Bilingualism: Language and Cognition

cambridge.org/bil

Research Article

*The Q-BEx Consortium: Nouhad Abou Melhem, Ashley Adams, Eva Aguilar-Mediavilla, Nadia Ahufinger, Shanley Allen, Llorenc Andreu, Effrosyni Froso Argyri, Sharon Armon-Lotem, Jacqueline Arsenault, Heather Baer, Colin Baxter, Lisa Bedore, Elma Blom, Mirjam Blumenthal, Ute Bohnacker, Claudine Bowyer-Crane, Krista Byers-Heinlein, Allegra Cattani, Shravasti Chakravarty, Vicky Chondroggiani, Dayna Camilleri Clarke, Joanne Cook, Daniel Cubilla-Bonnetier, Ineta Dabasinskiene, Kankan Das, Rachael Davis, Angela de Bruin, Annick De Houwer, Martine Diab, Maaike Diender, Helen Drummond, Heli Elovaara, Victoria Farrell, Karina Fascinetto-Zago, Giulia Filippi, Caroline Floccia, Max Freeman, Silke Fricke, Margaret Friend, Carolyn Gelenter, Samina Ghafoor, Kleanthes K. Grohmann, Matt Hall, Ewa Haman, Cornelia Hamann, Riikka Härkönen, Jeni Harrison, Carly Hartshorn, Maha Hassan, Erika Hoff, Nayr Ibrahim, Sladjana Indjin, Katy Isaac, Susan Joffe, Holly Joseph, Maria Kambanaros, Kathryn Kashyap, Olga Kepinska, Andrea Kewin, Ekaterini Klepousniotou, Tanja Kupisch, Francesca La Morgia, Gabrielle Lai, Nienke Lam, Jane Le Roux, Mia Le Roux, Annina Manninen, Virginia Marchman, Theodoros Marinis, Prisca Martin, Lourdes Martinez Nieto, Chantal Mayer-Crittenden, Natalia Meir, Liz Metcalfe, Eléonore Morin, Lina Mukhopadhyay, Victoria A. Murphy, Vishnu KK Nair, Marie Newton, Arja Nieminen, Silvia Nieva, Audrey Noel, Ewelina Urszula O'Donnell, Therese O'Sullivan, Ciara O'Toole, Weronika Ozpolat, Felicity Parry, Michelle Pascoe, Rupam Patel, Vrishali Patil Ingle, Sean Pert, Anne-Gaelle Piller, Eva Poort, Christine Potter, Eveliina Rantanen, Saiqa Riasat, Lidia Rodriguez, Wiebke Scharff Rethfeldt, Bethany Faye Schwartz, Irina Sekerina, Miquel Serra Raventós, Saleh Shaalan, Jinder Singh, Beth Skelton, Vicky Slaughter, Anne Margaret Smith, Sini Smolander, Neal Snape, Marina Sokolova, Antonella Sorace, Lisa Spinney-Hutton, Frances Jane Stokes, Luke Swift, Vasim Salim Tamboli, Elena Theodorou, Elin Thordardottir, Ianthi Tsimpli, Rachael Tuckley, Olga Urek, Liza van den Bosch, Klarien van der Linde, Laetitia Vanbruwaene, Josje Verhagen, Virve Vihman, Adriana Weisleder, Gillian Wigglesworth

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

Received: 29 January 2021
Revised: 5 May 2022
Accepted: 5 May 2022
First published online: 4 July 2022

Keywords:
quantifying bilingualism in children; Delphi consensus survey; researchers; speech and language therapists; teachers

Address for correspondence:
Cécile De Cat, University of Leeds, Michael Sadler Building, Woodhouse, Leeds LS2 9JT, United Kingdom; c.decat@leeds.ac.uk

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

CAMBRIDGE UNIVERSITY PRESS

How to quantify bilingual experience? Findings from a Delphi consensus survey

Cécile De Cat[1], Draško Kašćelan[2], Philippe Prévost[3], Ludovica Serratrice[4], Laurie Tuller[3], Sharon Unsworth[5] and The Q-BEx Consortium*

[1]University of Leeds and UiT Arctic University of Norway; [2]University of Leeds; [3]Université de Tours; [4]University of Reading and UiT Arctic University of Norway and [5]Radboud University

Abstract

While most investigations of bilingualism document participants' language background, there is an absence of consensus on how to quantify bilingualism. The high number of different language background questionnaires used by researchers and practitioners jeopardises data comparability and cross-pollination between research and practice. Using the Delphi consensus survey method, we asked 132 panellists (researchers, speech and language therapists, teachers) from 29 countries to rate 124 statements on a 5-point agreement scale. Consensus was pre-defined as 75% agreement threshold. After two survey rounds, 79% of statements reached consensus. The need for common measures to quantify bilingualism was acknowledged by 96% of respondents. Agreement was reached to document: language exposure and use, language difficulties, proficiency (when it cannot be assessed directly), education and literacy, input quality, language mixing practices, and attitudes (towards languages and language mixing). We discuss the implications of these findings for the creation of a new tool to quantify bilingual experience.

1. Introduction

Most investigations of bilingualism have moved away from classifying participants as bilingual without some documentation of language history and experience. There are several groups of professionals, such as researchers, teachers, and speech and language therapists (SLTs), who often have to document bilingual experience. Both within and across these groups, documenting bilingualism might be required for a range of different purposes: assessing children's development in each language, assessing the risk of a language disorder, assessing their learning potential, assessing their cognitive development, assessing their well-being, etc. This documentation is usually based on language background questionnaires. When studying children, the information is mostly obtained from caregivers, sometimes teachers, or even children themselves.

In a recent review of 48 questionnaires used to quantify bilingualism in children, Kašćelan, Prévost, Serratrice, Tuller, Unsworth and De Cat (2021) observed substantial variation in the documentation of key dimensions of bilingualism, such as language exposure and use, activities in each language, and language skills. For instance, across these questionnaires, exposure and use were documented with particular interlocutors, in specific contexts, during various activities, or as a combination of these. The level of detail varied greatly depending on the interlocutors, contexts, and activities specified in the questionnaire, and depending on whether informants could add categories (e.g., interlocutors) relevant to their circumstances. Language exposure and language use were usually treated as separate constructs but they were not always documented separately, depending on the wording of the relevant questions (which asked, for example, about the frequency of "speaking or hearing" each language). Frequency was usually documented on a Likert scale, but questionnaires used scales of different lengths and with different labels (e.g., quantifying adverbs, percentage-based points, a combination of frequency adverbs and percentages). Some questionnaires estimated frequencies through open-ended questions. Several questionnaires documented the time that the child spent with each interlocutor, so that the frequency of language exposure from that interlocutor could be adjusted for the proportion of the time spent with the child.

This considerable variability between questionnaires constitutes one of the major hurdles in making the resulting datasets truly comparable.[1] In line with Marian and Hayakawa (2020), we argue in Kašćelan et al. (2021) that bilingualism research is in need of a consensus on what aspects of bilingual experience to document and how. To achieve such a consensus, an

[1]Initiatives such as the BLC mini-series (Luk & Esposito, 2020) which aim to gather systematic collections of tools used to document bilingual experiences are a step in the right direction, but they do not aim to enhance the comparability of measures used across bilingualism research and across sectors.

inclusive and collaborative approach involving both researchers and practitioners (teachers and SLTs) is required. Such an approach can maximise the potential for cross-pollination between research and practice, and enable a construction of common tools to gather the relevant information about children's bilingualism.

We conducted an international Delphi consensus survey, with the aim of informing the creation of a modular tool for quantifying bilingual experience and achieve consensus between different groups (researchers and practitioners). In doing so, we contribute to the literature by "taking the pulse" of current research and practice, gathering experts' opinions on how bilingualism should be quantified. The survey was completed by 132 researchers, teachers and SLTs from 29 different countries who had worked with bilingual children of various ages and from different bi/multilingual contexts.

The paper is organised as follows. First, we briefly explain the Delphi method and review relevant literature. Second, we outline the methodology adopted in our study, and present the results obtained as well as the list of statements reaching consensus among the three stakeholder groups. Finally, we discuss the findings and outline future steps towards the creation of a common set of measures to quantify bilingual experience in children.

1.1 The Delphi method

The Delphi approach is commonly used across disciplines to explore the diversity of opinions among a group of stakeholders or lead them towards a consensus (Iqbal & Pipon-Young, 2009). The stakeholders (also known as panellists) are a representative group of experts on the relevant topic of interest. The technique involves an iterative process, in which the panellists express their opinions on a series of statements through two or more survey rounds. The initial set of statements is sometimes itself generated with an open-ended survey (Thangaratinam & Redman, 2005).

The initial set of statements is presented to the panellists via an online survey, in which they are asked to rate their level of agreement with each statement and (optionally) provide comments to justify their ratings. The following round includes a reduced set of statements, depending on the ratings and comments of the previous round. From one round to the next, the statements which have not yet reached agreement or which are in a predefined proximal zone (i.e., grey area) are retained. If required, they can be reformulated. New statements can also be added if necessary.

In the second (and subsequent) round(s), each statement is presented to panellists along with the distribution of responses from the preceding round as well as the panellist's own ratings. This allows them to reflect on their responses in light of group views, and give them the opportunity to maintain or change their rating as they see fit. This process can be repeated until the consensus is reached, until the predetermined number of rounds is completed, or until it is clear that greater agreement is not possible.

Anonymity is key. It guarantees parity among panellists by giving an equal voice to all, promotes freedom of expression, and limits the risk of bias. The online format allows greater inclusivity, as panellists can be recruited from different geographic areas and complete the survey at their own convenience (within a set time limit).

The design of the present study was informed by guidelines from Hasson, Keeney and McKenna (2000), Iqbal and Pipon-Young (2009), and Thangaratinam and Redman (2005), as well as the review by Diamond, Grant, Friedman, Pencharz, Ling Moore and Wales (2014). We modelled parts of our approach on Bishop, Snowling, Thompson, Greenhalgh and the CATALISE consortium (2016), as well as on Langlands, Jorm, Kelly and Kitchener (2008), and Spain and Happé (2019).

2. Methodology

Our aim was to identify the broad consensus on what needs to be documented rather than ask panellists to prioritise which aspects to document (as this would have depended on their particular area of specialism). To maximise the comparability of measures used across studies and practice, we aim to develop a customisable questionnaire allowing professionals to select what is relevant to their purpose from a large set of consensus-informed questionnaire components.

Ethical approval for the study was granted by the University of Leeds. Our procedure was as follows. To generate the first set of statements for the online survey, we organised a workshop with a group of experts. Subsequently, we conducted a two-round online survey in which a larger group of panellists rated the statements on a 5-point scale, with a possibility to leave comments. It is a common concern in the Delphi approach that repeated iterations can lead to increased panellist attrition (Walker & Selfe, 1996). Therefore, the end of round 2 of the online survey was set as a stopping criterion, no matter the number of statements that reached consensus. This gave panellists the opportunity to change their mind once, and only for those statements in the proximal zone.

In the next two sections, we present the panellist characteristics (section 2.1) and the survey design and procedure (section 2.2). How consensus was defined is explained in the results (section 3.1).

2.1 Panellists/Stakeholders

The initial workshop (which will be described in the next section) brought together researchers (n = 22) and practitioners (n = 14) who have worked with bilingual children and extensively used (and often designed) questionnaires to document bilingual experience. We tried to be as inclusive and representative as possible. The workshop participants included experts in typical as well as atypical language development, and had different lengths of experience in bilingualism research. Their expertise spanned different types of bilingual populations, including speakers of heritage languages with various levels of societal prestige, speakers of two majority languages with equal or unequal societal status, and bidialectal speakers. The researchers came from 11 countries: the UK (8), Germany (3), Canada (2), the Netherlands (2), the US (2), France (1), India (1), Israel (1), Poland (1), South Africa (1), and Sweden (1). The practitioners were mostly local (12 from the UK, 1 from Lebanon, and 1 from France), and they ranged from early career to more experienced.

Following the workshop, we used online expression of interest forms (separate ones for researchers, teachers, and speech and language therapists) to expand and diversify the sample of panellists for the online Delphi survey. The expression of interest forms were advertised on our project website and on social media (Facebook, Twitter). We also emailed 247 individual researchers, practitioners, and relevant organisations (e.g., Comité Permanent De Liaison Des Orthophonistes-Logopèdes de l'UE, Audiology and Speech-Language Pathology Associations around the world,

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

Table 1. Round 1 panellist breakdown per stakeholder group.

Categories	Sub-categories	Distribution by sub-category, Number (%)	Distribution by category, Number (%)
Researchers		68 (41%)	68 (41%)
Practitioners	Speech and language therapist/ pathologist/ logopedist	38 (23%)	58 (35%)
	Teacher	20 (12%)	
	Speech and language therapist/ pathologist/ logopedist, Teacher	0	
Researchers/ Practitioners	Researcher, Speech and language therapist/ pathologist/ logopedist	24 (15%)	38 (23%)
	Researcher, Teacher	12 (7%)	
	Researcher, Speech and language therapist/ pathologist/ logopedist, Teacher	2 (1%)	
	TOTAL		164

the Literacy Association of South Africa, the National Association for Language Development in the Curriculum in the UK).

The expression of interest forms contained questions about the prospective panellists' demographic background and about their experience working with bilingual children. Using these forms, registrations for the survey were submitted by 82 researchers, 61 SLTs, and 27 teachers. We applied the following exclusion criteria: having less than a year of work experience, having never worked with bilinguals, not being able to commit to both rounds of the Delphi survey. This resulted in 13 exclusions (1 researcher, 10 SLTs and 2 teachers), after which 157 stakeholders were retained. In addition, 22 researchers and 14 practitioners from the workshop, as well as 3 researchers who were invited to the workshop but could not attend, were invited to participate in the first round of the Delphi survey. Of the 196 invited panellists, 164 completed round 1 of the survey (response rate: 83%).

Table 1 summarises the distribution of round 1 panellists across self-assigned categories (allowing identification with multiple categories). They came from 30 countries: the UK (52), the US (17), the Netherlands (14), Canada (9), India (7), Spain (7), Finland (6), France (6), Germany (6), South Africa (6), Norway (4), the United Arab Emirates (4), Cyprus (3), Israel (3), Ireland (2), Italy (2), Lithuania (2), Reunion Island (2)[2], Australia (1), Egypt (1), Estonia (1), Greece (1), Japan (1), Lebanon (1), Malta (1), Mexico (1), Panama (1), Poland (1), Singapore (1), and Sweden (1).

In round 2, we invited the 164 panellists who had completed round 1. Of these, 132 completed round 2 (response rate: 80% of the round 2 panellist set, or 67% of the round 1 panellist set). This compares favourably with other Delphi consensus surveys: in a review of 100 Delphi studies, Diamond et al. (2014) reported that only five studies had ≥ 100 panellists in the final round, while the rest of the studies either had fewer participants (n = 90 studies) or the number was not specified (n = 5 studies). Table 2 summarises the distribution of round 2 panellists across self-assigned categories (allowing identification with multiple categories). Note that in this round, 10 panellists did not select identical stakeholder labels to describe themselves as in round 1. Round 2 panellists came from 29 countries: the UK (43), the US (15), the Netherlands (9), Canada (7), Finland (6), Germany (6), Spain (6), France (5), India (5), Cyprus (3), Israel (3), South Africa (3), Ireland (2), Norway (2), Reunion Island (2), United Arab Emirates (2), Australia (1), Egypt (1), Estonia (1), Italy (1), Japan (1), Lebanon (1), Lithuania (1), Malta (1), Mexico (1), Panama (1), Poland (1), Singapore (1), and Sweden (1).

[2]Although Reunion Island is an overseas department and region of the French republic, we counted it separately due to the geographical distance and potentially diverse experiences of the stakeholders in comparison to continental France.

2.2 Survey design and procedure

A Delphi consensus survey requires three types of contributors: panellists, moderators and an independent administrator. The panellists included the workshop participants who had informed the initial generation of statements, as well as those recruited through expression of interest forms. The Q-BEx team members (i.e., the main authors of this paper) acted as panellists but also as moderators: they organised and participated in the initial workshop, designed and administered the online survey, and analysed the data. To guarantee anonymity of participation, an independent administrator handled the correspondence with panellists at round 2 (as this included individualised reports). Anonymisation of the data ensured the moderators could not attribute any rating or comment to a particular panellist.

We outline the survey design and procedure below. The full protocol we followed is schematised in Figure 1. An important aim of the design was bias limitation. This was implemented by inviting a group of experts to inform the generation of the initial statements, and by adopting pre-defined procedures for moderation and analysis (as will be explained in the analytic strategy section below). The choice of what topics to include in the Delphi survey was informed by an in-depth review of existing questionnaires (Kašćelan et al., 2021) and by current research and practice (via the workshop).

Workshop

The three-day workshop was organised in Leeds in January 2020. The first two days of the workshop were attended by researchers, and the third one by practitioners. One practitioner and five researchers attended all three days. To inform the generation of statements for the online survey, the workshop was organised around thematic presentations by leading experts, a review of the state-of-the-art in bilingual experience questionnaires, issues raised by the participants, and guidance on the principles of questionnaire creation and validation by an expert in psychometrics (Kate Harvey). Throughout the workshop, participants were

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

Table 2. Round 2 panellist breakdown per stakeholder group.

Categories	Sub-categories	Distribution by sub-category, Number (%)	Distribution by category, Number (%)
Researchers	Researcher	57 (43%)	57 (43%)
Practitioners	Speech and language therapist/ pathologist/ logopedist	26 (20%)	40 (30%)
	Teacher	13 (10%)	
	Speech and language therapist/ pathologist/ logopedist, Teacher	1 (1%)	
Researchers/ Practitioners	Researcher, Speech and language therapist/ pathologist/ logopedist	24 (18%)	35 (27%)
	Researcher, Teacher	9 (7%)	
	Researcher, Speech and language therapist/ pathologist/ logopedist, Teacher	2 (2%)	
	TOTAL		132

invited to contribute their views verbally, in writing (using post-its and interactive virtual whiteboards), and through live polling (using Mentimeter). The thematic presentations focused on the themes in (1), and led to group discussions.

(1) Thematic presentations
   a. Capturing linguistic diversity (Ianthi Tsimpli)
   b. Using experience data to inform the assessment of risk of atypical development (Sharon Armon-Lotem)
   c. Measuring family socioeconomic status in studies of bilingual development (Erika Hoff)[3]
   d. Language mixing (Elma Blom)
   e. Input quality in relation to input quantity (Johanne Paradis)

Emerging themes were identified and discussed, leading to the generation of statements (individually and in small groups). Participants were then invited to contribute what they considered uncontroversial as well as controversial statements.

With the practitioners, the discussions were informed by a critical review of current practice and reflections on practitioners' needs. This also led to the generation of statements. At the end of the workshop, a combined list of 197 statements had been compiled.[4]

Following the workshop, the moderators excluded duplicates and unclear statements. Similar statements were merged, and some were reformulated for clarity, based on our notes from the workshop. This resulted in 53 statements, some of which consisted of several parts (see example (2)). To assess agreement with each part more precisely, every part of an overarching statement was assessed as a separate statement. Therefore, in what follows, we will refer to each part (i.e., sub-component of the original 53 overarching statements) as a separate statement. In this way, altogether, 112 statements[5] were included in round 1 of the Delphi survey.

Pilot

The survey was piloted by a group including the six moderators (i.e., the main authors of this paper), as well as six additional researchers and one SLT (the latter seven did not participate in the online study). The aim of the pilot was to check for any errors, assess the clarity of the statements, and optimise their order of presentation.

Delphi survey round 1

The first round of the online survey was administered in April and May 2020. Panellists were given five weeks to complete it (including a one-week extension). In addition to a personalised link to the survey, each panellist was emailed a Briefing Document (see supplement 1) explaining the aim of the study, and a Glossary of the technical terms appearing in the survey (see supplement 2). The panellists were asked to score the statements on a 5-point scale (1 = strongly disagree, 2 = disagree, 3 = I don't know, 4 = agree, 5 = strongly agree). Two statements were different: they asked panellists to indicate the optimal amount of time needed to complete a short and a long version of a bilingual experience questionnaire. Options for a short version included 5 minutes, 10 minutes, and 15 minutes, while options for the long version included 20 minutes, 30 minutes, 40 minutes, 50 minutes, and 60 minutes. There was a space for optional open-ended comments following each statement.

Delphi survey round 2

This round was administered in June and July 2020 and the panellists were given six weeks to complete it (including a one-week extension). Apart from a link to the survey, each panellist was emailed a personalised report of round 1, containing the distribution of responses for each statement, as well as their own scores for each statement (see supplement 3). In addition, the panellists were emailed a list of round 1 comments (see supplement 4). Finally, the panellists were sent Clarifications and Instructions for completing round 2 of the survey (see supplement 5).

In round 2, the panellists were asked to re-rate 27 of the round 1 statements, and to rate 10 reformulated statements (each one presented immediately following their original formulation), and four new statements. As explained below, we only reformulated statements at round 2 if there was reason to suspect that the lack of consensus at round 1 was due to lack of clarity (determined by participant comments).

[3]Erika Hoff was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant HD068421.

[4]We generally avoided negative statements, to prevent the complication of double negatives, which are hard to interpret ("I strongly disagree that this is not the case."). The only exceptions were statements s.5, s.13, s.35, s.45 and s.47.

[5]A complete list of round 1 and round 2 statements, panellists' response distribution and comments are available via the Open Science Framework: https://osf.io/2pd65/

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

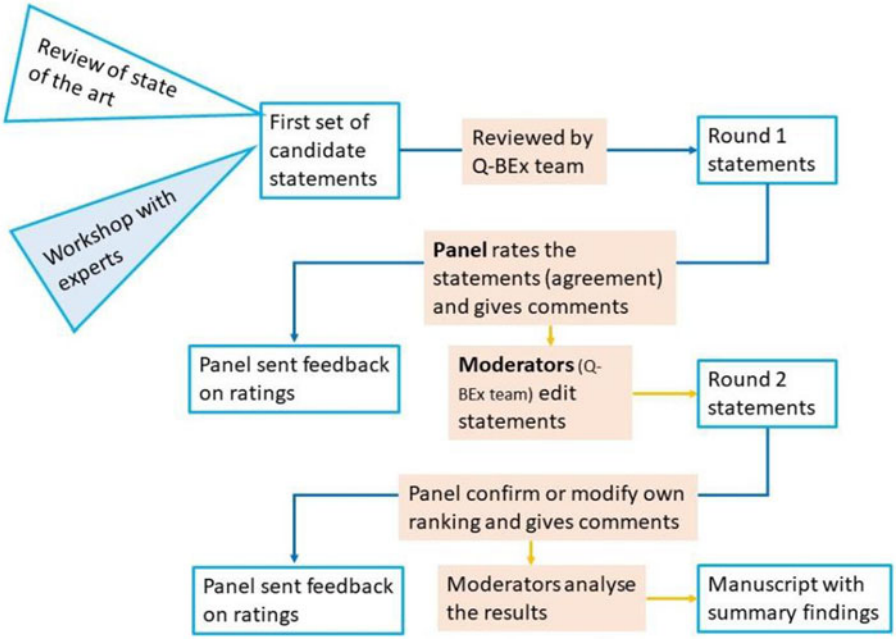


Figure 1. Procedure for the Delphi consensus survey, outlining the role of panellists and moderators[6]

3. Results

3.1 Analytic strategy and between-rounds moderation process

In their guideline paper on the Delphi methodology, Hasson et al. (2000) reported that values between 51% and 80% of agreement have been used as cut-off points for consensus in the literature. In order to determine the consensus cut-off point for our survey, we followed a review by Diamond et al. (2014). This review was based on 100 Delphi studies randomly selected from various disciplines that had been published between 2000 and 2009. Among those using a percentage or a proportion to define consensus, the median consensus threshold was 75%. Consequently, we applied the same criterion in our survey.

(Dis)agreement was defined as follows. Whenever a panellist selected '4 = agree' or '5 = strongly agree', we marked this as agreement with the statement. The round 1 statements reaching an agreement rate ≥75% were considered as having reached consensus and were excluded from round 2 (n = 74 statements), that is, they were immediately included in the final set of agreement-reaching statements (which we analyse in section 3.6). In addition, as an adaptation from Langlands et al. (2008) and Spain and Happé (2019), the round 1 items with an agreement score between 60% and 74%[7] were re-rated in round 2. The purpose of this approach was to reconsider what we refer to as the 'proximal zone statements' (i.e., those statements which were not that far from the designated consensus threshold) rather than dismiss them straight away. In round 2, the panellists were invited to reconsider their own scores for these statements, in light of the distribution of round 1 average scores and relevant comments: they could either confirm or modify their score based on this information. Twenty statements met the above 'proximal zone' criterion for inclusion in round 2.

A subset of panellists identified themselves as both a researcher and a practitioner (n = 38). We refer to these panellists as the 'dual interest' group. Since these panellists had insights from both perspectives, we decided to give stronger weight to their round 1 ratings in the following way: statements reaching proximal zone agreement levels (60%-74%) in the dual interest group were automatically included in round 2, even if they were below the 60% threshold in the panel as a whole. Eight statements were thus identified for inclusion to the proximal zone pool (in addition to the 20 mentioned above).[8] Statements reaching an overall agreement rate lower than 60% in round 1 were excluded from round 2 (n = 8).

Figure 2 presents the distribution of the statements by agreement rating in round 1. It shows that the pre-defined thresholds for consensus (75% agreement) or proximal zone (60% to 74% agreement) do not correspond to discrete breaks in the agreement distribution. The distribution is skewed (towards high agreement levels) but continuous.

Finally, we conducted a thematic analysis of the comments gathered in round 1. This qualitative analysis technique aims to identify the themes that emerge in a dataset inductively. In this approach, a theme is to be understood as a concept central to a mind map of related topics (i.e., a theme is underpinned by a core organising concept, related to other concepts – Braun & Clarke, 2019; Clarke & Braun, 2017). A theme can therefore capture a diversity of inter-connected meanings. The importance of a theme is determined by this web of relationships rather than just frequency of occurrence in the data. The flexibility of this technique allowed us not only to focus on the semantic content of the data, but also to consider the latent levels (i.e., going beyond the semantic content). Consequently, the need for new statements and reformulations could be identified even if they were not explicitly requested. For a further discussion of this approach, see Braun, Clarke and Hayfield (2019).

In round 1, there were 2,486 comments. One of the moderators selected a random subset of 11 statements (1.1, 1.2, 10.1, 10.2, 11, 12, 13, 14, 15, 16.1, 16.2) which in total contained 430 comments

[6]To illustrate our procedure, we adapted a flowchart Figure 1 from Bishop et al.'s (2016) CATALISE Delphi consensus study.

[7]The upper limit was 79% in Langlands et al. (2008) and in Spain and Happé (2019).

[8]One of these eight statements (statement 39.2) was excluded from round 2 as its related and preceding statement (i.e., statement 39.1) reached a very low overall agreement (32%). Consequently, it did not make sense to include statement 39.2 in round 2 for reconsideration.

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

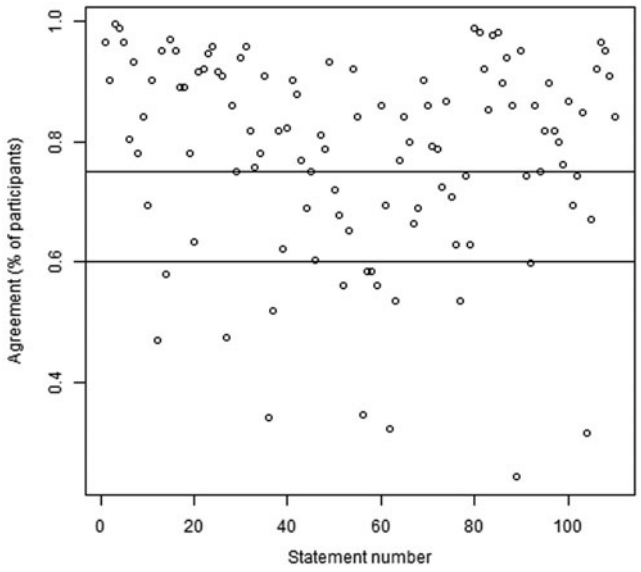


Figure 2. Distribution of round 1 statements by agreement rating (defined as the proportion of participants choosing "agree" or "strongly agree"). The horizontal lines indicate thresholds (for consensus: upper line, and for proximal zone: lower line).

(approx. 17% of the data). Based on the reading of these comments, 19 themes were identified and defined inductively. Two other moderators checked the thematic classification of 126 of these comments for consistency, and the list of themes was refined accordingly. The final list of themes and their definitions (provided in supplement 6) were used by all six moderators to classify the remainder of the comments. Each comment could be assigned up to three themes. The comments were then filtered by theme and analysed to identify opinions not yet represented by the survey statements. This resulted in 10 reformulations of the round 1 statements, as well as 4 new statements. Note that all 10 reformulations were rewordings of round 1 proximal zone statements. Consequently, in round 2, the panellists had to re-rate the original round 1 proximal zone statement as well as its reformulation.

In total, round 2 included 27 statements to be re-rated from round 1 (i.e., proximal zone statements based on the overall or the dual interest group ratings), 10 reformulations of the round 1 statements, and four new statements. Round 2 statements reaching an agreement rate ≥75% were considered to have reached a consensus.

3.2 Post-round 2 descriptive analyses

Overall, there were 55 overarching statements (53 in Round 1 and an additional two in Round 2). Many of these consisted of two or more parts (or sub-components), as in (2).

(2) The language(s) used at school should be documented as:
  a. language(s) used by teachers;
  b. language(s) used by the child;
  c. language(s) used by playmates.

For the purposes of our analysis, each part (i.e., sub-component) was counted separately, leading to a total of 126 statements (across the two rounds).[9]

After both rounds, consensus was reached for 79% of statements (98/124). As seen in Figure 3, the distribution of statements varied across agreement bands (with agreement defined as the proportion of panellists expressing agreement or strong agreement for a particular statement). The three highest agreement bands included approximately 85% of statements. By contrast, the 60-70% agreement band included only 6% of statements, and 8% of statements received agreement below 60%. None received less than 20% agreement.

Round 2 included 27 statements in their original formulation, 10 reformulations, and 4 new statements. Most of the original statements (93%, i.e., 25/27) were rated higher in round 2 than in round 1. However, this was not always sufficient for consensus to be reached (requiring a rating as "agree" or "strongly agree" by at least 75% of round 2 panellists). Only 52% of the original statements from round 1 (i.e., 14/27) reached consensus in round 2. From the reformulated statements, 7/10 yielded a higher agreement rating in round 2 than the round 1 originals, while 6/10 reached consensus. Two of these six reformulated statements (statements 31.1 and 49.1) reached consensus both in their round 1 formulation (proximal zone statements) and in their round 2 reformulation. All the new round 2 statements (n = 4) reached consensus, bringing the overall round 2 consensus to 59% (24/41).

Figure 4 shows the difference in agreement ratings of proximal zone statements between the rounds. The statements are presented in the ascending order of the difference in ratings. From the 14 proximal zone statements that reached consensus following round 2, the following three had a jump higher than 15% between the rounds:

- 42. There should be a question on attitudes to language mixing (b) within the local community (including school).
- 5. The questionnaire should not aim to measure the child's language proficiency. This should be done by other means.
- 36.2 [The literacy practices of parents] need to be documented independently of parental education and socioeconomic status.

A visual summary of the entire study can be seen in Figure 5.

3.3 Attrition analysis

Out of the 164 panellists in round 1, 32 did not respond to round 2. To assess the risk of bias at round 2 as a result of panellist attrition, we reanalysed the data from round 1 by including only responses from panellists who had taken part in both rounds. The results from this subset of panellists differed from the results from the whole round 1 panel in three respects.

First, out of the 74 statements that reached consensus at round 1 in our original analysis, four of them (22, 25, 27e, and 51.1.a) remained marginally below the consensus threshold in the subset analysis (reaching 71.21%, 73.48%, 74.24%, and 72.72% agreement respectively) when the panellists who responded in round 1 only were excluded. While there is no way of knowing whether re-rating these 4 statements in round 2 would pass the 75% consensus threshold, these high levels of agreement (71%-74%) suggest that it is likely that they would. Indeed, as shown in Figure 4,

[9]Of these, 124 statements were rated on a 5-point agreement scale (strongly disagree, disagree, I don't know, agree, strongly agree). The remaining 2 statements were rated on a time-length scale, as they inquired about preferred time lengths of the short and the long versions of the questionnaire. These 2 statements are excluded from the analyses below. For the distribution of responses on these 2 statements, see supplement 7.

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

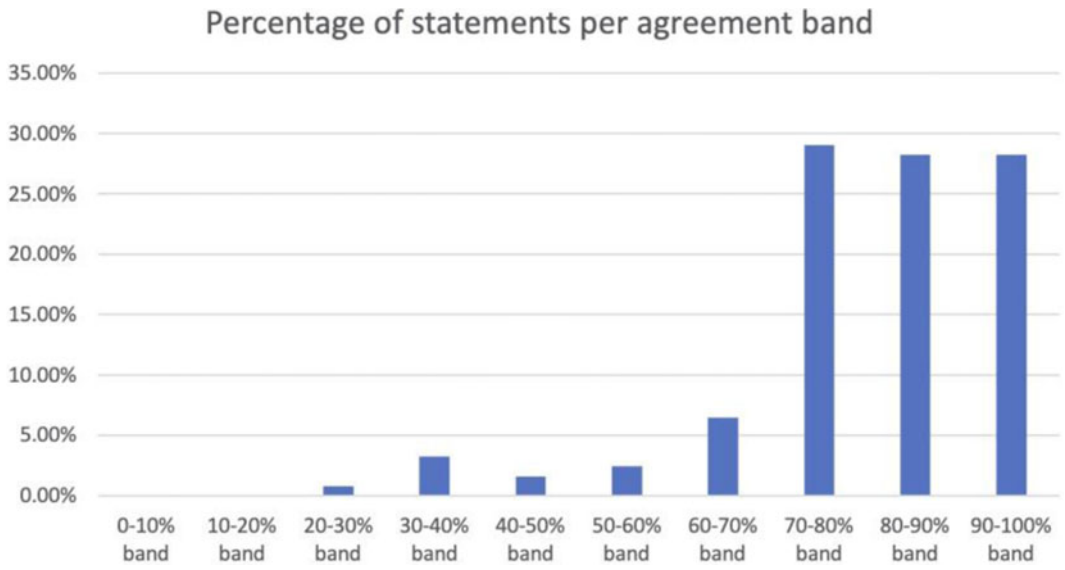


Figure 3. Percentage of statements per agreement band following the two rounds of the Delphi survey

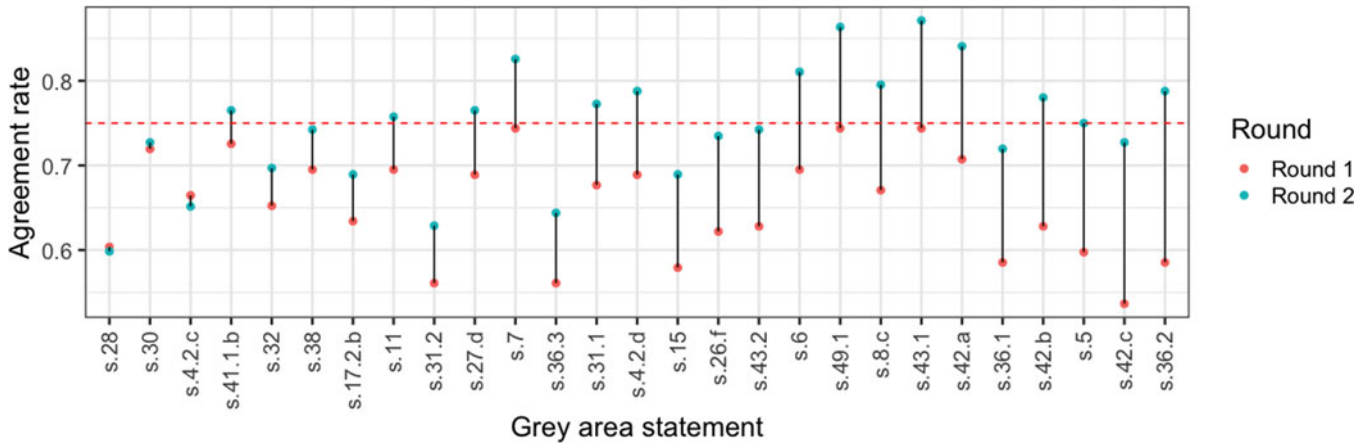


Figure 4. Difference in agreement ratings of proximal zone statements between the rounds (ordered ascendingly by difference size). The dotted horizontal line indicates the consensus threshold.

most statements that were re-rated in round 2, increased in the agreement rate by 5% or more.

Second, out of the 28 proximal zone statements identified above, 2 statements (49.1 and 7) already reached consensus at round 1 in the subset analysis (reaching 77.27% and 75% agreement respectively). This is of no consequence, as the re-rating of these statements at round 2 allowed them to reach 86.36% and 82.57% agreement respectively. Finally, statement 31.2 reached the proximal zone in the original analysis, but not in the subset analysis. Again, this is of no consequence as the re-rating of this statement at round 2 did not reach consensus, even in its reformulation (62.87% and 68.18% agreement respectively). Therefore, we conclude that attrition did not increase the risk of bias at round 2.

3.4 Subgroup analysis

To assess the extent to which the views of researchers aligned with those of the panel as a whole, we analysed the level of consensus among researchers at round 1 (including those from the dual interest group). This revealed that the exact same set of statements reached consensus at round 1 even if only researchers were included. For 92% of statements, there was less than a 5% difference in agreement level between researchers and the panel as a whole. The average difference was 3.4%, and the maximum 10.6%. The direction of the difference was almost evenly balanced (with 53% of statements achieving a higher level of agreement among researchers compared to the panel as a whole, and 47% thus showing a lower level of agreement among the researchers). Differences greater than 5% were always towards greater agreement among researchers than the group as a whole.

3.5 Polarisation analysis

To ascertain whether there was any polarisation of opinion in the statements that reached consensus, we inspected the distribution of responses for those statements. The data can be seen in Figure S3 (supplement 8), in which statements are ordered according to the level of disagreement (with the would-be controversial statements appearing in the top-most part of the plot). None of the consensus-reaching statements was rated as "strongly disagree" by more than 6% of panellists. We conclude that there

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

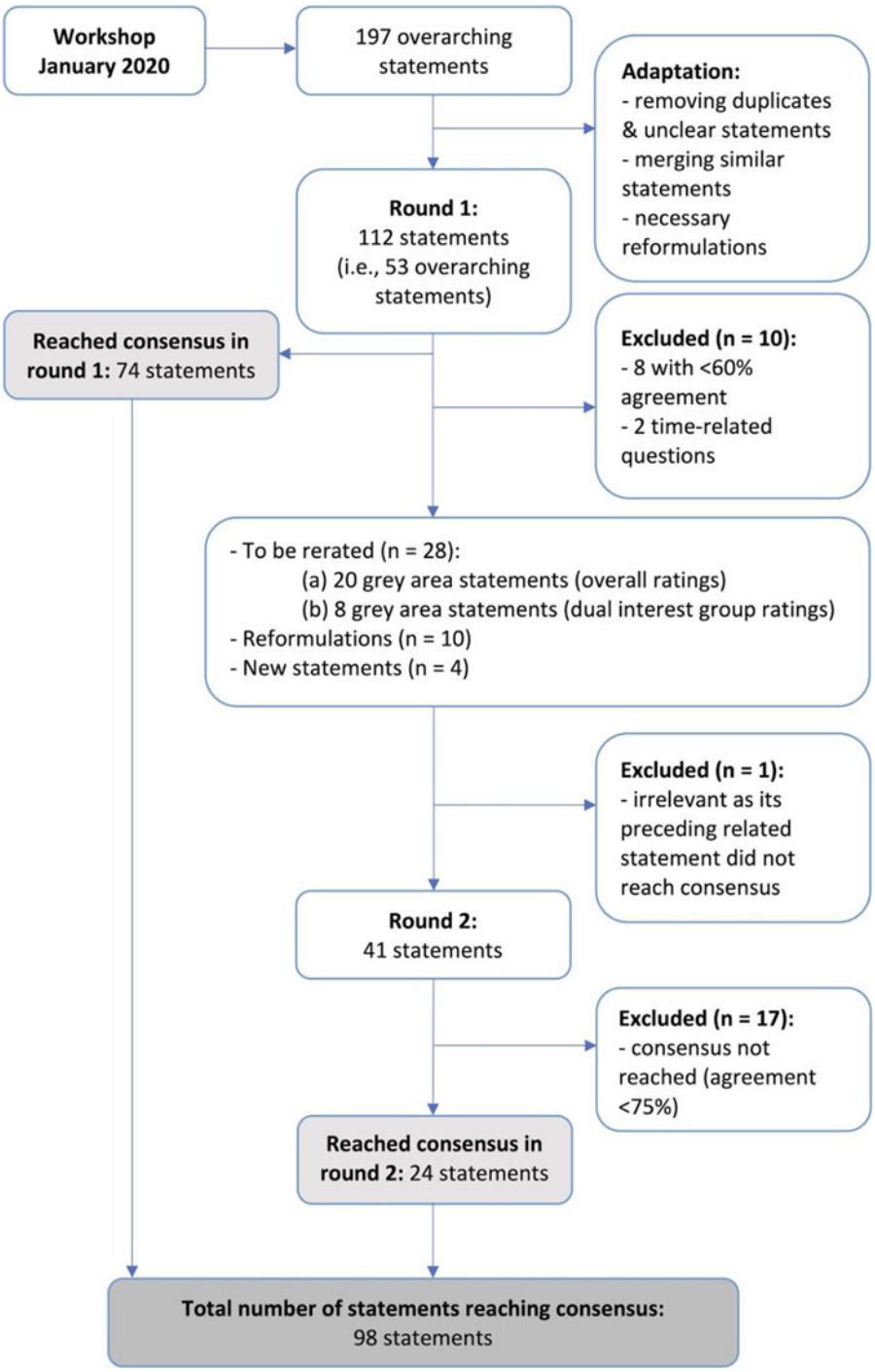


Figure 5. Consensus by stage of the Delphi survey

was no substantial polarisation with respect to the consensus-reaching statements.

3.6 What is the consensus?

In this section, we present the 98 statements that reached consensus following both rounds of the Delphi survey (i.e., with an agreement rate ≥75%). We list them in sections grouped around specific topics, along with the level of agreement reached (expressed as the proportion of panellists who agreed or strongly agreed). We also indicate if the statement was the original round 1 statement, a reformulation, or a new statement added in round 2. For ease of reference, the statement numbers below are as per those in the online survey. A short commentary is provided at the end of some sections as a reflection on the consensus. The complete datasets from both rounds of the online survey can be found via the Open Science Framework: https://osf.io/2pd65/

Mandate for a new tool

1.1 There needs to be 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%, original]

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

1.2 These measures should be applicable to children who speak more than two languages. [90%, original]
2. The questionnaire should be accompanied by a tool yielding automatic calculation of objective scores (in each language) of:

(a) current language exposure [95%, original]
(b) current language use [96%, original]
(c) cumulative language exposure [91%, original]
(d) cumulative language use [91%, original]

3. The tool should provide clear guidance about how to interpret the data it produces (where possible). [93%, original]
53. Questions should be as concrete as possible (e.g., asking about daily routines rather than asking the respondent to estimate frequencies in percentages). [87%, original]

Language difficulties

4.1 The questionnaire should contain a section to identify children who might be at risk for a Developmental Language Disorder. [77%, original]
4.2 This should include:

(a) early language milestones [84%, original]
(b) hearing difficulties [80%, original]
(d) issues related to trauma, attachment, prematurity [79%, original]
(e) family history of learning difficulties or speech/language delays [92%, new]

7. The questionnaire should ask about difficulties the child may have (had) with language, in order to identify what might require further assessment by specialists. [83%, original]
55. If the questionnaire includes a section documenting language and/or developmental difficulties, this should under no circumstances be used as a diagnostic tool. [87%, new]

Child's proficiency

5. The questionnaire should not aim to measure the child's language proficiency. This should be done by other means. [75%, original]
6. The child's proficiency should be documented for the language (s) that cannot be tested directly. [81%, original]

The consensus was that questionnaires should document a child's proficiency only when the languages in question cannot be tested directly. Further support for this interpretation comes from the lack of consensus on the reformulation of statement 5 (The questionnaire should not aim to document the child's language proficiency [in any of the languages]. This should be done by other means) and of statement 6 (The child's proficiency should be documented for all their languages).

Exposure and use

8. Exposure and use should be measured (for each language):

(a) over an average week [85%, original]
(c) over holiday and school periods separately [80%, original]
(d) over home and school separately [92%, original]

9.1 The language(s) used at school should be documented as:

(a) language(s) used by teachers [96%, original]
(b) language(s) used by the child [95%, original]
(c) language(s) used by playmates [91%, original]

9.2 Frequencies of use should be documented separately for each type of interlocutor. [84%, original]
10.1 The languages used outside school should be documented as:

(a) languages used with parents [99%, original]
(b) languages used with siblings [99%, original]
(c) languages used with other carers in the home [96%, original]
(d) languages used with friends of the family [80%, original]
(e) languages used amongst the child's friendship groups [93%, original]
(f) languages used within the neighbourhood [78%, original]

10.2 Frequencies of use should be documented separately for each type of interlocutor. [84%, original]
11. The amount of overheard speech (between parents) needs to be estimated. [76%, original]
12. The child's digital language exposure and use needs to be measured (e.g., Internet, social media, gaming). [90%, original]
14. Changes in the child's language exposure over time should be documented. [95%, original]
16.1 The child's first exposure to each language should be documented, in terms of:

(a) age [97%, original]
(b) context (e.g., childminder, pre-school, etc.) [95%, original]

16.2 A list of potential contexts should be provided. [89%, original]
17.1 Time spent in school should be quantified. [89%, original]
17.2 This should be done in:

(a) hours per week [78%, original]

In addition to these statements, it is important to highlight statement 13, which did not reach consensus: the precise measuring of what happens during holidays is unnecessary. It is enough to document whether or not the child travels to the "home" country or has ties with "manifestations of the home country" (e.g., regular contacts online in the home language with dispersed family members). As the statement had a low agreement rate (47%), the documentation of exposure- and use-related practices during holidays might need to be considered.

Child's education and literacy

18. Any prolonged period in which the child did not attend formal education should be documented. [91%, original]
19. The questionnaire should ask if the child attended school in another country. [92%, original]
21. The child's frequency of reading in each language should be measured. [86%, original]
22. Child's frequency of writing in each language should be measured. [75%, original]
23. The questionnaire should document any home-language classes that children are attending:

(a) in school [94%, original]
(b) outside school [96%, original]

36.1 Any literacy activities that the parents engage in with the child should be documented. [86%, reformulation]
36.2 This needs to be done independently of parental education and socioeconomic status. [79%, original]

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

Input quality

24. Input quality should be measured as far as this is feasible. [82%, original]
25. There needs to be agreement on a global/composite measure of input quality. [76%, original]
26. The following aspects are indicative of input quality:

(a) parental education [78%, original]
(b) interlocutor proficiency in each language [91%, original]
(e) pre-literacy and literacy activities [82%, original]
(g) playing with peers [82%, original]
(h) number of interlocutors who interact with the child in each language [86%, new]

27. The language proficiency of the child's interlocutors should be documented (based on the respondent's estimate). This should include estimates for:

(a) each parent [90%, original]
(b) any siblings [88%, original]
(c) other members of the households [77%, original]
(d) playmates [77%, original]
(e) teachers [75%, original]

28. The language proficiency of the child's interlocutors should be estimated in relation to typical and representative members of the population/region in which the child lives. [76%, reformulation]
29.1 The types of activity carried out in each language should be documented (e.g., storytelling, video games, play, etc.). [81%, original]
29.2 A predefined list of activities should be provided to ensure comparability. [79%, original]
37. Both mother's and father's education need to be documented. [86%, original]
38. If a parent was educated in more than one language, education in each language should be documented separately. [76%, reformulation]

The statements above demonstrate a general requirement across the sectors for estimating child's input quality. While statement 26 outlines the aspects which according to the panellists are indicative of input quality, some of the suggestions did not reach consensus. These were: interlocutor accent in each language (34%), language mixing (52%), and digital exposure (73%).

Language mixing

31.1 Language mixing should be estimated (in terms of exposure and use). [77%, original]
31.1 Language mixing (heard or produced by the child) should be estimated as part of the child's language exposure and language use. [78%, reformulation]
32. Language mixing that the child is exposed to should be documented separately from the language mixing that the child produces. [77%, reformulation]
33. The questionnaire should ask if the parents use one language in conversation and the child responds in the other. [92%, original]

It is worth noting that in contrast to the statements on language mixing above, where consensus was reached, only about a third of stakeholders agreed with the related statement on mixing in highly multilingual contexts (i.e., in densely multilingual societies, language mixing need not be measured (because it is so prevalent)).

Attitudes

34. There should be a question probing whether the child is unwilling to speak one of his/her languages. [84%, original]
40. There should be a question on attitudes towards each of the child's languages

(a) within the family (at home) [90%, original]
(b) within the local community (including school) [86%, original]
(c) within the broader society [79%, original]

41.1 There should be a question about the status of each of the child's languages

(a) within the local community (including school) [79%, original]
(b) in the "home" country (if applicable) [77%, original]

42. There should be a question on attitudes to language mixing

(a) within the family (at home) [84%, original]
(b) within the local community (including school) [78%, original]

43.1 Parents should be asked if they feel pressurised to speak the societal/majority language. [87%, original]

Background information

41.2 The child's languages should be identified precisely (e.g., variety, dialect). [87%, original]
44. The following demographic information should be collected:

(a) child's date of birth [99%, original]
(b) the date of filling in the questionnaire [98%, original]
(c) child's sex [92%, original]
(d) child's birth order [85%, original]
(e) child's arrival into the country (if relevant) [98%, original]
(f) period of the child's life spent in other countries (if relevant) [98%, original]
(g) name of each country in which the child has lived (if relevant) [90%, original]

45. The questionnaire should not label languages as societal, heritage, minority or majority languages. The language names (provided by the respondents) should be used throughout the questionnaire to identify these languages where relevant. Labels can be applied post-hoc by the researcher or practitioner as required. [94%, original]
46. A range of labels for the child's carers should be allowed by the questionnaire, to better document different family constellations (e.g., other than mother + father). [86%, original]

In addition to the above, only 24% of the panellists agreed with statement 47 (The questionnaire should not document the immigration history of the child). Such an outcome could be expected as immigration history often complements or is a part of the data on the child's language exposure and use.

Questionnaire versions

48. The questionnaire should be available in an online version, in a paper version, and as an interview protocol. [95%, original]

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

49.1 The questionnaire should be available in different lengths. [86%, original]
51.1 The questionnaire should be available in different versions for different respondents. This should include:

(a) a version to be used with 5-7-year-old children [75%, original]
(b) a version to be used with 8-12-year-old children [82%, original]
(c) a version to be used with parents/carers (also usable by adolescents) [90%, original]
(d) a version to be used with teachers [82%, original]

51.2 The child-focused versions will need to be complemented by a brief parental questionnaire. [80%, original]
52. The questionnaire in which the child is the respondent should be administered as an interview with the child. [76%, original]

Questionnaire modularity

49.1 The questionnaire should contain thematic sections (e.g., on language exposure/use, on proficiency, on attitudes, etc.). Each section should be optional, and it should be up to the researchers/practitioners to select which section to use. [87%, reformulation]
50. Some sections of the questionnaire should be optional. The researcher/practitioner should be able to exclude some sections. [86%, original]
54. Some parts of the questionnaire should not be administered if asking these specific questions is not adequate (for safety, political, personal, or any other ethical or relevance reasons). [88%, new]

4. Discussion

This Delphi consensus survey aimed to ascertain the level of agreement among researchers and practitioners regarding how bilingual experience should be documented. Our aim was to identify the broadest consensus possible, in order to inform the design of a new, customisable questionnaire allowing researchers and practitioners to select the components relevant to their purpose.

The round 1 statements were formulated based on the outcomes of an exploratory workshop with 22 researchers and 14 practitioners. In two rounds of the online survey, a diverse set of 132 panellists from 29 countries rated a total of 126 statements[10] (112 original statements, four new statements, and 10 statement reformulations). Furthermore, 27 original round 1 statements were re-rated in round 2. Several measures were adopted to limit the risk of bias: we used a diversification strategy for the recruitment of panellists; the data was anonymised upon collection; we used pre-defined criteria for consensus and for selecting statements to be re-rated; and we carried out post-hoc bias analyses. While it is impossible to avoid bias entirely, we believe these measures give credence to our approach. We are confident we have captured a range of opinions reflecting a variety of centres of interests among an international community of researchers and practitioners.

Overall, the level of consensus was high: 79% of statements (i.e., 98/124) reached consensus across rounds, and only 8% of statements (10/124) received less than a 60% agreement rate. This suggests that the views of the experts from the initial workshop were fairly representative of those of the wider, more diverse panel participating in the online survey, including a solid proportion of uncontroversial views but also some more controversial ones.

[10]Two of these were not rated on an agreement scale and were excluded from the consensus calculations below.

4.1 The consensus

There was an almost unanimous call for a set of common measures of bilingual experience, enabling greater comparability across studies, and facilitating exchanges and cross-pollination across sectors. Also, almost unanimous was the desire for an automated calculator of language exposure and language use, and the need to allow for measuring multilingual (not only bilingual) experience.

Consensus was also reached regarding the need to document a number of aspects of bilingual experience. This includes, for each language: language exposure and use, language mixing, language difficulties experienced, proficiency (if it cannot be assessed by other means), education and literacy, indices of input quality, language mixing practices, and attitudes (towards each language and towards language mixing). Consensus levels were the highest in relation to language exposure and use, and the need to document them in detail (i.e., across interlocutors, in different contexts, and over time).

The variability observed with respect to other aspects is likely a reflection of the fact that they have hitherto been researched less systematically. The survey results indicate that researchers across the board agree these aspects are likely to be important but require more supporting evidence and/or more scrutiny. For example, input quality and language mixing are starting to attract more attention in bilingualism research (see, e.g., Unsworth, 2016 and Byers-Heinlein, 2013, respectively), and this is reflected in a set of quality-related and language mixing statements which reached consensus. Nonetheless, several statements relating to the documentation of language mixing did not pass the consensus threshold. For instance, there was no agreement about the documentation of language mixing for each interlocutor in interactions with the child (statement 30); nor was consensus reached on documenting language mixing through examples of different types of mixing or their frequency (statement 31.2 and its reformulation). This might reflect scepticism as to whether questionnaires can reliably document these aspects of language mixing. Further research will be necessary to elucidate these points.

Another aspect eliciting diverse reactions was language proficiency. Round 2 reformulations revealed that this was due to a caveat: if language proficiency could not be assessed by other means, the agreement was that it should be documented via a questionnaire. Similarly, as long as the questionnaire is not seen as a diagnostic tool, the consensus was that the child's difficulties with language should be documented.

The need for a flexible or modular questionnaire elicited strong consensus. The constant tension between the level of detail aspired to and the constraints inherent to data collection will often result in having to forgo the documentation of some aspects. A modular questionnaire will allow for this. The exclusion of some aspects will also partly depend on foci of interest and the purpose of documentation (e.g., screening for language difficulties vs. informing a study on a particular aspect of bilingual experience). Different versions of the questionnaire were also considered necessary to adapt to different types of respondents (e.g., caregiver, child, teacher).

The apparent contradiction between the call for a flexible and modular tool and the acknowledgement of the necessity to

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

document a wide range of aspects of bilingual experience brings us to the question of what is 'core' vs. what can be considered optional in the documentation. While this question was not asked directly in the survey, we believe variations in level of consensus can be interpreted as useful indicators. Language exposure and use, as well as some indicators of language difficulties and some indicators of input quality, thus seem to emerge as core aspects of the quantification of bilingual experience. Ultimately, though, the identification of an essential 'core' is an empirical question, which will require comparing data from large and diverse groups of bilinguals and multilinguals, using identical measures. And this empirical investigation will need to ascertain not only which aspects of bilingual experience are part of the 'core', but also the minimum level of detail required to measure (or document) them reliably. Indeed, the cognitive load involved in completing bilingual experience questionnaires can be quite complex, as respondents are required not only to recall language practices over long periods of time and many different contexts, but also to estimate their frequency. In addition, they might not have been direct witnesses of these practices (e.g., in the case of a parent estimating what happens at school), or they might not have been fully aware of them (e.g., in the case of language mixing). This is likely to result in unavoidably high levels of error in the measurements. It is therefore not necessarily the case that a more detailed questionnaire will elicit more precise information. Here too, empirical investigation will be necessary to identify the optimal level of detail to be targeted by bilingual experience questionnaires.

4.2 Limitations

While the Delphi method provides a flexible approach to accommodate the needs of various fields, a clear limitation is the lack of agreed standards on what should count as consensus and how it should be interpreted (Iqbal & Pipon-Young, 2009). To limit the risk of bias, we adopted a pre-defined criterion informed by Diamond et al.'s (2014) review, which identified 75% agreement among panellists as the median consensus threshold among the publications in which consensus was defined as a percentage or proportion. Furthermore, we adopted a pre-defined criterion to identify round 1 statements in the proximal agreement zone (i.e., the proximal zone statements): informed by the approach of Langlands et al. (2008) and Spain and Happé (2019), any statement reaching agreement between 60% and 74% of round 1 panellists was automatically selected for re-rating at round 2. We also pre-defined a dual interest group (identifying as both researchers and practitioners), whose proximal zone statements (i.e., statements with the agreement rate 60%-74%) were automatically selected for round 2, in addition to those of the panel as a whole. These procedures enabled us to mitigate the risk of a potentially too conservative consensus threshold.

In spite of our diversification strategy, the panel remained predominantly Western-centric (both in the workshop and in the online survey). The use of English and the use of an online platform to conduct the study aimed to increase inclusivity, but at the same time it discriminated against non-English speaking stakeholders and individuals without access to the internet. Future work should therefore seek to improve representation, as many significant voices have likely not been included.

We also note that there are no validated quality indicators of Delphi studies. However, Diamond et al. (2014) proposed a set of four elements to include in Delphi publications to increase their value: (1) provide reproducible criteria for panellist selection, (2) state the number of rounds performed, (3) provide clear criteria for dropping the items, and (4) clarify whether there is a stopping criteria other than the number of rounds specified. We reported on each of them in this manuscript.

The ultimate limitation of the Delphi method is that, in spite of all attempts at bias limitation, it remains possible that the correct answer or opinion was not identified (Hasson et al., 2000; Iqbal & Pipon-Young, 2009). Opinion is also likely to evolve as bilingualism research progresses. However, we believe the findings of this survey could enable a step-change in bilingualism research through the adoption of a common method to document bilingual experience, thereby enabling greater comparability across studies (and across populations), and increased synergy between practitioners and researchers. The validity of the tool developed on the basis of the Delphi consensus will need to be assessed based on new empirical evidence. It may turn out that some of the aspects of bilingual experience which were deemed necessary to document (as per the consensus reached by the Delphi survey) can in fact not be reliably documented via questionnaires. If that is the case, it will indicate that opinions need to change. This paper represents the first step in this long process, that is, the documentation of current opinions.

There are also limitations inherent to the quantification of (aspects of) the bilingual experience, which was at the heart of this Delphi survey. Such an approach might not be sufficient to capture key aspects of bilingualism associated with variation across socio-cultural contexts, such as non-industrialised countries without a robust education system. Ethnographic approaches (not considered in this survey) might offer insightful alternatives in some contexts. It is possible that standardised questionnaires are inadequate to document bilingualism in some populations. The extent to which bi/multilingualism in these populations can be conceptualised along the same lines as in other populations is an important question beyond the scope of this paper and will require further research.

4.3. Next steps

The main authors of this paper are currently designing an online questionnaire[11] and background calculator meeting the requirements identified via the Delphi consensus survey and informed by a review of the state of the art in bilingualism questionnaires (Kašćelan et al., 2021), and by methodological insights from the literature on psychometrics (e.g., DeVellis, 2017). This questionnaire will allow some level of customisability, so that professionals can choose the components relevant to their research objectives, clinical practice, or educational aims. To limit the burden on respondents, we will also carry out a cost-benefit analysis aiming to identify the optimal balance between informativity and error margin. This is particularly important as lengthy questionnaires are likely to be too burdensome for some marginalised communities, which would result in less representative population samples. As is the case with the creation of any standardised tool, the domain of applicability of this new questionnaire will need to be assessed empirically. In the long term, a multi-team approach will be necessary for full validation and optimisation.

The creation of a new tool raises several methodological challenges. First, the constructs that make up bilingual experience

[11]The design phase took place while the manuscript for this paper was under review. The new online tool is now available for free at https://q-bex.org.

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press

need to be operationalised precisely and concretely to avoid reification. Second, there is a risk of normalisation inherent to the creation of a standard questionnaire. This risk must be limited by embedding into the questionnaire design the intention to capture the diversity of bilingual experiences. This will however need to be balanced with the need to keep the questionnaire sufficiently short and clear for respondents.

5. Concluding remarks

The need for standardisation in how bilingualism is characterised and categorised has become uncontroversial. This Delphi consensus survey has highlighted the readiness of bilingualism researchers and practitioners from both clinical and educational settings to adopt common methods for the documentation of bilingual experience, to enhance the generalisability of research findings and facilitate exchanges between research and practice.

Several new profiling measures have been proposed recently, as global indices of bilingualism: for example, the LSBQ composite score (Anderson, Mak, Keyvani Chahi & Bialystok, 2018; Anderson, Hawrylewicz & Bialystok, 2020), language entropy (Gullifer & Titone, 2020), or a possible bilingualism quotient (Marian & Hayakawa, 2020). Independent of their conceptual validity, the reliability of these measures will be determined by how their components are documented and quantified. This in turn will require comparability of measures, which will only be possible if similar documentation tools are used across research teams. The results of this Delphi consensus survey constitute a first step in that direction.

Acknowledgments. Warm thanks to the Q-BEx Consortium for contributing to the Delphi survey (twice!), and to the workshop participants for their insights and enthusiasm. We are grateful to Bissera Ivanova for helping us communicate anonymously with the panellists. We thank the reviewers for inviting us to go deeper into the discussion and to clarify the remit of our enterprise. This project is funded by the Economic and Social Research Council (grant reference: ES/S010998/1), which is gratefully acknowledged.

Supplementary Material. For supplementary material accompanying this paper, visit https://doi.org/10.1017/S1366728922000359

Data availability statement. The data that supports the findings of this study are openly available via the Open Science Framework at https://osf.io/2pd65/.

Competing interests declaration. The authors declare none.

References

Anderson JAE, Hawrylewicz K and Bialystok E (2020) Who is bilingual? Snapshots across the lifespan. *Bilingualism: Language and Cognition* **23**, 929–937. https://doi.org/10.1017/S1366728918000950

Anderson JAE, Mak L, Keyvani Chahi A and Bialystok E (2018) The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods* **50**, 250–263. https://doi.org/10.3758/s13428-017-0867-9

Bishop DVM, Snowling MJ, Thompson PA, Greenhalgh T CATALISE consortium (2016) CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. *PLoS ONE* **11**(7), 1–26.

Braun V and Clarke V (2019) Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* **11**(4), 589–597, DOI: 10.1080/2159676X.2019.1628806

Braun V, Clarke V and Hayfield N (2019) 'A starting point for your journey, not a map': Nikki Hayfield in conversation with Virginia Braun and Victoria Clarke about thematic analysis. *Qualitative Research in Psychology*, DOI: 10.1080/14780887.2019.1670765

Byers-Heinlein K (2013) Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition* **16**, 32–48. https://doi.org/10.1017/S1366728912000120

Clarke V and Braun V (2017) Thematic analysis. *The Journal of Positive Psychology* **12**(3), 297–298, DOI: 10.1080/17439760.2016.1262613

DeVellis RF (2017) *Scale Development Theory and Applications* (4th ed.). SAGE Publications, Inc.

Diamond IR, Grant RC, Friedman BM, Pencharz PB, Ling SC, Moore AM and Wales PW (2014) Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology* **67**, 401–409.

Gullifer JW and Titone D (2020) Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition* **23** (2), 283–294. https://doi.org/10.1017/S1366728919000026

Hasson F, Keeney S and McKenna H (2000) Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing* **32**(4), 1008–1015.

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

Kašćelan D, Prévost P, Serratrice L, Tuller L, Unsworth S and 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. https://doi.org/10.1017/S1366728921000390

Langlands RL, Jorm AF, Kelly CM and Kitchener BA (2008) *Journal of Affective Disorders* **105**, 157–165.

Luk G and Esposito A (2020) BLC mini series: Tools to document bilingual experiences. *Bilingualism: Language and Cognition* **23**(5), 927–928.

Marian V and Hayakawa S (2020) Measuring bilingualism: the quest for a "bilingualism quotient". *Applied Psycholinguistics*, 1–22. doi:10.1017/S0142716420000533

Spain D and Happé F (2019) How to Optimise Cognitive Behaviour Therapy (CBT) for People with Autism Spectrum Disorders (ASD): A Delphi Study. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*. https://doi.org/10.1007/s10942-019-00335-1

Thangaratinam S and Redman C (2005) The Delphi technique. *The Obstetrician & Gynaecologist* **7**(2), 120–125.

Unsworth S (2016) Quantity and Quality of Language Input in Bilingual Language Development. In Nicoladis E and Montanari S (Eds.), *Bilingualism Across the Lifespan: Factors Moderating Language Proficiency* (pp. 103–121). De Gruyter Mouton.

Walker AM and Selfe J (1996) The Delphi technique: a useful tool for the allied health researcher. *British Journal of Therapy and Rehabilitation* **3**, 677–680.

https://doi.org/10.1017/S1366728922000359 Published online by Cambridge University Press