i.e., age of first words, age of first short sentences) and concerns about language development before the age of four. They are subsumed under a module entitled *Risk factors*. Note that the term "risk factors" is never visible to respondents (none of the module names is). Care was taken to word the questions as neutrally as possible, to avoid worrying the respondent.[2]

Parental evaluation of the child's language abilities is also probed in two other modules: the *Language proficiency* module and the *Attitudes and satisfaction with child's language* module. The *Language proficiency* module includes questions about the child's proficiency in each language (with and without a comparison/reference group). The *Attitudes and satisfaction with child's language* module includes questions about satisfaction with a child's language speaking and understanding skills among other attitude-related questions.

The content, formulation and order of questions related to child's language abilities and the risk of DLD (appearing across the three modules of the current version) was optimised through consultations with specialists of language disorders in bilinguals: Sharon Armon-Lotem, Elma Blom, Ute Bohnacker, Daniela Gatt, Ewa Haman, Camille Moitel Messarra, and Hadar Oz. For the *Attitudes* module, we also consulted Aleksandra Tomić, while Elma Blom and Maria del Carmen Parafita Couto provided feedback on questions related to language mixing attitudes.

There was consensus in the Delphi study that the richness of the child's language environment and experience should be documented. There is a growing body of research focusing on "richness" or "input quality" (see, e.g., the double special issue in the *Journal of*

---

[2] See the section below on the *Evaluation phase*, in which we asked caregivers if any of these questions made them feel uncomfortable.

Child Language: Blom & Soderstrom 2020a, 2020b), but no consensus has been reached yet on what constitutes "rich" or "high-quality" input. Literacy activities are often considered to contribute to a rich language environment, but whether or not the medium is important remains unclear: can equal richness be obtained from oral practices (e.g., storytelling, poetry) compared to book-based practices? Do both provide equally useful scaffolds for the child's literacy development in the school language? Another frequently used indicator is socio-economic status (often operationalised as maternal education). In spite of a large body of research documenting the impact of socio-economic status on language development (in monolinguals as well as bilinguals), more work is required to understand the extent to which socio-economic status is a reliable proxy for "input quality", and if so, which factors associated with socio-economic status drive this effect. In particular, what standards should we use to inform the criteria used to capture richness of language experience? In her ethnographic study of two very different rural communities in South Carolina, Brice-Heath (1983) demonstrates that different types of experiences can be equally rich, but not in a way that is detectable or understood by those who assess them.

Our approach to the *Richness of language experience* module was to document the diversity of interlocutors in each language (how many and how proficient), the types of language use (media, gaming, literacy activities), and the level of education of each caregiver in each language. We expect this module will evolve as new research emerges in relation to the themes highlighted in the previous paragraph. The method we adopted to calculate richness scores for each language is based on the approach used in ALEQ (Paradis, 2011), where the sum of scores from richness-related questions is divided by the sum of maximum scores that can be achieved on these questions. Thus, the richness score range for each language is between 0 and 1, with scores closer to 1 indicating a richer environment for that particular language.

For the *Language mixing* module, not many of the relevant statements reached consensus in the Delphi survey. We believe there are two explanations for this. First, there is scepticism about whether language mixing can be documented accurately with a questionnaire. Second, language mixing has not yet been widely studied in bilingual children (at least not as much as in adults) so panellists may not have felt sufficiently informed to make a judgement. Our aim with this optional module is to *attempt* the detailed documentation of language mixing, and to invite validation studies focusing on that aspect.

As for the previous module, we expect this one to evolve as new research emerges, clarifying the extent to which this aspect of bilingual experience can be documented reliably through a questionnaire. To optimise the questions in this module, we consulted with Elma Blom and Maria del Carmen Parafita Couto. The consensus was to ask separately about language mixing practices at home and outside the home, including questions about the types of switches (i.e. consisting in one word, two-three words, or a whole sentence). Provided that the users use single-clause sentences to exemplify these switches in the questionnaire (which we strongly encourage), the first two examples represent estimates of INTRA-clausal switches (i.e., within a clause), while the last one represents an example of an INTER-clausal switch (i.e., between clauses). As an illustration, here we provide examples[3] which could be used in case of bilinguals speaking Spanish and English[4]:

- One-word switch: Puedes darme la drink?
- Two-three-word switch: I really like your nueva camisa.
- Sentence (i.e., clause) switch: Podemos comenzar ahora. Or we can wait a bit longer.

The contexts within which we probe the attitudes towards language mixing are: at home, at school, and in the local community (outside school), to align with the contexts in which we ask about preferred languages. Both sets of attitude-related questions (about language mixing and about preferred languages) are in the *Attitudes and satisfaction with child's language* module.

Obligatory modules

The questionnaire is composed of seven modules:

- Background information
- Risk factors
- Language exposure and use
- Language proficiency
- Richness of linguistic experience
- Attitudes and satisfaction with child's language
- Language mixing

The modules (and their sub-modules) are optional, except for two obligatory modules: *Background information* and *Risk factors*. These are obligatory for functional reasons, as some of the information they collect is variable-setting: it determines or conditions the formulation of questions in other modules. For instance, the *Background information* module asks for the names and numbers of household members, the names of languages, the birth date of the child, the country of residence. These are used to customise the formulation of questions in subsequent modules.[5] The *Risk factors* module asks about the age at which the child produced first words in any language. This determines the age from which cumulative language use is calculated for the child. As an illustration, imagine the caregiver specifying that the age of their

---

[3] The questionnaire is programmed to use the same three examples when asking about both the home and the outside of home context. Therefore, when creating switching examples for your project, use the ones that could be heard/said in both contexts.

[4] In the examples we provide here, if we observe the switches in a linear way (left-to-right), the first and the third example are switches from Spanish to English, while the second one is from English to Spanish. We recommend that in your examples you use the direction of switching common for the community that you study. If unfamiliar with the community practices, make sure to consult a member of that community regarding your examples.

[5] Note that people's names are not stored in the data. No record of these is kept after the questionnaire is completed.

child's first words is 1 year and 4 months. Consequently, in the *Language exposure and use* module, when asked about the language exposure/use in the past (i.e., cumulative estimates), for the period from the child's birth until the age of 1 year and 4 months, the caregiver will only be asked about the languages that the child was exposed to, but not about the languages that the child used (because she/he was non-verbal in this period). Note that in the child version of the questionnaire, the *Risk factors* module cannot be included, as we expect that children are unlikely to know answers to these questions. For the purposes of cumulative language calculations in the child version, we therefore had to set a research-informed age of first words. We set it as 1 year and 1 month, as this corresponds to the average age when neurotypical bilinguals produce their first words (e.g., Paradis et al., 2010; de Almeida et al., 2017).

Lessons from the psychometric literature

The initial Delphi workshop (in January 2020) included a keynote presentation by health psychology expert Professor Kate Harvey, on the guiding principles of questionnaire creation and validation. Following this, we reviewed relevant psychometric literature: DeVellis (2017), Dörnyei and Taguchi (2010), and Dillman et al. (2014). As our aim was to build a modular questionnaire documenting various aspects separately, rather than eliciting a scale providing one index score (e.g., an overall bilingualism score), the guidelines in Dillman et al. (2014) were the most adequate for the purposes of the project.

While within our team we had had experience of constructing or adapting questionnaires to document bilingualism in children (e.g., BiLEC, PaBiQ), turning to the psychometric literature enabled us to consider all possible sources of error more systematically. The Total Survey Error approach (see Dillman et al., 2014) aims to reduce all possible error sources through meticulous survey planning, sample selection, design of the questionnaire itself, questionnaire distribution and data analysis. Dillman et al. (2014, pp. 3-8) outline four main types of errors that need to be minimised when trying to estimate true values of variables in population: coverage error, sampling error, nonresponse error, and measurement error. Here we report on how the steps taken in our questionnaire design to minimise potential sources of measurement error and on how we sought to minimise the risk of non-responses, through the optimisation of the questionnaire's online interface. Some issues relating to coverage, sampling, and nonresponse error are discussed in relation to our questionnaire in De Cat et al. (2021).

Various circumstances can lead to measurement errors. For instance, the concepts might be measured in inadequate ways; the content of questions might encourage certain responses to make the respondent look more favourable; the question wording might be confusing; the order of questions or the visual layout might affect the answers; the survey mode can play a role in how participants respond to a survey. Further in this document we outline the steps we took to minimise potential sources of measurement error, following guidelines from the psychometric literature.

Following the literature review (Kašćelan el al., 2021) and the Delphi study (De Cat et al., 2021), we knew that our new questionnaire was likely have these characteristics:

- Two versions, depending on respondent type: a caregiver version (about their child's bilingualism) and a child version (about their own bilingualism).
- Administered online, likely only on computers, laptops and tablets. This particular characteristic immediately implies the presence of coverage error, as this mode of distribution will exclude bilinguals with no access to the internet and/or computers. On the other hand, computerisation brings several benefits. Some of these include minimising errors of commission (i.e., respondents answering questions they are not supposed to answer) and minimising errors of omission (i.e., respondents accidentally skipping the questions they were supposed to answer). Computerisation also allows for personalisation of the questions, using fills (i.e., using information provided by the respondent in follow-up questions). For instance, if a caregiver specifies that there are 4 people (other than the target child) in the child's household, follow-up questions can be tailored to ask about these 4 people.
- Modular: some of the sections/modules will be obligatory while others will be optional, depending on the needs of the professional using it.
- Complemented by a backend calculator to derive scores from the raw data, for example: language exposure and use in the current year of child's life, cumulative language use and exposure, richness scores for each language.
- Self-administered rather than as an interview protocol.

These features determined which parts of the psychometric literature to follow. Users wishing to adapt the Q-BEx questionnaire for use in paper version or as an interview protocol, or wishing to optimise it for administration via mobile phones, will need to consult additional sources from that literature. We are planning to create an interview protocol (in collaboration with teachers and SLTs) which could be used with the existing computerised version of the questionnaire.

In constructing the questions and response scales, we were guided mostly by chapters 4, 5, 6, 7, and 9 from Dillman et al. (2014), which covered the following topics: the fundamentals of writing questions, writing open- and close-ended questions, aural and visual design of questions and questionnaires, ordering of questions, web questionnaires and implementation. While these chapters contain a comprehensive set of guidelines and their elaboration, here we briefly outline the ones that influenced our design directly. We also specify when we decided not to follow some of the guidelines, if we estimated that their implementation was unlikely to minimise measurement error. We address these guidelines thematically in the next four subsections: question and scale writing, question order, visual design, and online implementation.

Question and scale writing

Close-ended vs. open-ended format. The open-ended question format allows respondents to write their own answer and allows nuanced responses, but it can discourage respondents due to the effort required from them, especially if it is overused. This format also requires coding of the data, which would be a burden for the questionnaire users and risks being less systematic. For these reasons, we decided to use mostly the close-ended format throughout the questionnaire (i.e., making respondents choose from a list of answers). The open-ended format was reserved only for a very small set of questions (thus following the guideline to *use it sparingly*). For instance, open-ended questions were used when the respondents have to input the name of the target child (CQ.15), the names of other household members (NQ.17, NQ.18), or to specify the name of a language that the child speaks/understands if the language is not listed in a drop-down menu (Q.72).

Explicit frames of reference. To ensure that the data is collected as consistently as possible and fit to inform calculations (where required), the wording of questions and response choices needed to be precise and coherent in relation to the targeted unit of measurement. For instance, to inform calculations of current and cumulative language exposure and use, the information elicited needed to be expressed within a specific time frame in a consistent way. Therefore, every time we asked about current language exposure and use in different contexts and with different individuals, we prefaced the questions with an overarching statement: *Think about a typical week in the current year*. In this way, the estimates collected from the respondents are anchored within the same time frame rather than leaving the time frame open to interpretation, where *current* could mean the last few days to some respondents, while it might mean the last few weeks, months or even years to some others.

Inclusivity: making questions applicable to all respondents. While we attempted to create a tool suitable to document children’s bilingual experience across as many contexts and populations as possible, a truly universal tool might be impossible to achieve. The questions were formulated so as to make sense to respondents from diverse backgrounds, but within a core set of assumptions (e.g., existence of a household, maximum number of adults and children per household[6]). The options available to describe family constellations are as inclusive as possible and allow respondents to choose *other* in case the labels offered in the list of options do not correspond to their reality. Respondents also have the option of listing a language other than those provided in the dropdown menu. When documenting language exposure and use, the contexts considered by the questionnaire were informed by the Delphi consensus study (De Cat et al., 2021) but flexibility needed to be allowed. The questionnaire was designed for children up to the age of 18 as long as they attend daycare or school. Two other general assumptions were (i) the existence of a local community (e.g., neighbourhood) and (ii) the existence of a yearly period with no school or day care (*the holidays*). Questions informing cumulative estimates of language exposure and use allow for the specification of periods in the child’s life, defined on the basis of changes to their patterns of language experience. The questionnaire also asks if there were any prolonged periods during which the child did not attend school. It is hoped that these two aspects will allow sufficient flexibility to document the language experience of refugee or immigrant children, for instance.

Further flexibility is afforded by the possibility for the researcher or practitioner to exclude particular modules or sub-modules, and by the availability of response choices indicating that something *(almost) never* occurred (e.g., doing homework in a particular language; language mixing). Some questions also included response options such as *not*

---

[6] A maximum of five adults and five other children can be specified.

relevant - for instance, in relation to the child's writing skills (Q.196): some children are too young to write, and some languages have no written form. In this way, based on respondents' answers, we can infer which contexts/questions do not apply to them or which situations never happen to them.

Asking one question at a time. While this guideline seems rather straightforward, it is not always that easy to follow. Asking many similar questions in succession (each focusing on a different individual or a different context, for instance) can lead to respondent fatigue or error, as they repeatedly have to identify new information in otherwise identical formulations. In such cases, we grouped several potentially separate questions into one. Similar activities were also grouped by categories with examples. For instance, to minimise the risk of diverse interpretations of what computer/technology-related activities referred to, and to maximise efficiency, we listed the types of activities that we wanted the respondents to consider when answering this question (Q.19). However, note that this affects the level of detail of the information collected (e.g., gaming cannot be distinguished from watching movies).

Clarity of the questions: using concrete words wherever possible; avoiding ambiguity; using complete sentences with simple sentence structures; avoiding double negatives. The latter should ensure a *yes* response corresponds to an affirmative statement (and *no* to a negative one). The formulation of all the questions and answer options was checked for readability to be up to and including level B2 proficiency according to the Common European Framework of Reference for Languages (see section *Readability updates* below). Clarity of the questions was further improved during the evaluation phase, in which we carried out cognitive interviews with adult and child respondents (see section *Evaluation phase* below).

Using as few words as possible to pose questions. A compromise needed to be found between this guideline and the need for clarity. In some cases, it was necessary to include examples or clarifications to make sure that the respondents understood the question or the reference points in the same way.

Organising questions in a way to make it easiest for respondents to comprehend the task. We grouped questions thematically: for instance, questions about language mixing types (a word switch, a two-three-word switch, a sentence switch) all fit on one page. Similarly, questions about the use of and exposure to languages with different groups within the school context are also grouped on the same page.

Minimising the number of questions that require respondents to calculate. As the tool aims to quantify bilingual experience, it is impossible to avoid asking respondents to estimate quantities or frequencies (e.g., how many weeks in the last year has the child been on school holidays - NQ.3). However, wherever possible, we tried to avoid asking the respondent to make complex calculations. For instance, to elicit estimates of language exposure and use, we designed an intuitive method to obtain precise estimates without using numbers. In all questions quantifying exposure and use, the respondent sees one slider per language (i.e., two or three horizontal sliders stacked vertically, depending on whether the child is bilingual or trilingual). A pie chart appears on the same screen, divided equally for each language to start with. For each language, the respondent is asked to move the slider to the left (indicating less of that language) or to the right (indicating more of that language). As they do that, the pie chart is updated as a visual representation of the proportion of each language determined by the respondent. The data output stores these estimations in percentages which add up to 100%, i.e., the coherence of quantities reported across languages is automatically guaranteed. That way, it is not possible to report that two languages are used *most of the time* - which can be interpreted either as *they are used equally* or *the respondent made a mistake*). The burden of calculations is shifted to the back-end calculator associated with the online questionnaire (see below).

Stating both the positive and the negative side in the question stem. This aims to avoid biasing respondents towards one end of the response scale. For instance, rather than asking *How satisfied are you with…* a more balanced way would be to ask *How satisfied or dissatisfied are you with….* Or instead of asking *How concerned are you that…* a more balanced approach would be to ask *How concerned, if at all, are you that….* While we planned to implement this guideline, once we started formulating the questions by following this rule, we found that many questions became burdensome to read, which would breach previous guidelines about the length, simplicity or clarity. Furthermore, as we knew from the outset that the tool will be translated in multiple languages, we noticed that similar issues would emerge in some of the languages that the tool is to be translated into. Therefore, in order to be consistent and clear throughout the questionnaire, we decided not to implement this guideline.

Including answer categories that cover all possible scenarios. This was a crucial guideline to follow, considering that most questions in Q-BEx are in the close-ended format. By relying on the team's previous experience, the review of the literature we had conducted (Kašćelan et al., 2021), and the Delphi study (De Cat et al., 2021), we tried to provide as exhaustive response options as possible. In some cases, this required allowing nonsubstantive options (e.g., *I don't know*; *no opinion*). We limited the use of these to reduce the risk of sufficing (i.e., the respondents choosing a nonsubstantive option rather than doing the mental work necessary to report accurate answers, see Dillman et al., 2014, p. 136). Nonsubstantive options were included when we judged that excluding them substantially increased the risk of measurement error. For instance, as language mixing is often an unconscious practice, it might be hard to report its frequency in certain contexts. Forcing a choice of frequency could therefore result in too much inaccuracy. Similarly, when asking the caregivers to compare their child to other bilingual/multilingual children in the country of residence regarding their language skills, we found that there is a high chance that some caregivers will not be able to do this; therefore, including a nonsubstantive option was necessary.

Including answers that are mutually exclusive. We were careful to abide by this guideline throughout the questionnaire. Combining this with the *clarity* guideline, we used numeric ranges rather than vague quantifiers where possible. For instance, when asking about the number of speakers in each language that talk to the child on a regular basis (Q.33), we used options 0, 1-2, 3-5, 6-10, more than 10. Implementing this guideline was challenging for some of the questions, such as for those asking about the frequency of activities in each language. Here we used the following answer options: *(almost) never*, *once or twice a month*, *once or twice a week*, *several times a week*, *every day*. As we go up the frequency scale, most options logically include the previous one (e.g., if an activity is done twice a week, it is also true that it happens at least twice a month). However, pragmatically, selecting a higher frequency band implies that those bands preceding as well as those (even higher ones) coming after do not apply. A challenge for such ordinal scales is to achieve a sufficient but not unrealistic level of detail (i.e., respondents have to be able to make the required estimations, sometimes across large time periods and highly variable contexts). The level of magnitude separating points on the same scale to the next was determined on the basis of these considerations. For instance, in the activities frequency scale (*(almost) never, once or twice a month, once or twice a week, several times a week, every day*), the distance between *once or twice a month* and *once or twice a week* is not the same as between *once or twice a week* and *several times a week*.

Quantifiers (e.g., rarely, often, etc.) vs. a natural metric (e.g., never, once, twice, etc.). Throughout the questionnaire, we avoided vague quantifiers wherever possible, as their interpretation can be inconsistent across respondents and contexts. For instance, saying that a language is used often in a certain context can mean one thing in the UK and another thing in more multilingual contexts, such as South Africa. Even within the same community, respondents might interpret vague quantifiers in diverse ways. Therefore, for most questions, we relied on natural metrics (e.g., *once or twice a week*, *every day*, etc.). Nevertheless, for a small number of questions, it was challenging to identify a natural metric that would be applicable to different respondents and contexts and that would not pose a severe burden on respondents in making a choice. As noted by Dillman et al. (2014, p. 151), vague quantifiers might be an optimal solution when it is impossible to get precise measurements. Therefore, we kept the vague quantifiers only in the following six questions: the frequency of overheard speech at home (Q.177), the proportion of speakers with a high proficiency in a specific language (Q.34), preferred languages in particular contexts (Q.62, Q.64, Q.66), and child's willingness to speak a language (Q.70).

Removing numeric labels from vague quantifiers. In the few cases where we used vague quantifiers (e.g., *sometimes*, *all of them*, etc.) we did not label these categories with numbers (e.g., from 1 to 4) in order to avoid giving an impression that they are equal distance from each other. The same point should be considered in the data analysis, that is, the users should avoid making the assumption of equal distance between these points.

Choosing between unipolar or bipolar scales. In unipolar scales, zero falls at one end of the scale (e.g., from *not at all important* to *very important*), while in bipolar scales, the zero point falls in the middle or it tips from positive to negative (e.g., *very likely*, *somewhat likely*, *somewhat unlikely*, *very unlikely*). Mixing these two approaches within a single scale breaks the ordinality of the scale. We avoided such mixing within scales throughout the questionnaire.

Choosing appropriate scale length. Dillman et al. (2014, p. 152) generally advise limiting the number of points to four or five. All our ordinal scales contain up to five points. This count does not include nonsubstantive options (e.g., *I don't know*), and as advised by Dillman et al. (2014, p. 154), when these options appear, we placed them at the end of the scale rather than in the middle. In terms of directionality of the scales, we always started with a negative end (e.g., from *never* to *every day*), apart from a few questions which have a binary yes-no response format, where the positive option (*yes*) comes first. Questions about preferred languages in three different contexts also had a specific response format in which the scale starts with the most positive option for one of the selected languages yet implying the most negative option for other languages (see questions Q.62, Q.64, Q.66).

Providing balanced scales with relatively equal distances. We applied this guideline wherever possible. However, as discussed above, the use of natural metrics does not allow for equidistance between the points of the scale. This needs to be taken into account when interpreting the data. In four cases, we used four-point scales to avoid the possibility of a neutral option (Q.182, Q.183, Q.184, Q.185). These questions aim to assess the child's skills in comparison to other bi/multilinguals, as well as whether caregivers have any concerns about the child's language development. The scale is therefore split into two points with a negative connotation (e.g., *a child has many more difficulties* or *a child has a few more difficulties*), and two points with positive connotation (*a child is no different from others* or *a child is better than others*). This symmetry allows the scales to be balanced, in spite of the absence of a mid-point.

Choosing direct or construct-specific labels to mark the points on the scale. We avoided the common practice of relying on the agree/disagree format for response options, as it tends to elicit ambiguous responses. For instance, asking *To what extent do you agree or disagree that the child is willing to speak English?* does not directly estimate the extent of the child's willingness to speak that language. Construct-specific labels in the answer options (e.g., *willing* as in *(almost) never willing* to *(almost) always willing*) are more appropriate and were therefore used in the questionnaire. Additionally, throughout the questionnaire, we avoided the agree/disagree format to minimise the risk of acquiescence (i.e., respondent bias towards agreement).

Choosing appropriate answer spaces. This guideline refers to the use of visual analog scales, radio buttons, or drop-down menus. As explained above, we used visual analog scales (i.e., sliders) for questions quantifying language exposure and use to minimise the calculation effort on the participants, which would likely exist with radio buttons or drop-down menus. For questions where response scales were ordinal or nominal, we found the use of radio buttons and drop-downs more adequate than turning visual analog scales from a 0-100 continuum into a slider with four or five points. In this case, the sliders would be rather inadequate, as visually they would imply equal distance between the points, which is not necessarily the case, as discussed above.

As suggested by Dillman et al. (2014, p. 141), there is no consensus whether the radio button format or the drop-down menu format takes longer to complete. They do point out that the drop-down format might cause some difficulties for participants using a scrolling mouse (as it can cause an inadvertent change of their answer). Also, the long drop-down menus might bias the respondents to select the options that appear first on the list before reading through all the options. In our questionnaire, the point about a long list of options is relevant only when asking about the country of residence or birth. However, it is unlikely that the respondents will be biased in these cases by the first few countries that they see, as they will not need to read through all the options to know which option applies to them.

As in some sections of the questionnaire we included an overarching question with three or four sub-questions (e.g., asking about the frequency of three types of language mixing), visually, drop-down menus were more adequate, as radio buttons would spread out the sub-questions on the screen. With drop down menus, respondents can easily process all the sub-questions before even seeing the response options. Therefore, we chose drop-down menus as a default. However, we used radio buttons in a few instances when deemed necessary, especially when more than one option can be selected (e.g., select one or two main caregivers of the child - NQ.19). In some questions, it was also necessary to combine the two formats. For example, when answering about a preferred language in a certain context, the participants first need to choose if they prefer a language almost always or often by selecting it in a radio button format and then from a drop-down list of languages select which language is preferred (Q.62, Q.64, Q.66).

Labelling all categories verbally. No matter the number of points on our scales, we made sure to label each of them rather than just labelling the ends of the scale (or the ends and the middle) and leaving the unlabelled values open to interpretation.

This section has outlined how the psychometric literature informed the formulation of the questions and response scales and some aspects of the layout and functionality of the online questionnaire. We attempted to follow best practice, and to strike an optimal balance between conflicting desiderata, with the aim of minimising the risk of measurement error by reducing the cognitive burden on respondents and striking a realistic balance between precision and realism (in what we are requesting respondents to achieve).

Question order

The ordering of questions was determined by three principles: (i) limiting cognitive effort, (ii) limiting response bias, and (iii) meeting technical requirements of the online implementation. Regarding the first two principles, we followed the guidance of Dillman et al. (2014, Chapter 7).

Questions were arranged thematically, as per the matrix defined as outcome of the Delphi study (De Cat et al., 2021). The questionnaire starts with the background information module, as it asks about information likely to be salient and straightforward to provide (e.g., child's name, country of birth, etc.). From a technical point of view, this module also needs to come first as it is variable-setting for the rest of the questionnaire: some of the answers to these questions directly condition either response options or question wordings in some follow-up modules. For instance, the names of household members listed in this module will appear in the question stems later on when asking about the child's language exposure and use with each household member.

To limit cognitive effort, we organised questions by time and place within each module. Language exposure and use is first documented in the current period, and then chronologically for each past period defined by the respondent. Questions about exposure and use are grouped by physical context (e.g., home vs. school vs. the local community) rather than by exposure vs use. Within each context, the first question asks about the use of each language when an interlocutor or a group of interlocutors is addressing the child, followed by a question about languages used by the child when addressing each of these interlocutors or groups of interlocutors. Schematically this results in the nesting shown below:

- Typical week in the current year
    - Context 1 (home)
        - Exposure (experienced by the child)
            - Interlocutor 1
        - Use (produced by the child)
            - Interlocutor 1
        - Exposure (experienced by the child)

- Interlocutor 2
- Use (produced by the child)
  - Interlocutor 2
- Context 2 (school/daycare)
  - Exposure (experienced by the child)
    - Teachers
  - Use (produced by the child)
    - Teachers
  - Exposure (experienced by the child)
    - Friends
  - Use (produced by the child)
    - Friends
- Context 3 (local community, excluding school/daycare and home)
  - Exposure (experienced by the child)
    - Friends
  - Use (produced by the child)
    - Friends
  - Exposure (experienced by the child)
    - Adults
  - Use (produced by the child)
    - Adults
- School/daycare holidays in the last 12 months
  - Exposure (experienced by the child)
    - Adults
  - Use (produced by the child)
    - Adults
  - Exposure (experienced by the child)
    - Other children
  - Use (produced by the child)
    - Other children

To limit potential response bias, the questions about language proficiency precede the ones about parental satisfaction with the child’s development in each language. The module about attitudes towards language mixing precedes the module about the frequency of language mixing, because the latter module is prefaced by a note saying that language mixing is a natural phenomenon in many bilingual/multilingual communities (to encourage respondents to not under-report language mixing practices). That statement had to appear after the respondent had documented their own attitudes towards language mixing.

Originally, we had intended to ask about risk factors at the end of the questionnaire, to avoid asking parents about their concerns regarding their child’s language development at the outset (which might affect how they report on the child’s proficiency later in the questionnaire). But the risk factors module had to be fore-fronted to appear just after the background information, for technical reasons: the age of first word combinations (which is part of the risk factors module) is used to automatically determine the age from which language production is documented in the language exposure and use module (in relation to the onset of cumulative use of languages by the child).

Visual design

As noted by Dillman et al. (2014, p. 172), the visual design plays an important role in self-administered surveys since it can help or hinder the response process. The visual design of the questionnaire was informed by Dillman et al.'s Chapter 6 and by advice from the professional application designers (CastlegateIT).

To enhance clarity, questions were presented one-per-screen, except sequences of questions which were very closely related (as this would have resulted in a lot of redundancy of wording from one screen to the next). If there were many closely related questions, they could be distributed across more than one screen. For instance, questions about language exposure and use with teachers and school friends were grouped together on one screen when asking about school, while questions about language exposure/use with adults and other children during school holidays were presented on another screen. In cases when several sub-questions were presented on one screen, we made sure that this never included more than four sub-questions. Each sub-question was also framed individually to make them prominent. Apart from these thematic sub-question clusters, we implemented the *one question per screen* rule.

In the question stems, which set the frame of reference for a group of questions, we used bold fonts to emphasise important points. For instance, an overarching statement that applies to all sub-questions on the screen would often be put in bold (e.g., *Think of a typical week in the current year*). Bold face was also used to highlight which word(s) varied across a set of questions with very similar formulation (e.g., the name of a language, of a context, or a person).

We minimised the use of matrices and grids. These were only used for questions related to the daily schedule of the child where respondents need to indicate with whom the child spends each hour between 6 am and midnight.

Any special instructions, explanations or definitions were integrated in the questions rather than singled out as free-standing before the question, to maximise the likelihood that respondents read them. Examples of these include, for instance: questions about certain activities (where in the question stem we listed the activities to consider); the question about the main caregivers of the child (where we included the definition of the main caregiver); several questions including a note to move to the next question if the current question is irrelevant.

In the few open-ended questions where a single answer is required, we provided a single box rather than multiple or larger boxes. Where applicable, we also specified the required unit next to the box. For instance, the age when the child spoke their first words in any language is documented in two adjacent boxes: one for years and one for months. We also conditioned these boxes to accept only numbers rather than text. In the boxes requiring dates as answers, we included the required date format within the box (e.g., mm/yyyy or dd/mm/yyyy).

In those questions where we used radio buttons, we used circular radio buttons when only one option can be selected (for instance, when we ask whether the child goes to school or not - yes or no, CQ.12); the square radio buttons were used when more than one option can be selected (for instance, when indicating up to two main caregivers, NQ.19, or the members of the household who do not speak, CQ.19).

Finally, at the bottom of the screen, we included a progress bar whose sections correspond to the number of modules in the questionnaire. Once a module is completed, a section of the progress bar fills up. The size of the sections on the progress bar are proportional to the length of the respective module. In this instance, we acted against the advice of Dillman et al. (2014, pp. 325-326), who recommend avoiding progress bars. This was based on consultation with experts and on the results of our evaluation study (see below).

Online implementation

We collaborated with professional web developers CastlegateIT to create an optimised, user-friendly online tool. This allowed the customisation of the questionnaire (e.g., to determine what modules and sub-modules could be excluded, without jeopardising the functionality of the online questionnaire) as well as personalisation to the respondent (e.g., using the information they provided to determine how subsequent questions are worded or presented). Online implementation also allowed the use of intuitive methods of quantification that prevented inconsistent responses. For instance, the sliders and associated pie-charts to document language exposure and use automatically adjust the values for one language according to those of the other language, so that respondents cannot report, e.g., 60% exposure to Russian at the same time as 60% exposure to Estonian for the same child in a particular context or with a particular interlocutor.

As noted before, online implementation also allowed the automated calculation of specific scores, such as a richness score in each language, current weighted measures of language exposure and use, as well as cumulative measures of language exposure and use. These features are likely to make the tool more appealing to many users but also guarantee comparability across studies in terms of how particular scores were calculated. Finally, all responses get anonymised automatically. In the raw data file and in the calculations file, each child is automatically assigned a pseudonym ID. In a separate .csv file, children’s individual names are linked to their IDs so that they can be tracked by authorised researchers/practitioners if necessary.

Evaluation phase

Cognitive interview principles

To assess the clarity and ease-of-use of the questionnaire, we ran a series of cognitive interviews (following Dillman et al., 2014, pp. 243-244 and Wills, 2015). Cognitive interviews differ from regular interviews (in which an interviewer asks a question from the tool and a respondent answers) in their use of verbal report methods (Willis, 2004, p. 24). Two methods can be used: think-aloud verbal probing. In the think-aloud method, a respondent is asked to verbalise their thoughts as they answer the survey questions. During the verbal probing, however, after a respondent provides an answer to a survey question, the interviewer asks additional probe questions for the purposes of revealing the respondents' thought process or interpretation of the survey question. While these two report methods seem rather clear cut, in practice there is no generally accepted standard in terminology or in the practice of cognitive interviews. In fact, cognitive interviews often include a combination of a think-aloud and verbal probing, and the ways of conducting them, of coding and analysing the data can vary widely (see Willis, 2004, p. 25; DeMaio & Landreth, 2004).

As explained by Wills (2015), the purpose of a cognitive interview can be reparative or descriptive. In the reparative approach, the aim is to inspect each question, identify their flaws and find a way of improving them to avoid ambiguity, unintended interpretations, or any other inadequacies. The descriptive approach aims to understand better how each question works, and what impact the existing formulations can have on respondents. It doesn't necessarily aim to modify or improve the survey items. In practice, the two approaches are on a continuum as many studies include elements of both (Willis, 2015, p. 20). While our goal was to implement primarily the reparative approach, inevitably some of the data were looked at from the descriptive point of view, with the aim to acknowledge potential limitations of the questionnaire.

As both think-aloud and verbal probing have advantages and disadvantages (summarised in Willis, 2015, p. 38, Table 3.2), we opted for a hybrid approach combining the two methods. We asked respondents to think aloud as they answered the questions, with further verbal probes by the interviewer as required. While some probes were immediate (i.e., asked straight after the target survey question), we tried to follow a modification of Willis' (2015, p. 44) recommendation to use a hybrid approach of retrospective probing. That is, after each of the seven modules, we probed the respondents about any outstanding points which were not clear from their think-aloud or some immediate probes used throughout the module. In this way, the respondents' thought process was not constantly interrupted with immediate probes, but probes were not delayed until the end of the questionnaire, by which point respondents might no longer remember the details of their previous reasoning. Any data collected after the completion of the questionnaire was in relation to respondents' overall impression and any outstanding comments that they wanted to make.

Our method

Two interviewers were in charge of conducting the evaluation of the Q-BEx questionnaire. The first interviewer prepared general guidance on conducting the interview primarily based on Willis (2015). He also drafted the interview protocol, which was then approved by the rest of the Q-BEx team. The protocol consisted of two parts: (a) the instructions for the interviewers, and (b) a list of probe questions. The instructions outlined guidance for interviewers on how to perform before, during and after the interview. The list of probes included more generic (neutral) probes, which could be asked in relation to most questions and likely in an immediate format (e.g., *What made you select that option?*, *What does X bring to mind?*, *Tell me more about that.*). There was also a list of more specific probes relating to questions within each module (e.g., for Q.33: *You were asked about the number of people who speak to the child in a specific language on a regular basis. What do you consider 'regular basis'?*). These specific probes were asked retrospectively following each module in case they had not been addressed already via think-aloud or some of the immediate probes.

The first interviewer (Draško Kašćelan) trained the second interviewer[7] by providing them with a list of relevant literature and notes to digest, as well as by going through the interview protocol with them. They also ran a test interview together in which the second interviewer could familiarise herself with the protocol and ask any questions about potential difficulties that might arise. Considering the fact that two interviewers were included in the data collection, this could have consequences on potential differences in their note taking (i.e., data collection). In order to minimise this as much as possible, both interviewers were using the same interview protocol. Furthermore, after each interviewer conducted three interviews, they met to discuss their note-taking approach, as well as the issues that they had encountered and how they went about them. Throughout the rest of the study time, the interviewers were closely communicating in case they needed to consult on how to go about certain situations (for instance, when it might be necessary to guide participants in case the question interface confuses them to the point that they cannot continue on their own, and how to document this in the interview notes). Originally, we planned to have each interviewer conduct the same number of interviews. However, due to individual schedules of the interviewers, as well as the availability of the respondents, this was not achieved. The first interviewer conducted 22 interviews (with 18 caregivers and with 4 children) and the second interviewer conducted 8 (with 7 caregivers and with 1 child).

All 30 interviews[8] took place over a period of three weeks (in June and July 2021). Before the interviews, the caregivers were sent a briefing document and a consent form. Following that, the interview date and time were arranged with each participant. Each participant was rewarded with a gift voucher for their time. This study was approved by the University of Leeds ethics committee (reference: FAHC 20-074).

---

[7] We are grateful to Anna Hamilton, who was the second interviewer in the evaluation study.

[8] In addition to these 30 interviews, one caregiver completed the questionnaire on their own and sent us their feedback via a Google Form feedback form. This caregiver and their child were based in Spain, and the languages of the child were Galician and Spanish.

Descriptive analysis

Among the 30 participants, there were 25 caregivers and 5 children. One caregiver completed 4 out of 7 modules due to the shortness of the session that was arranged. Other 29 participants completed all 7 modules (i.e., the complete questionnaire in its longest version). The participants were recruited from 14 countries in order to obtain opinions from bilinguals coming from different contexts. The distribution of participants in relation to their country of residence can be seen in Figure 1 (for caregivers) and in Figure 2 (for children). The languages spoken/understood by the children included: Afrikaans, Arabic, Catalan, Chinese, Danish, Dutch, English, French, Georgian, Hindi, Icibemba, Italian, Japanese, Korean, Malay, Marathi, Northern Sami, Norwegian, Potwari, Russian, Serbo-Croatian, Spanish, Telugu, Xitsonga.

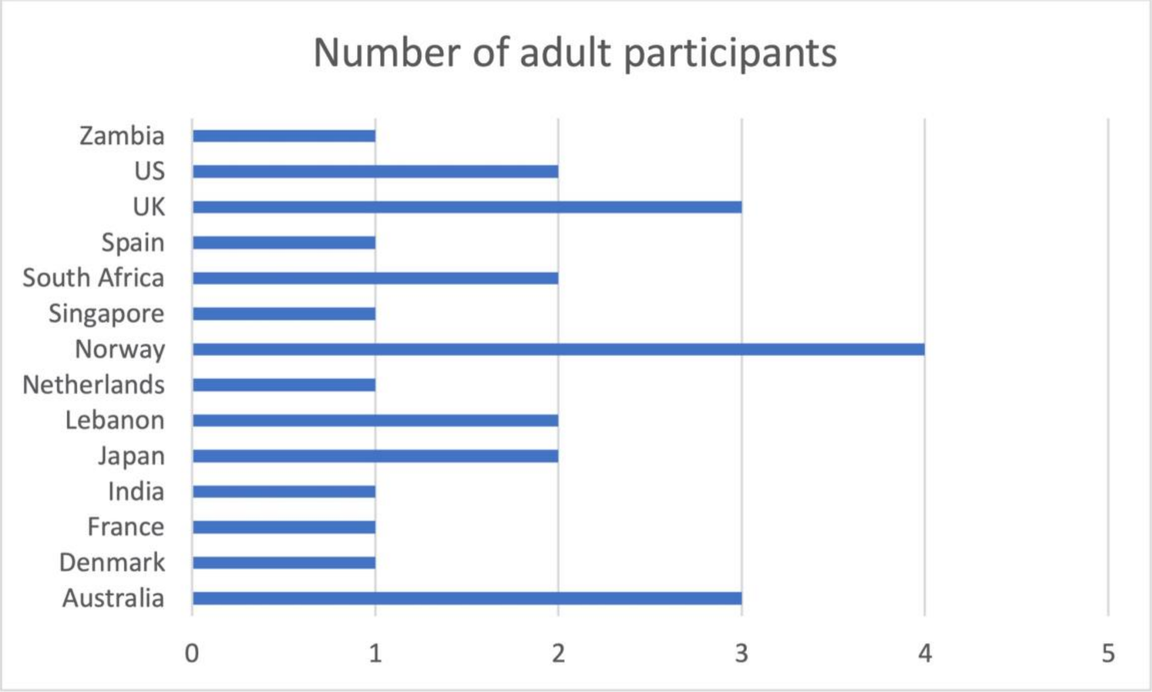

Figure 1. Distribution of caregiver participants per country of residence

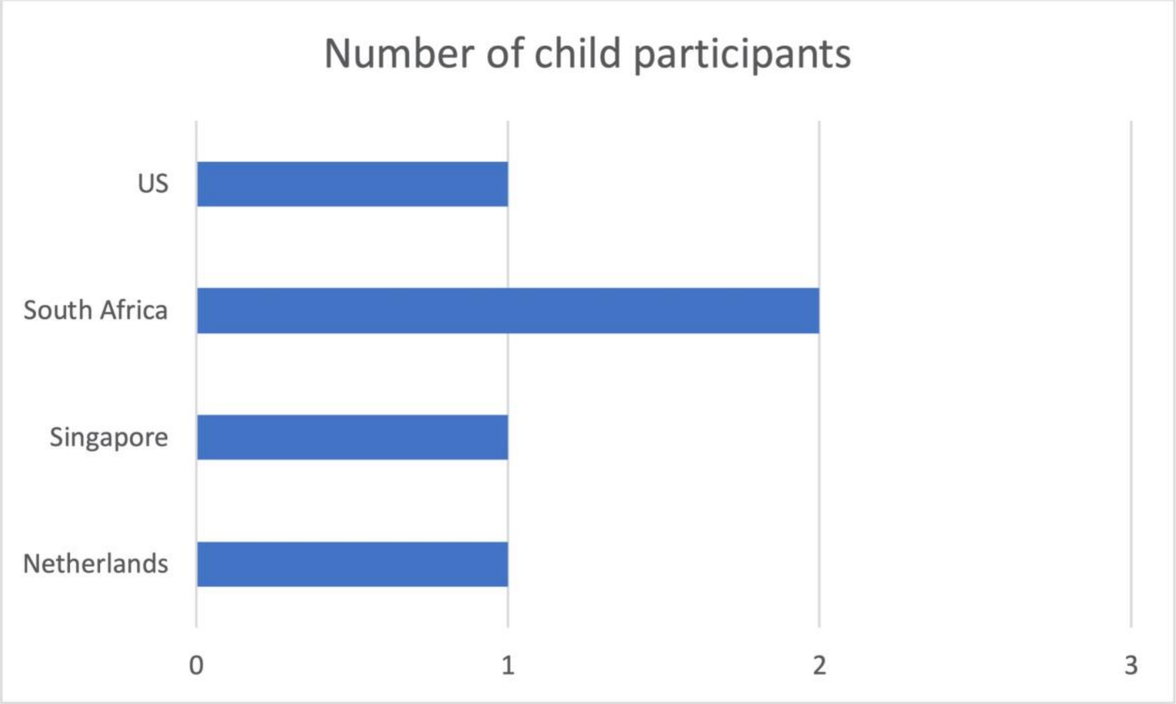

Figure 2. Distribution of child participants per country of residence

Among the 30 interviewed participants, 8 of them were 4 caregivers and their 4 children. In these 4 pairs, both the child and the caregiver completed the respective questionnaire version about the same child.

Data documentation included note taking by the interviewers as well as video recording. The recordings were used to enrich the notes and clarify any ambiguities in notes when required. Following this, the recordings were deleted. As suggested by Willis (2015, pp. 57-58), we found it unnecessary to transcribe the recordings, especially considering that our primary aim (i.e., the reparative aim) could be achieved more successfully by relying on notes which include not only participant's thoughts/opinions/verbatim statements, but also the interviewer's observations.

Willis (2015, pp. 58-124) presents five analytical models for the processing of cognitive interview data: text summary, cognitive coding, question feature coding, theme coding, and pattern coding. While text summaries are classified as an uncoded approach, the remaining four are coded approaches (the former two are top-down - that is, they include assigning codes to the data; the latter two are bottom-up - that is, they require building codes from the data). Considering that the main aim of our evaluation was to repair any issues in relation to question interpretation and the usability of the tool, we found text summaries to be the most adequate choice. By not reducing the data to individual codes, we kept the richness of the dataset which enabled us to identify specific deficiencies with the questions. Initially, we planned to compose a one-page text summary of each interview. Nevertheless, considering the length of the questionnaire, this could have led to the reduction of question-specific points to potentially uninterpretable notes for the adequate diagnosis of problems to fix. Therefore, our notes represent a rich summarization with references to specific questions and a distinction between respondents’ comments and interviewers’ observations.

Main analysis

Following the data collection, the notes were read through by the first interviewer. Next to each identified issue, the interviewer noted the problem and left a suggestion of how this might be repaired. While the frequency of each issue partly played a role in our estimation of how serious those issues were, we avoided relying on quantification. In this way, we enabled identifying major issues even if they occurred only in one instance (i.e., with one respondent). An example of this included a case of a parent who was using the Internet Explorer to fill in the questionnaire. By using this browser, all language exposure and use questions which included the sliders and pie charts showed numbers (percentages) as the parent was moving the sliders. This was problematic since the parent believed that for each context/person, they needed to add up exposure or use of all three languages to 100%, without realising that the platform adjusts this automatically. Such a predicament made the experience of answering these questions burdensome for the parent. It also shifted their focus on thinking about these estimations in percentages, which was not necessarily the way in which other respondents who did not see the numbers on the sliders conceptualised and approached these questions. Therefore, even though this issue occurred only with one respondent, it was important to acknowledge it as a major problem due to its implications for the user experience and potential measurement errors.

In addition to reading through notes from each interview and flagging potential issues, we created a text summary of issues encountered in each module in order to obtain a general overview for each part of the questionnaire. This was helpful for two reasons. First, it allowed us to observe if we missed something generally problematic in each module which might have been missed in the reading of individual notes. Second, this summary contains points regarding certain questions which might cause issues in the future use of the questionnaire. Some of these will be discussed in the limitation section (in a future version of this document, to be added after the validation phase).

Resulting changes

From the issues identified, the major ones were addressed with the Q-BEx team and decisions were made about required changes. Minor issues were discussed between interviewer 1 and the project PI, who made decisions on required follow-ups. Here we thematically summarise some of the main changes implemented following the cognitive interviews:[9]

- Clarifying certain questions or parts of the question. For example:

---

[9] A spreadsheet containing specific changes for each modified question is available upon request.

    - Adding a note that the languages of the child should be listed no matter how well or how often they are spoken;
    - Adding a definition explaining what we mean by *a caregiver;*
    - Changing *full name* to *name and surname;*
    - Clarifying that by *holidays* we mean *school or day care holidays;*
    - Adding a note in the child's daily timetable that the slots when the child is sleeping or is alone should be left unmarked.
- Changing a number of points on a response scale
    - Adding a *non-binary* option on the question about the gender identity of the child;
    - Adding option 0 for a number of regular speakers in a particular language and updating the rest of the options accordingly;
    - Adding options *in three to five conversations per day* and *in more than five conversations per day* in the language mixing module.
- Adding questions
    - about the child's country of residence;
    - about members of the household who do not speak.
- Removing questions
    - about the child's history of previous residence.
- Adding error or warning messages (e.g., encouraging caregivers to guess the age of first words and sentences of the child in case they try to leave these questions unanswered).
- Putting in bold specific words or phrases to make them stand out (e.g., adults, local community).
- Reversing the order of a response scale to match the order of other scales in the questionnaire (negative to positive).

The implementation of these modifications produced the final version of the questionnaire, which was then tested for functionality and proofread by the Q-BEx team before the official release on 02 August 2021. The next step included final readability updates before the validation of the tool with bilinguals in the UK, France and the Netherlands.

Readability updates

The final check of the questionnaire by the Q-BEx team revealed that some formulations were not optimal in terms of readability: the structure of some sentences was unnecessarily complex, and some words might be challenging for non-native speakers. Consequently, we reworked the questionnaire according to the following principles:[10]

- Changing any passive sentences into active ones.
- Fronting context-setting adverbials (e.g., putting a phrase *In regular school/day care* at the beginning of a relevant question rather than in the middle).

---

[10] A spreadsheet containing specific changes for each modified question is available upon request.

- Ensuring that the vocabulary used did not exceed the competence expected at level B2 of the Common European Framework of Reference for languages.
- Eliminating some logical fallacies from the child version (e.g., asking children whether they go to day care, even though the child version is to be distributed to teenagers).
- Simplifying complex phrases/sentences wherever possible (e.g., changing *of the same or similar age* to *of a similar age*).

Work in progress: Questionnaire validation

We are currently conducting the Q-BEx questionnaire validation study in the UK, France, and the Netherlands, with monolingual, bilingual, and trilingual children between the ages of 5 and 9. We expect the findings of the study to be available by October 2023.

During the validation study, based on our experience of using the tool as well as the experience and feedback from other tool users, we made the following changes to the questionnaire functionality in order to improve its usability:

- **We added a *Save and abandon* button.** The users of the questionnaire (i.e., researchers and practitioners) can now include a *Save and abandon* button when creating a questionnaire link. In this way, the questionnaire respondents can leave the survey before completing it and save the answers which they have provided. Note, however, that the respondents will not be able to return to the survey at a later time. A pop-up message will inform the respondents of this before they confirm leaving the survey and saving their responses. We hope that this button will be of use to researchers/practitioners who need to administer longer versions of the questionnaire. By using this button, they may collect at least some data from respondents in case they don't answer all questions.
- **We replaced the existing *Calculator* output with the *Ordered response data and calculations* output.** The initial *Calculator* output included calculations and raw data of all participants. However, as each participant has a unique family constellation and circumstances (e.g., different number of people at home, varying language history), answers from each respondent will likely produce a different number of variables in the output. The initial *Calculator* output first presented variables that were relevant to all participants, followed by variables unique only to some of them. However, this made the output file difficult to navigate and inspect. In the new *Ordered response data and calculations* output, we include raw data and calculation variables always ordered in the same way. If certain variables are irrelevant to some participants (e.g., if they don't speak a third language, or if they don't have five other children at home), those cells will remain empty for those specific participants. The only data not included in the *Ordered response data and calculations* output is the average time that the child spends with each individual, group of people, or context on a typical weed day, on odd days, on a typical weekend day, and on a typical day during holidays. This data can be found in the raw data files of each participant.

- **We added a *Log in* button on the project website.** After registration, the questionnaire users (i.e., researchers/practitioners) had to use a specific link rather than the project website link to access the questionnaire platform. We have now introduced a *Log in* button on the questionnaire website, which allows the users to log in and access the questionnaire platform in a more straightforward way.
- **We added short report outputs aimed at practitioners (i.e., teachers and speech and language therapists).** By considering feedback from practitioners, we designed a short report which contains a summary of some language history data of relevance to teachers and speech and language therapists. The short report can be downloaded in English or Dutch. A translation in French is planned for near future. The reports contain the following data:
    - The child's name
    - Basic demographic data about the child's age, languages, country of birth, number of siblings, date the child started school, languages spoken by their caregivers, age of first exposure to each language
    - If relevant (sub-)modules are distributed: cumulative estimates for exposure and use in each language
    - Concerns about the child's language development and periods when the child missed school (if relevant)
    - If relevant (sub-)modules are distributed: current weighted estimates for exposure and use in each language
    - If relevant (sub-)modules are distributed: current unweighted estimates for exposure and use in 4 contexts (home, local community, school, during holidays)
    - If relevant (sub-)modules are distributed: richness of linguistic experience in each language
    - If relevant (sub-)modules are distributed: proficiency estimates for each language (listening, speaking, reading, writing)
    - Further information on each of the above measures

Next steps: Post validation documentation and limitations

Upon completion, the findings and the design implications of the validation study will be documented here together with a list of limitations of the Q-BEx tool.

References

Blom, E., & Soderstorm, M. (2020a). Introduction to the Special Issue on the Influence of Input Quality and Communicative Interaction on Language Development Part 1. *Journal of Child Language*, *47*(1), 1-4.

Blom, E., & Soderstorm, M. (2020b). Introduction to the Special Issue on the Influence of Input Quality and Communicative Interaction on Language Development Part 2. *Journal of Child Language*, *47*(2), 265-266.

Brice-Heath, S. (1983). *Ways with Words: Language, Life and Work in Communities and Classrooms*. Cambridge University Press.

de Almeida, L., Ferré, S., Morin, E., Prévost, P., dos Santos, C., Tuller, L., Zebib, R., & Barthez, M. A. (2017). Identification of bilingual children with specific language impairment in France. *Linguistic Approaches to Bilingualism*, *7*(3-4), 331-358.

De Cat, C., Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & The Q-BEx Consortium (2021). How to quantify bilingual experience? Findings from a Delphi consensus survey. https://osf.io/ebh3c/

DeMaio, T. J., & Landreth, A. (2004). Do Different Cognitive Interview Techniques Produce Different Results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer (eds), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 89-108). John Wiley & Sons, Inc.

DeVellis, R. F., (2017). *Scale Development: Theory and Applications*. Fourth Edition. SAGE Publications, Inc.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Fourth Edition. John Wiley & Sons, Inc.

Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in Second Language Research: Construction, Administration, and Processing*. Second Edition. Routledge.

Iqbal, S., & Pipon-Young, L. (2009). The Delphi method. *Methods*, *22*(7), 598-601.

Kašćelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & De Cat, C. (2021). A review of questionnaires quantifying bilingual experience in children: Do they document the same constructs? *Bilingualism: Language and Cognition*, 1–13.

Li, P., Sepanski, S., & Zhao, X. (2006). Language history questionnaires: A web- based interface for bilingual research. *Behavior Research Methods*, *38*(2), 202–210.

Paradis, J. (2011). Individual Differences in Child English Second Language Acquisition: Comparing Child-Internal and Child-External Factors. *Linguistic Approaches to Bilingualism*, *1*(3), 213–237.

Paradis, J., Emmerzael, K., & Sorenson Duncan, T. (2010). Assessment of English Language Learners: Using Parent Report on First Language Development. *Journal of Communication Disorders*, *43*, 474–497.

Tuller, L. (2015). Clinical use of parental questionnaires in multilingual contexts. In S. Armon-Lotem, J. de Jong and N. Meir (eds), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 301–330). Multilingual Matters.

Unsworth, S., (2016). Quantity and quality of language input in bilingual language development. In E. Nicoladis and S. Montanari (eds.), *Bilingualism Across the Lifespan: Factors Moderating Language Proficiency* (pp. 103–121). De Gruyter Mouton.

Willis, G. B. (2004). Cognitive Interviewing Revisited: A Useful Technique, in Theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer (eds), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 23-44). John Wiley & Sons, Inc.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.