4 Mapping Teacher Beliefs and Practices About Multilingualism: The Development of the MultiBAP Questionnaire

Pia Sundqvist, Henrik Gyllstad,
Marie Källkvist and Erica Sandlund

Introduction and Aims

While language-diverse English classrooms are under-researched in Sweden (Källkvist *et al*., 2017), teachers are gaining firsthand experience through teaching language-diverse student groups, thus gaining experience and knowledge that warrants documentation. Such knowledge is often conceptualized as beliefs and practices (Borg, 2006), and a suitable instrument for mapping that knowledge among large numbers of teachers quickly is the questionnaire (Dörnyei & Taguchi, 2010; Phakiti, 2015). Questionnaires, like any other instruments, must be capable of yielding reliable data through which valid inferences can be drawn, and scholars have recently called for increased methodological and statistical awareness in Applied Linguistics and Second Language Acquisition (SLA) (Norris *et al*., 2015; Plonsky, 2015), where the use of questionnaires is widespread (Dörnyei & Taguchi, 2010; Phakiti, 2015). In a similar vein, as pointed out by Borg (2015: 494; our emphasis), in many self-report instruments, there is room for quality enhancement, and 'a first requirement for researchers wanting to use questionnaires […] to study teachers' beliefs is to ensure they understand – theoretically and in practice – *how to design a robust instrument*'. Similarly, Gu (2016) and Valeo and Spada (2016) have called for more attention to how questionnaires are designed, analyzed and validated. Careful reporting of procedures and instruments used also make replication studies possible (Mackey, 2012; Marsden *et al*., 2018).

Following in the wake of increased mobility, language teachers are experiencing a shift towards greater linguistic diversity in additional language (L2) classrooms (Busse *et al*., 2020). At the same time, while there is extensive research on teacher beliefs about L2 teaching/learning in general (see, e.g. Borg, 2006, 2015; Pajares, 1992), there is little research on teacher beliefs specifically about the role of multilingualism in L2 classroom contexts (though see Lundberg, 2019). Prior research reveals that the classroom is 'a key site where policies become action' where teachers exercise agency (Hult, 2014: 159; see also Borg, 2006).

In response to the above calls, this chapter provides a detailed description of the methodology behind the development of a new questionnaire instrument called MultiBAP (Multilingualism: Teacher Beliefs And Practices). As part of the school-based research project MultiLingual Spaces (see Källkvist *et al*., 2017) – in which *multilinguals* are defined as learners of English who use Swedish and one or several additional languages (e.g. Arabic, Finnish or Somali) in their everyday life – MultiBAP was designed to map L2 English teachers' *beliefs* about multilingualism in individuals, in classrooms and schools and in Swedish society at large, and practices in their classrooms and schools. Pajares (1992: 316) suggests *teacher beliefs* be defined as 'an individual's judgment of the truth or falsity of a proposition' and are constructed in everyday practice (van Lier, 2006). Thus, it is relevant to map teacher practices while examining their beliefs, even though beliefs are not always mirrored in their practices (Borg, 2015).

Consequently, the present chapter aims to contribute to developing questionnaire research methodology in L2 language education. In pursuing this aim, we:

(1) describe the development of the instrument MultiBAP,
(2) critically evaluate each step of the development process and
(3) provide a step-wise validation of MultiBAP.

In what follows, we focus on methodological aspects of questionnaire development and then provide an account of the construction and validation of MultiBAP. We end by discussing possible uses of MultiBAP, including the need for further development and validation.

Questionnaires in Research on Beliefs and Practices – Methodological Considerations

Questionnaires have been used extensively in SLA research (e.g. Winke, 2011) and in research on teacher beliefs, although not as frequently. In Borg's (2015) account of 20 studies of L2 teachers' beliefs, half were qualitative, 7 were mixed-method and 3 were quantitative. Of these, 16 investigated the beliefs of in-service teachers. Nine had sample sizes of

fewer than 10 participants; three had 11–50 participants; four were composed of 51–100 participants and four > consisted of 100 participants.

Kern (1995), Levine (2003), De Angelis (2011) and Bailey and Marsden (2017) are examples of studies that focus on teacher beliefs, including topics such as comparisons between learner and teacher beliefs, beliefs about target and first language (L1) use and beliefs about anxiety and the role of prior knowledge in learning. Generally speaking, in such studies, validation procedures used are rarely addressed. Furthermore, with relevance to the current study, Norris *et al.* (2015: 472) stress the importance of providing evidence regarding both the consistency of the measurement instruments used and the validity of 'the intended construct interpretations being made in the actual study with the actual population sample'.

Studies discussing reliability and validity in more depth include Graus and Coppen (2016), Loewen *et al.* (2009), Lee and Oxelson (2006), Spada *et al.* (2009) and Winke (2011). Graus and Coppen (2016) investigated the beliefs of student teachers of English as a foreign language (N = 832) about grammar teaching. A questionnaire consisting of three parts was developed and validated, and following piloting and revisions, 24 five-point Likert-scale items were retained. Reliability values (Cronbach's alpha) between 0.735 and 0.864 were observed, and items had moderate to large loadings on their respective factors.

Loewen *et al.* (2009) studied learner beliefs about the role of grammar instruction and error correction. University students (N = 754) responded to a questionnaire containing 37 Likert-scale items (information about the range of the scale is missing) and 4 prompts (open-ended). The quantitative data were submitted to an Exploratory Factor Analysis (EFA) and '[f]actor loadings of .30 or greater on the obliquely rotated factor matrix were considered significant' (2009: 95), identifying six underlying factors, with a Cronbach'alpha of 0.84 for the questionnaire overall. No reliability values for the subscales are given.

Lee and Oxelson (2006) studied teachers (N = 67) responding to 35 questions about their students' heritage language maintenance, 11 about practices and 7 about demographics (plus 3 open-ended questions). A seven-point Likert scale was used. In total, 290 questionnaires were distributed. A rather low return rate (29%) was expected due to timing and an assumed lack of interest in the topic (heritage languages). The questionnaire had eight constructs of which reliability values were satisfactory for six (Cronbach's alpha ranged from 0.76 to 0.85), but low for two (0.51 and 0.53). The researchers used a Varimax principal component factor analysis and report eight underlying factors, highlighting the highest factor loading for each item. Items with a factor loading below 0.40 were excluded from the analysis. There is no further comment on the validity and reliability of the instrument.

Spada *et al.* (2009) centered on developing and validating an instrument for measuring L2 learner preferences for two types of form-focused instruction, 'isolated' or 'integrated', including 294 respondents. Three kinds of validity evidence were gathered: content, reliability and construct. Regarding content validity, 12 expert judges were asked to assess whether items should belong to the 'isolated' or 'integrated' scale. Only items for which there was 70% agreement or higher were kept. To calculate internal consistency reliability, Cronbach's alpha was used, and for construct validity, principal component analysis (PCA) was used. The authors initially created 44 items (5-graded Likert scale), but after several rounds of vetting, the instrument adopted included 20 for practicality reasons. Cronbach's alpha value for 10 items was 0.63 and for the other 10 items was 0.69. With regard to the PCA used for construct validity, 14 items with loadings of 0.30 or above were retained (two subscales, seven items per subscale). These explained 43.35% of the item variance, and the Eigenvalue for the 'integrated' component was 3.77 and for the 'isolated' was 2.30. Even though there were only seven items in each subscale, reliability values around 0.7 were claimed to be 'respectable […] for a new questionnaire with a small number of items' (Spada *et al.*, 2009: 78).

Winke (2011) included 267 respondents answering a questionnaire about the validity of the English Language Proficiency Assessment (ELPA) test. The questionnaire included three parts corresponding to the social, ethical and consequential dimensions of ELPA test. It had 40 belief statements, asking respondents to mark their answer on a 10-graded Likert scale. Based on the reported figures about the distribution of the questionnaire (Winke, 2011: 637), the response rate appears to have been 15.1% (an initial 2508 questionnaires, minus 585 that bounced back and 156 non-respondents). Internal consistency was 0.94 (Cronbach's alpha) overall based on 134 respondents (due to missing data) and 0.95 when missing values were replaced by the series mean. An EFA resulted in a five-factor solution, explaining 72% of the variance. The Eigenvalues ranged from 1.18 to 11.43. Regarding factor loadings, items with loadings of 0.5 or above were kept.

In sum, it seems that dominating reliability/validity analyses comprise the use of item analysis (item-total correlation, internal consistency reliability and internal vetting), expert judgments (content validity) and various types of factor analysis (underlying constructs). One observation relates to the type of EFA used. Specifically, the use of PCA over a common factor EFA model (e.g. Lee & Oxelson, 2006; Spada *et al.*, 2009) has been questioned (see the section *Item analysis and factor analysis* (p. 66) on the appropriateness of using PCA). Finally, details from piloting rounds are seldom reported, and the response rates are either not reported at all or vary in the way they are reported. In developing MultiBAP, we included item analysis, expert judgments, and an EFA.

Construction and Validation of the MultiBAP Questionnaire

On reviewing prior questionnaire research, it was clear that no extant instrument would capture the type of questions we intended to address. Therefore, we constructed and validated MultiBAP with the purpose of yielding generalizable, quantitative data. The target statistical population was secondary school (Grades 6–9) L2 English teachers in Sweden. A questionnaire cannot possibly cover everything in broad fields, but it may examine some aspects of the fields well, namely the targeted constructs (see below).

The process of creating MultiBAP breaks down into five carefully planned phases, outlined in Table 4.1, in line with important methodological considerations addressed by Wagner (2015). In Phase I, we decided on the parts to be included. In Phase II, we identified the constructs that the instrument was intended to tap into and generated a pool of items for each construct, which was then vetted in the research group. Finally, we asked two raters to link items to the constructs, which led to the final content of the PILOT Questionnaire. Phase III consisted of piloting MultiBAP using a sample of teachers from the same population as the one intended for the FINAL Questionnaire. Based on these data, we analyzed the feedback solicited from the respondents, carried out item analysis and created a Draft FINAL Questionnaire, which an external expert (specialized in multilingualism, L2 learning and translanguaging) was

Table 4.1 Phases and steps in the questionnaire construction and validation

Phase	Steps
Phase I	Deciding on questionnaire parts
Phase II	Theory-driven content specification (constructs)
	Item generation (multi-item scales)
	Internal vetting of items in the research group
	Decision on final content and design of PILOT Questionnaire
	Building of online version of the PILOT Questionnaire
Phase III	Administration of PILOT Questionnaire
	Analysis of teacher feedback on PILOT Questionnaire
	Validation: Item analysis
	Validation: Use of two raters – relating items to constructs
	Validation: Feedback from external expert
	Decision on content and design of FINAL Questionnaire
	Building of online version of FINAL Questionnaire
Phase IV	Administration of FINAL Questionnaire
	Item analysis and EFA of FINAL Questionnaire
Phase V	Content and design of MultiBAP Questionnaire

asked to critique. Following feedback, we decided on the content of the FINAL Questionnaire. In Phase IV, we administered the FINAL Questionnaire, followed by item analysis, an EFA and a reliability analysis. Finally, in Phase V, we decided on the design and content of the MultiBAP Questionnaire.

Phase I: Outlining the questionnaire instrument

Based on best practice for questionnaire design (Dörnyei & Taguchi, 2010; Wagner, 2015), MultiBAP was designed to capture data on beliefs, practices and background information, such as years of teaching experience. Beliefs are essential as they are known to underpin practices (Borg, 2006) of how to teach multilingual groups of students. Such contexts provide opportunities to use pedagogical translanguaging involving teachers' and students' background languages, defined as languages learned prior to classroom exposure to English (Bardel *et al.*, 2013). Demographic background data were deemed important to enable correlation analyses, for example, correlating teachers' experience with their self-reported beliefs and practices.

We used closed-ended items combined with a small number of open-ended items, thereby adopting so-called intramethod mixing (Johnson & Christensen, 2017). For closed-ended items, we used Likert scales with six steps, ranging from 'I fully disagree' to 'I fully agree'. Opinions vary as to whether scales should have an even or odd number of steps; we base our decision on the potential problem of having respondents overusing a middle category (Dörnyei & Taguchi, 2010). Leung (2011) found no clear negative effects of the use of even-numbered scales compared to odd-numbered scales, and by having a six-step scale, we forced respondents to place themselves either to the left or the right of the middle. In Part B, a seventh 'not relevant/don't know' option was included *but separated from the scale*, a procedure in line with Spratt (1999) and recommended by Broca (2015).

Regarding other design aspects, we considered the time needed by respondents to fill in the questionnaire. Dörnyei and Taguchi (2010) suggest that no questionnaire should take more than 30 minutes; knowing of teachers' heavy workload and valuing the need for as high a response rate as possible, our target was 20 minutes. Other considerations concerned starting from a theory-driven list of constructs/concepts/subjects/topics, creating a logical structure, using multi-item scales for constructs and using both positively and negatively worded items.

Phase II: Identifying the constructs and generating questionnaire items

Dörnyei and Taguchi (2010) recommend starting building a questionnaire by identifying critical concepts. This part of our work was guided both by

a research problem formulation in the parent study, MultiLingual Spaces, broadly relating to how teachers and students use their linguistic repertoires to facilitate the learning of English, and by research on multilingualism. We now turn to the six constructs that emerged as relevant.

The constructs

The first construct, *Openness towards other cultures*, has bearings on inclusiveness and attitudes towards other cultures other than one's own. Inclusive practices have been identified as fundamental to education (OECD, 2012) and entail using means to meet the range of natural variation among students in a classroom (Swedish National Agency for Education, 2013). In present-day Sweden, this variation in the range of background languages in the same classroom may include, for example, Arabic, Bosnian, Dari, Farsi, Persian, Polish, Serbian, Swedish and Vietnamese (Gunnarsson *et al*., 2015). Lindberg and Hyltenstam (2013) argue that a resource attitude to diversity and collectively striving for utilizing all students' varied experiences 'is a prerequisite for successfully teaching students with different linguistic and cultural backgrounds than the homogeneous Swedish one' (Lindberg & Hyltenstam, 2013: 126, our translation). Similarly, Edstrom (2006) argues that acknowledging students' L1(s) is teachers' moral obligation; students are then recognized as individuals and treated with respect. On this research background, we generated items aimed at tapping into teachers' attitudes to, *inter alia*, people from other cultures, having contact with them, visiting foreign countries, respecting people with views other than one's own and adapting to other people's habits and needs. This construct was targeted by 10 items in the pilot version (Appendix 1).

The second construct is *Multilingualism in general*, formed against the backdrop of multilingualism being the norm worldwide (Grosjean, 2008). Items were generated asking, for example, whether multilingualism is something positive, whether it is important to be multilingual in today's world and whether multilingual individuals are more likely to succeed in the future. Like the first construct, 10 items target this construct in the pilot version (Appendix 1).

The third construct centers on the current language situation in Sweden, which is characterized as rapidly growing in multilingualism due to refugee migration. Multilingualism researchers Lindberg and Hyltenstam (2013: 122, our translation) suggest that multilingualism be viewed as an asset, whereas in practice, they claim multilingualism involving migrant, minority languages to be commonly 'connected with problems and deficiencies' (our translation).

The fourth (4) and fifth (5) constructs tap into beliefs and practices to do with the use of background languages in learning an additional language. Specifically, whereas Construct 4 deals with learning any

additional language, Construct 5 targets English in particular. As to practices, research shows that bi- and multilingualism have a positive effect on the acquisition of additional languages (Cenoz & Genesee, 1998); there is strong evidence that bi-/multilingual users cannot completely deactivate their prior languages when processing information in a target language (see Källkvist *et al*., 2017). Further, the L1 has been shown to be an effective way 'of communicating meaning' (Nation, 2003: 5).

In terms of beliefs, teachers typically harbor positive beliefs about multilingualism. Research has shown that most teachers are hesitant towards allowing languages that are not known by them in the classroom (De Angelis, 2011; Heyder & Schädlich, 2014). For the beliefs part of MultiBAP, we generated items targeting whether drawing on background languages is good or bad, whether just in general or specifically in the classroom and whether additive multilingualism exists and whether specific language skills (speaking/reading/listening/writing/vocabulary/grammar) may benefit from involvement of background languages. Eleven and 19 items were created for Constructs 4 and 5, respectively, for the pilot version (Appendix 1).

Finally, the sixth construct has to do with monolingual beliefs. Here, it was possible to draw on an existing questionnaire (Pulinx *et al*., 2015), which focuses on Flanders, Belgium, a region where educational policies are predominantly based on a monolingual ideology. We saw an opportunity of replicating part of Pulinx *et al*. by gathering data from Sweden, where there has been some policy support for multilingualism in that mother-tongue tuition has been offered since 1977. We saw this also as a way of anchoring MultiBAP in an already existing questionnaire.

From constructs to item generation

Our initial goal was for items in Part A (beliefs) to mirror items in Part B (practices). However, it soon became clear that this would only be meaningful for Constructs 1, 2, 5 and 6. Thus, Constructs 3 (the language situation in Sweden) and 4 (using background languages to facilitate learning of an additional language) are included in Part A only.

Next, items were generated aiming to come up with multi-item scales for each construct, that is, 'a cluster of differently worded items that focus on the same target' (Dörnyei & Taguchi, 2010: 24) with no less than 3–4 items be used for each construct. We thus developed 7–10 items for each construct (pilot version) allowing us to, at a later stage, select 3–4 items (final version). Once items had been created, an internal vetting process was carried out, resulting in our PILOT Questionnaire (for all items, see Appendix 1), comprising 64 items in Part A, 40 items in Part B, 19 questions in Part C and 9 questions in Part D. The final step in Phase II was to build an online version of the PILOT using the software Survey&Report (Artologik, n.d.).

Phase III: Administering and evaluating the PILOT Questionnaire

Administration

Prior to its distribution, the PILOT Questionnaire went through 'technical piloting' among colleagues in order to ascertain that it functioned well regardless of the device used when responding. For distributing the PILOT Questionnaire, 45 English teachers from our professional networks were approached, asking them to respond to the extensive pilot version. In total, 23 teachers replied (response rate: 51%). The data collected were exported into statistical software for the analytical work (IBM SPSS 25).

Analysis of teacher feedback in the PILOT Questionnaire

Part D included evaluative questions, including specific questions about each of the six constructs, to find out to what extent the respondents thought they had answered questions about these. The means for the six constructs ranged from 4.39 (*SD* = 1.78) for Construct 3 (*The current language situation in Sweden*) to 5.91 (*SD* = 0.29) for Construct 4 (*Use of background languages when learning an additional language*). In short, the responding teachers stated that they had answered questions about all six constructs. The greatest spread in answers was found for Construct 3 (the language situation in Sweden), with answers scattered across the whole scale. Thus, items in Construct 3 were less salient to the respondents than items belonging to the other constructs.

As expected, the PILOT Questionnaire took a long time to answer, ranging from 15 minutes to more than 40. Thus, several items were deleted when creating the final version.

Item analysis

Item ananlysis was important and entailed analyzing the items in relation to the assumed multi-item scales. Corrected item-total correlations and reliability coefficients were computed in SPSS. The items were then perused in a step-wise process as to their fit into the multi-item scale. The goal was to reach as high a reliability as possible with a scale consisting of 3–5 items. As an example, the items aimed at targeting Construct 3 (*The current language situation in Sweden*) are provided in Table 4.2. The initial scale consisted of six items, and the reliability was 0.574, which is on the low side. The removal of Item A3.2 (see Table 4.2) increased the reliability to 0.735, and reliability was observed at 0.822 through the removal of Item A3.4. As can be seen in Step 3, an even higher reliability could be reached by deleting Item A3.1, but this was felt to have a detrimental effect on the dimension targeted in the construct, as well as bringing the number of items down to three. Therefore, no further deletions were made. The same procedure was subsequently used for all the other scales. The resulting list of items is attached in Appendix 1.

Table 4.2 Cronbach's alpha for items in Construct 3

	Cronbach's alpha if item deleted		
Item	**Step 1**	**Step 2**	**Step 3**
A3.1. In Sweden, it is important that students with another home language than Swedish to keep this language alive	0.547	0.719	0.900
A3.2. In Sweden, in addition to Swedish, it is more important to know English than any other language	0.735	DELETED	DELETED
A3.3. In Sweden, your chances of getting a job increase if you are multilingual	0.405	0.602	0.710
A3.4. I think that the status of the Swedish language is threatened by other languages	0.636	0822	DELETED
A3.5. If you learn English well. your chances of getting a job in Sweden increase	0.403	0.659	0.776
A3.6. If you learn several languages, your chances of getting a good job in Sweden increase	0.352	0.579	0.690
Total Cronbach's alpha	0.574	0.735	0.822

Validation: External raters relating items to constructs

To investigate content validity, data were collected from two external raters. Rater 1 was a senior Humanities researcher, and Rater 2 was a junior scholar in the field of English Linguistics, with expertise in statistics. The raters were presented with all the items in the PILOT Questionnaire, alongside the six constructs, and were asked to categorize each item into these constructs. The external ratings were then compared to that of the research group. According to Altman (1991), pair-wise correlations between 0.60 and 0.80 are considered 'good'. Here, all pair-wise correlations fell within this range (Rater1×ResearchGroup: $r_s = 0.655$; $p < 0.001$; $N = 98$; Rater2×ResearchGroup: $r_s = 0.778$; $p < 0.001$; $N = 102$ and Rater1×Rater2: $r_s = 0.716$; $p < 0.001$; $N = 98$). Using Krippendorff's alpha (Hayes & Krippendorff, 2007), inter-rater reliability for the three ratings reached 0.72, a modest but acceptable result, which was considered satisfactory.

External expert

Another strategy to enhance content validity involved asking a linguist, external to the research group, with expertise in multilingualism to assess the quality of the questionnaire ('external audit', Johnson & Christensen, 2017: 299), leading to further changes. For instance, we streamlined terminology and specified definitions (*multilingualism*, *background languages*).

Content and design of FINAL Questionnaire

Based on the above analyses and steps, the FINAL Questionnaire consists of 39 items in Part A (64 in PILOT), 38 items in Part B (40 in PILOT) and

19 questions in Part C (19 in PILOT). The FINAL Questionnaire was built in Survey&Report (Artologik, n.d.).

Phase IV: Administering and evaluating the FINAL Questionnaire

Administration

A stratified random sample of L2 English teachers was drawn using statistics from Statistics Sweden coupled with school data from the National Agency for Education. This resulted in the questionnaire being distributed to 441 teachers. It remained open for four weeks, with reminders issued at the end of the first and second weeks. A few automated responses were received from teachers on leave; teachers could also opt out of responding. This lowered the number of respondents to 321. When the questionnaire closed, 139 (43%) teachers had responded, which is a respectable number compared with other studies (e.g. Granfeldt *et al.*, 2019, 35%; Henry *et al.*, 2018, 44%) and higher than rates reported in the studies reviewed above. The sample consisted of 103 women (74.1%), 32 men (23%) and 4 individuals who preferred not to state their gender (2.9%). In sum, it was reasonable to consider the random sample representative of the statistical population (see Appendix 2).

Item analysis and factor analysis

Like the PILOT data, items in the FINAL Questionnaire were subjected to item analysis. As a first step, the scoring of items with a reversed phrasing was corrected as such items, if uncorrected, are known to affect reliability (Field, 2013). Next, a reliability coefficient (Cronbach's alpha) for all the 76 items (Parts A and B) was computed and observed at 0.88. The reliability statistics of the 10 multi-item scales are provided in Table 4.3.

As can be seen, most reliabilities were acceptable, with many values close to or well above 0.7. However, scales for B1 (*Openness towards other cultures*) and B2 (*Multilingualism in general*) were clearly below levels aspired to. A reasonable explanation is that teachers' classroom practices do not necessarily mirror school practices. For MultiBAP, we include items from B5 (*Use of background languages in learning and using English*) and B6 (*Monolingual beliefs in education*), because these scales were reliable.

Table 4.3 Cronbach's alpha reliability for multi-item scales in the FINAL Questionnaire

	Multi-item scales									
	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**B1**	**B2**	**B5**	**B6**
Alpha	0.68	0.58	0.71	0.60	0.88	0.72	0.30	0.46	0.86	0.71

Even though MultiBAP was based on six assumed constructs, we could not be sure whether the items technically mapped onto the constructs. One reason was that most items used had not previously been part of a questionnaire. Therefore, we carried out an EFA rather than a confirmatory factor analysis (CFA).

Factor analysis (FA) comprises 'an array of multivariate statistical methods used to investigate the underlying correlations among a set of observed variables' (Loewen & Gonulal, 2015: 182) and can be divided into EFA and CFA. As we could not ascertain the number and nature of underlying factors, an EFA rather than a CFA was used. Furthermore, EFA can be divided into EFA and PCA. Conceptually, the difference between PCA and EFA has to do with how the models treat variance; PCA analyzes variance, whereas EFA analyzes covariance. In other words, PCA does not differentiate between variance that is shared versus unique among variables, but EFA does. In many cases, PCA and EFA results are very similar, but not always. Conway and Huffcutt (2003) advise that researchers whose purpose it is to understand the underlying structure of a set of variables should decide on a common factor model (EFA) such as principal axis or maximum likelihood factoring, whereas purposes of pure reduction of variables calls for PCA. We therefore opted for an EFA common factor model.

In preparation for running the EFA, we concluded that many Part B items involve reported practice in the respondents' classrooms, but also practices at their schools, and beliefs presumably held by principals. Responses to such disparate items may not necessarily correlate. For this reason, we carried out the EFA only on Part A items.

First, it was necessary to investigate the factorability of the data. A wide range of scholarly advice is given in this regard. In the case of sample size, Loewen and Gonulal (2015) conclude that suggestions for minimum absolute sample sizes vary between 100 and 500. An alternative is to consider the number of respondents per item, where recommendations also vary. Based on their review of the literature, Loewen and Gonulal (2015) report on a range between 3 and 20, and Field (2013) report on a range between 10 and 15. In MultiBAP, Part A (beliefs) included 3.66 respondents per item, thus somewhat low. However, not only absolute sample size matters, and when in doubt, a number of statistical tests should be run. Therefore, the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was used. KMO values range from 0 to 1; the higher the value, the better sampling adequacy. Our value was 0.78, which is considered 'good', bordering on 'great' (Loewen & Gonulal, 2015: 187). Furthermore, to test for undesirably low correlations overall, a Bartlett's test was used. The result was significant, with $\chi^2\,(703) = 2807.346$, $p < 0.001$, meaning that the variables were sufficiently correlated and suitable for EFA. A related problem involves variables being too highly correlated (multicollinearity), with coefficients of around ±0.90. Only one case of such high

correlation was found (Q8 and Q9). Removing one of them did not improve the determinant, but this single case was deemed unproblematic in the light of the high number of items.

As FA seeks to determine 'the fewest number of variables that will still explain a substantial amount of variance in the data' (Loewen & Gonulal, 2015: 182), we employed several criteria to arrive at a decision that would chime well with that aim. One is based on a minimum Eigenvalue cutoff level. According to Kaiser's method, factors with Eigenvalues greater than 1 are retained; Appendix 4 shows that this would leave us with 11 factors. An 11-factor solution was deemed excessive, however, as we observed 1-item factors and factors in which the items were very disparate. Notably, the use of a Eigenvalue >1 in FA is referred to as 'inappropriate' by Pedhazur and Schmelkin (1991: 594), and Field (2013) argues that Kaiser's criterion works well with fewer than 30 items and sample sizes over 250. Another similar method is called Joliffe's criterion, by which factors with Eigenvalues greater than 0.7 would be retained. This would mean keeping even more factors (15); working with these many factors was not deemed feasible. We subsequently tried several analyses with 9, 8 and 7 factors. However, it was still difficult to arrive at satisfactory solutions. An important aspect for deciding on factors to retain is cumulative percentage of variance (CPoV). Plonsky and Gonulal (2015) report that the average CPoV in L2 research is approximately 60%, while Field (2013) suggests a minimum of 55–65%.

Adhering to these guidelines, with a cumulative percentage of 55.11%, six factors can be retained. Next, we checked communalities (h^2) as these can provide an indication of the relationship of each variable to the entire dataset. High communalities are desired, and the mean value for our 38 items after extraction was 0.47 (SD = 0.21). A final potential criterion is a scree plot, where the point of inflexion indicates the cutoff point for selecting factors (Figure 4.1). Scree plots are notoriously difficult to interpret and should only be used in light of other selection criteria (Loewen & Gonulal, 2015). In our case, there were many potential cutoff points, and in our interpretation, the plot did not yield a clear picture.

Through a concerted approach, then, drawing on Kaiser's test, Bartlett's test, CPoV and a scree plot, we ultimately decided to retain six factors. This yielded a respectable CPoV of 55% (in line with Field, 2013). As the extraction method, we used maximum likelihood factoring for the analysis of the 38 items in Part A (beliefs). We used oblique rotation, as high correlations were expected for our data (see Loewen & Gonulal, 2015: 197). The rotated factor loadings for the six factors are provided in the form of a pattern matrix in Appendix 3. This type of factor loading matrix is often considered more meaningful and interpretable. As suggested in Loewen and Gonulal (2015), all loadings of < 0.30 have been suppressed. As seen in the matrix, there were deviations from the intended subscales for the 38 items in the sense that the items did not always load

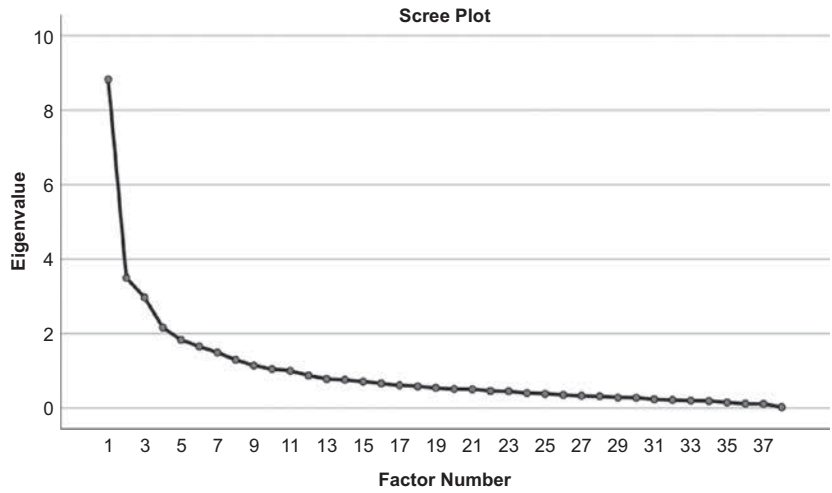

Figure 4.1 Scree plot of components and their associated Eigenvalues

on our six hypothesized constructs. The question code in the left-most column reveals the deviations (items starting with number 1 = *Construct 1*, items starting with number 2 = *Construct 2*, etc.).

As argued by Loewen and Gonulal (2015), when naming a factor, it is important to come up with a descriptive label that represents all items loading onto that particular factor, paying particular heed to items that have the highest load. The four items from the hypothesized Construct 1 (*Openness towards other cultures*) all mapped on Factor 6. In addition, so did one item from Construct 2 (*Multilingualism in general*) and one item from Construct 5 (*Use of background languages in learning and using English*). An analysis of what these items focus on resulted in the factor label *Openness towards other cultures*.

For Factor 5, high loadings from four items from three different hypothesized constructs were observed. These items seemed to focus on the importance of maintaining other languages than the majority language (Swedish). Factor 5 was consequently labeled *Importance of maintaining other languages than the majority language (Swedish)*. Four items from the hypothesized Construct 6 loaded highly on Factor 4. What these items seemed to have in common was *The importance of proficiency in the majority language*.

Regarding Factor 3, four items loaded on this factor, dominated by three from the hypothesized Construct 3 (*The current language situation in Sweden*), and with one item from the hypothesized Construct 2 (*Multilingualism in general*). These items rendered the label *Importance of multilingualism for future employment and success in Sweden*.

For Factor 2, no less than 11 items were observed with high loading: 8 items from the hypothesized Construct 5 (*Monolingual beliefs in*

education), 2 items from Construct 4 (*The use of background languages when learning an additional language*) and 1 item from Construct 6. The common denominator was seen as *Positive attitudes to background languages when learning English*.

Finally, for Factor 1, four items were observed to have high loadings. They all came from the hypothesized Construct 5 (*The use of background languages when learning and using English*). An analysis rendered the following label: *Importance of background languages for receptive and productive English skills*.

Phase V: Content and design of MultiBAP questionnaire

The analysis accounted for above resulted in a set of 33 multiscale items for MultiBAP Part A (beliefs). In order to check the reliability of the new subscales, Cronbach's alpha was computed (see Appendix 3). The reliability values observed were 0.84, 0.80, 0.81, 073, 0.68 and 0.75, with a mean of 0.77. This is wholly satisfactory as most guidelines point to 0.7 as a desirable minimal level (Dörnyei & Taguchi, 2010).

As regards Part B (practices), there was no EFA to rely on. However, the multiscale item reliability analysis revealed that two of the 'original' constructs (B5, *Use of background languages in learning and using English*, and B6, *Monolingual beliefs in education*) in the FINAL Questionnaire were reliable, and it was therefore decided to include them in the MultiBAP Instrument (see Appendix 1). Altogether, Part B of the MultiBAP Instrument includes 31 closed items and 1 open question. In sum, then, the MultiBAP Instrument contains two parts: 'Beliefs' (33 closed + 1 open) and 'Practices' (31 closed + 1 open), in total 66 items/questions (64 closed + 2 open). Note that both original constructs B1 (*Openness towards other cultures*) and B2 (*Multilingualism in general*) were unreliable and therefore excluded. However, although excluded as 'scales', individual questions in B1 and B2 may nevertheless be useful in future studies, as answers to the various questions can be informative. For example, in multilingual settings, to what extent do schools view students' cultural backgrounds as resources (see B1.3, Appendix 1)? In addition to the MultiBAP *Instrument*, the full-length MultiBAP *Questionnaire* also contains the items/questions included in B1, B2 and Part C of the FINAL Questionnaire (see Appendix 1).

Discussion

We aimed to account for the development and initial validation of MultiBAP, a questionnaire instrument designed to map teacher beliefs and practices, as well as school practices, about multilingualism. A review of existing instruments revealed a lack of one that served the purposes of our parent study, MultiLingual Spaces (Källkvist *et al.*, 2017). The

construction process was guided by best practice advice *inter alia* in Dörnyei and Taguchi (2010), Loewen and Gonulal (2015) and Plonsky and Gonulal (2015). The result is the questionnaire instrument named MultiBAP, included as Appendix 5.

Initial validation of MultiBAP entailed going from *a priori* postulated constructs and pertinent multi-item scales to an evidence-based modification of these. This modification entailed revising the content in Part A in the light of an EFA. Such analysis provided construct-related validity in the sense that we sought to investigate what latent traits our instrument was tapping into. The EFA made us modify the way in which items were linked to assumed constructs. For example, all the items assumed to relate to the *a priori* construct *Openness towards other cultures* clustered together with one item from the *a priori* construct *Multilingualism in general*, and another from *Use of background languages in learning and using English*. There were also some interesting groupings of items, such as the separation of items related to the importance of drawing on background languages for receptive English skills from items related to the importance of drawing on background languages for productive skills. The mean scores of the items linked to those two factors reveal that items targeting receptive skills received higher scores than items targeting productive skills. This could emanate from a belief that receptive skills such as listening and reading may involve an individual's background languages more so than the productive skills.

In terms of reliability, the multi-item scales in MultiBAP Part A rendered respectable coefficients, as did two of the scales in Part B. Thus, this aspect of validity is promising. However, the type of reliability used is sample-dependent, and technically not really a characteristic of the instrument itself, but rather of the sample scores. As suggested by Knoch and McNamara (2015), this can be overcome through the use of Item Response Theory (IRT) approaches, such as extended Rasch models. Consequently, such analyses could provide for further validation of MultiBAP.

Limitations

Some limitations need to be addressed. For example, it was not possible to carry out a factor analysis of Part B items. Thus, only results from reliability analyses of the FINAL Questionnaire are available. Although the overall reliability of Part B was good (0.894), the reliability of constructs B1 (0.3) and B2 (0.465) was unsatisfactory. Thus, if used, this must be kept in mind. In contrast, the reliability values of constructs B5 (0.855) and B6 (0.712) were high, so those constructs can be used. Another potential limitation is the number of respondents. Admittedly, a higher number would have been preferred, but considering the time and effort invested in establishing a random sample, the outcome was satisfactory, in particular in light of multilingualism in Swedish schools being a politically charged

topic at the time (and still is). Finally, the number of respondents comes out well in comparison with previous questionnaire studies of teacher beliefs (cf. Borg, 2015), and the response rate is in line with similar studies (Granfeldt *et al*., 2019; Henry *et al*., 2018).

Suggested use

The developed and validated questionnaire consists of Parts A, B and C, of which the first two constitutes the MultiBAP Instrument. For example, MultiBAP can be adapted to mapping beliefs and practices about multilingualism in teaching other additional languages by replacing 'English' by another language. MultiBAP can also be used by teachers as a means of raising awareness and initiate professional discussion about prevailing beliefs in specific contexts. Similarly, Part C can be modified. Most likely, nine of the C-items (i.e. C1–C2, C5, C8, C11–C13, C17 and C19) target background variables that are core to many studies.

Conclusion

We have accounted for the construction, development and initial validation of MultiBAP, aimed at mapping teacher beliefs and practices about multilingualism. Care was taken to consider essential methodological procedures, and comprehensive reporting was provided for steps taken. It is hoped that our detailed appendices will aid future similar questionnaire design and validation projects. Suggestions for its use have been offered, outlining straightforward adaptations to contexts. Seeing the pursuit of validity (including reliability) as a perpetual process, initial evidence presented here is promising but may be extended, for example, by using interviews and think-aloud data from respondents while filling out MultiBAP. Finally, it goes without saying that mapping the beliefs and practices among the teachers in our sample is the ultimate aim of this research. These results gained from MultiBAP will be reported in Sundqvist *et al*. (in preparation).

Acknowledgements

The MultiLingual Spaces project was funded by the Swedish Research Council (Reg. No. 2016-03469). Appendices 1–5 are available at Multilingual Matters Resources: https://www.multilingual-matters.com/page/pttmep/

References

Altman, D.G. (1991) *Practical Statistics for Medical Research*. Boca Raton, FL: Chapman & Hall/CRC.

Artologik (n. d.) Survey&Report [computer software]. Retrieved from https://www.artologik.com/en/SurveyAndReport.aspx?pageId=223

Bailey, E.G. and Marsden, E. (2017) Teachers' views on recognising and using home languages in predominantly monolingual primary schools. *Language and Education* 31 (4), 283–306.

Bardel, C., Falk, Y. and Lindqvist, C. (2013) Multilingualism in multicultural Sweden. In D. Singleton and L. Aronin (eds) *Current Multilingualism: A New Linguistic Dispensation* (pp. 247–269). Boston, MA: De Gruyter Mouton.

Borg, S. (2006) *Teacher Cognition and Language Education: Research and Practice*. London: Continuum.

Borg, S. (2015) Researching teachers' beliefs. In B. Paltridge and A. Phakiti (eds) *Research Methods in Applied Linguistics: A Practical Resource* (pp. 487–504). London: Bloomsbury Academic.

Broca, Á. (2015) Questionnaires on L2 learning and teaching practices: Rating responses on frequency and opinions. *TESOL Quarterly* 49 (2), 429–440.

Busse, V., Cenoz, J., Dalmann, N. and Rogge, F. (2020) Addressing linguistic diversity in the language classroom in a resource-oriented way: An intervention study with primary school children. *Language Learning* 70 (2), 382–419.

Cenoz, J. and Genesee, F. (1998) Psycholinguistic perspectives on multilingualism and multilingual education. In J. Cenoz and F. Genesee (eds) *Beyond Bilingualism: Multilingualism and Multilingual Education* (pp. 16–32). Clevedon: Multilingual Matters.

Conway, J.M. and Huffcutt, A.I. (2003) A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods* 6, 147–168.

De Angelis, G. (2011) Teachers' beliefs about the role of prior language knowledge in learning and how these influence teaching practices. *International Journal of Multilingualism* 8 (3), 216–234.

Dörnyei, Z. and Taguchi, T. (2010) *Questionnaires in Second Language Research: Construction, Administration, and Processing* (2nd edn). London: Routledge.

Edstrom, A. (2006) L1 use in the L2 classroom: One teacher's self-evaluation. *Canadian Modern Language Review* 63 (2), 275–292.

Field, A. (2013) *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll* (4th edn). Los Angeles, CA: SAGE.

Granfeldt, J., Sayehli, S. and Ågren, M. (2019) The context of second foreign languages in Swedish secondary schools: Results of a questionnaire to school leaders. *Apples – Journal of Applied Language Studies* 13 (1), 27–48.

Graus, J. and Coppen, P.-A. (2016) Student teacher beliefs on grammar instruction. *Language Teaching Research* 20 (5), 571–599.

Grosjean, F. (2008) *Studying Bilinguals*. Oxford: Oxford University Press.

Gu, Y. (2016) Questionnaires in language teaching research. *Language Teaching Research* 20 (5), 567–570.

Gunnarsson, T., van de Weijer, J., Housen, A. and Källkvist, M. (2015) Multilingual students' self-reported use of their language repertoires when writing in English. *Apples – Journal of Applied Language Studies* 9 (1), 1–21. Retrieved from http://apples.jyu.fi/article/abstract/367

Hayes, A.F. and Krippendorff, K. (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1 (1), 77–89.

Henry, A., Korp, H., Sundqvist, P. and Thorsen, C. (2018) Motivational strategies and the reframing of English: Activity design and challenges for teachers in contexts of extensive extramural encounters. *TESOL Quarterly* 52 (2), 247–273.

Heyder, K. and Schädlich, B. (2014) Mehrsprachigkeit und Mehrkulturalität–eine Umfrage unter Fremdsprachenlehrkräften in Niedersachsen. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 19 (1), 183–201.

Hult, F.M. (2014) How does policy influence language in education? In R.E. Silver and S.M. Lwin (eds) *Language in Education: Social Implications* (pp. 159–175). London: Continuum.

Johnson, R.B. and Christensen, L. (2017) *Educational Research: Quantitative, Qualitative, and Mixed Approaches* (6th edn). Thousand Oaks, CA: SAGE Publications.

Källkvist, M., Gyllstad, H., Sandlund, E. and Sundqvist, P. (2017) English only in multilingual classrooms? *LMS Lingua* (4), 27–31.

Kern, R.G. (1995) Students' and teachers' beliefs about language learning. *Foreign Language Annals* 28 (1), 71–92.

Knoch, U. and McNamara, T. (2015) Rasch analysis. In L. Plonsky (ed.) *Advancing Quantitative Methods in Second Language Research* (pp. 275–304). London: Routledge.

Lee, J.S. and Oxelson, E. (2006) 'It's Not My Job': K–12 teacher attitudes toward students' heritage language maintenance. *Bilingual Research Journal* 30 (2), 453–477.

Leung, S.O. (2011) A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research* 37 (4), 412–421.

Levine, G.S. (2003) Student and instructor beliefs and attitudes about target language use, first language use, and anxiety: Report of a questionnaire study. *The Modern Language Journal* 87 (3), 343–364.

Lindberg, I. and Hyltenstam, K. (2013) Flerspråkiga elevers språkutbildning. In A. Flyman Mattsson and C. Norrby (eds) *Language Acquisition and Use in Multilingual Contexts: Theory and Practice* (Vol. 52, pp. 122–141). Lund: Travaux de l'Institut de linguistique de Lund.

Loewen, S., Fei, S.L., Thompson, A., Nakatsukasa, K., Ahn, S. and Chen, X. (2009) Second language learners' beliefs about grammar instruction and error correction. *The Modern Language Journal* I (1), 91–104.

Loewen, S. and Gonulal, T. (2015) Exploratory factor analysis and principal components analysis. In L. Plonsky (ed.) *Advancing Quantitative Methods in Second Language Research* (pp. 182–212). London: Routledge.

Lundberg, A. (2019) Teachers' beliefs about multilingualism: Findings from Q method research. *Current Issues in Language Planning* 20 (3), 266–283.

Mackey, A. (2012) Why (or why not), when and how to replicate research. In G. Porte (ed.) *Replication Research in Applied Linguistics* (pp. 21–46). Cambridge: Cambridge University Press.

Marsden, E., Morgan-Short, K., Thompson, S. and Abugaber, D. (2018) Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning* 68 (2), 321–391.

Nation, P. (2003) The role of the first language in foreign language learning. *Asian EFL Journal* 5 (2), 1–8. Retrieved from www.asian-efl-journal.com/june_2003_pn.pdf

Norris, J.M., Plonsky, L., Ross, S.J. and Schoonen, R. (2015) Guidelines for reporting quantitative methods and results in primary research. *Language Learning* 65 (2), 470–476.

OECD (2012) *Equity and Quality in Education*. OECD.org: OECD Publishing.

Pajares, M.F. (1992) Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research* 62 (3), 307–332.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design, and Analysis: An Integrated Approach* (Student edn). Hillsdale, NJ: Lawrence Erlbaum Associates.

Phakiti, A. (2015) Quantitative research and analysis. In B. Paltridge and A. Phakiti (eds) *Research Methods in Applied Linguistics: A Practical Resource* (pp. 27–47). London: Bloomsbury Academic.

Plonsky, L. (2015) Introduction. In L. Plonsky (ed.) *Advancing Quantitative Methods in Second Language Research* (pp. 3–8). London: Routledge.

Plonsky, L. and Gonulal, T. (2015) Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning* 65 (S1), 9–36.

Pulinx, R., Van Avermaet, P. and Agirdag, O. (2015) Silencing linguistic diversity: The extent, the determinants and consequences of the monolingual beliefs of Flemish teachers. *International Journal of Bilingual Education and Bilingualism*, 1–15.

Spada, N., Barkaoui, K., Peters, C., So, M. and Valeo, A. (2009) Developing a questionnaire to investigate second language learners' preferences for two types of form-focused instruction. *System* 37 (1), 70–81.

Spratt, M. (1999) How good are we at knowing what learners like? *System* 27 (2), 141–155.

Sundqvist, P., Gyllstad, H., Källkvist, M. and Sandlund, E. (in preparation) Multilingual classrooms in Sweden: English teachers' beliefs and practices.

Swedish National Agency for Education (2013) *Research for Classrooms: Scientific Knowledge and Proven Experience I Practice*. Stockholm: Swedish National Agency for Education.

Wagner, E. (2015) Survey research. In B. Paltridge and A. Phakiti (eds) *Research Methods in Applied Linguistics: A Practical Resource* (pp. 83–99). London: Bloomsbury Academic.

Valeo, A. and Spada, N. (2016) Is there a better time to focus on form? Teacher and learner views. *TESOL Quarterly* 50 (2), 314–339.

van Lier, L. (2006) Preface. In P. Kalaja and A.M. Ferreira Barcelos (eds) *Beliefs About SLA. New Research Approaches* (pp. vii–viii). New York, NY: Springer Science+Business Media.

Winke, P. (2011) Evaluation the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly* 45 (4), 628–660.