Regular Article

Identifying the Core Indicators of Migrant and Refugee Children's Integration Using the Delphi Method: A Multi-Input Strategy for Definition of Consensus

International Journal of Qualitative Methods
Volume 22: 1–11

DOI: 10.1177/16094069221149487
journals.sagepub.com/home/ijq
SAGE

Eva Bajo Marcos, Ángela Ordóñez-Carabaño, Elena Rodríguez-Ventosa Herrera, and Inmaculada Serrano

Abstract

This paper presents the Delphi methodology employed to select a final dashboard of 30 indicators on the socio-educative inclusion of refugee and migrant children in Europe. Firstly, a procedure for identifying Key Performance Indicators (KPIs) was carried out, including a specialized scientific literature review, the mapping of previous indicators, and qualitative workshops with key stakeholders at micro, meso, and macro levels in six countries. Then, a Delphi design was implemented to assess, rate, and provide meaningful qualitative improvements to a pool of pre-selected indicators. The Delphi methodology involved a group of international experts on the matters of inclusive education or migration, researchers, NGOs, and public officers. As an alternative to traditional "benchmark-based" consensus, we introduced the use of a) the CARA model and b) an alternative multi-input and mixed-method consensus-building procedure. The results provided a significant contribution to qualitative methods on the one hand and to migration and integration literature on the other. The methodological innovations, the diversity of experts' perspectives involved in the process, and the structured nature of the method constituted significant advantages to improve the robustness of the Delphi methodology for selecting and validating indicators. Future research involving a Delphi methodology can benefit from applying the present procedure.

Keywords

delphi method, migrant children, co-creation, consensus building, qualitative methods

Introduction

Migration is a phenomenon currently marking European countries' international and national agendas. As of January 2020, 23 million non-EU citizens were living in the European Union, making up to 5.3% of the EU population, representing the highest percentage reached in history so far (Eurostat, 2021a). Approximately one-third of them are children, many of whom are unaccompanied (accounting for almost 14,000 children in 2020) or seeking refuge (almost 130,000 children in 2020) (Eurostat, 2021a, Eurostat, 2021b, UNICEF, 2022). In the face of this demographic pressure, the successful integration of these children is key to the region's social cohesion and the sustainable development of European societies in the future (European Migration Network, 2022).

Despite the international agreement on the benefits of integration and intercultural dialogue, assessing and monitoring the inclusion of migrant and refugee children has been problematic. The reason lies in the lack of relevant data to describe their current regional circumstances and the vast diversity within this group (You et al., 2020). Migrant and

Comillas Pontifical University, Madrid, Spain

Authors by alphabetical order.

Corresponding Author:
Ángela Ordóñez-Carabaño, MSc, Comillas Pontifical University, C. de Alberto Aguilera, 23, Madrid 28015, Spain.
Email: aordonez@comillas.edu

refugee children face several new hurdles related to adapting to the new country and how the institutions and citizens welcome them, which jointly shape the integration process of children. Classic examples of these challenges are adapting to a new school system and syllabus, sometimes in a different language, making new relationships, and finding a way to integrate the differences between the host culture, its customs and traditions, and their own ones (OECD, 2018). Consequently, these children often develop a series of academic, social, and psychological needs that need to be covered.

Addressing the diversity of students has become a priority for educational communities. Thus, schools are turning to inclusive educational models able to provide a framework to successfully engage challenged students in the current educational programs (European Commission/EACEA/Eurydice, 2019). However, no consensus strategy has been reached on how to apply these models. This is largely due to a deficiency of high-quality data and monitoring tools on migrant integration (Asis et al., 2018; UNECE, 2019). In this regard, indicators represent a valuable tool optimal for assessing and monitoring the degree of socio-educational integration of migrant and refugee children in schools. Specifically, indicators are crucial tools to monitor performance and inform strategic decision-making while providing data on the quality of an activity, project, or program. They must be concise, meaningful, and feasible measures informing credible and precise data (Ager & Strang, 2004; Booth & Ainscow, 2002; Huddleston et al., 2013; Thoreau & Liebig, 2018). These indicators must have a high level of validity, that is, the extent to which a measure accurately represents the intended concept (in this case, socio-educational inclusion) (OECD, 2008).

Several sets of indicators have been developed to measure the integration of migrants (mainly identifying structural and cultural dimensions of integration). For instance, in the European context, the most relevant sets of indicators are the Zaragoza Declaration (Council of the European Union, 2010), the Migrant Integration Policy Index (Huddleston et al., 2015), and the Settling in indicators (OECD/EU, 2018). These indicators systems provide a comparative perspective on the socio-political contexts of integration and their outcomes across different countries in the region from a macro-level approach. They focus on legislation and migration control; the social services involved and how they perform in different life domains; or the demographic and economic balance derived from the incorporation of migrants into the host country. However, all of them are built to reflect adult realities, focusing on key aspects such as employment and unemployment rates, qualification levels, housing conditions, and the possibility to vote, among others. None have been specifically designed to measure the inclusion of migrant and refugee children. Also, neither have they included the children's perspectives in the definition of how integration materializes for them (Bajo Marcos et al., 2022). In this regard, the available evidence from childhood and migration studies suggests that indicators aiming to reflect migrant children's integration experiences need to focus on different dimensions (Van Vooren & Lembrechts, 2021). In particular, they stress the need to focus more on the educational and socio-emotional aspects of children's lives on top of the more political and intercultural aspects that are also relevant for adults (MiCreate, 2019).

This paper describes the Delphi methodology employed to obtain valid indicators of migrant and refugee children integration in Europe. The Delphi study presented in this paper was designed as means to achieve one of IMMERSE project's main objectives: creating of a dashboard of socio-educational integration indicators for migrant and refugee children in Europe. The project aims to map these children's integration in different European host countries through the definition of new research methods and tools, incorporating children's voices throughout the process and collecting data representative of the European reality. Its ultimate goal is to make policy recommendations and encourage relevant stakeholders in their adoption. A child-centred approach was adopted, including a participatory design with children and other relevant stakeholders that enriched the interpretative analysis by fostering a dialectic exchange between researchers and participants (MiCreate, 2019).

The origin of the name "Delphi" derives from the Geek mythology and is related to the oracle of Delphi; therefore, it represents a method that pursues the forecasting of a particular future event or decision. This method was firstly introduced in the 60s by the seminal work of Dalkey and Helmer (1963) and the subsequent developments done by Linstone and Turoff (1975), and it stems from the epistemological assumption that scientific consensus reflects shared knowledge among experts (Miller, 2013). This methodology relies on structured discussion rounds among a group of experts dealing with a complex problem to come to terms about a specific scientific question (Brady, 2015; Linstone & Turoff, 1975). This is achieved through the design of a questionnaire in which participant's anonymity is guaranteed and all their responses are re-organised and provided back to the participants for them to revise their previous responses and provide further feedback until options are narrowed down, and consensus is reached (Linstone & Turoff, 1975). Thus, according to hybrid theorists of consensus and dissent, the Delphi method meets the three conditions that ensure shared knowledge as the best explanation for a consensus. These conditions are social calibration, apparent consilience of evidence and social diversity (Miller, 2019).

In our study, the Delphi methodology sought consensus on the key determinants and outcomes in the socio-educational integration of migrant children, starting with an initial list of 57 indicators.[1] Although the notion of consensus is fundamental to Delphi studies, what constitutes consensus and how it is reached remains contested and one of the least standardized aspects of the methodology (Boulkedid et al., 2011; Diamond et al., 2014; Jünger et al., 2017; Miller, 2019). As reported in different systematic reviews, most studies set an a

priori criterion or cut-off, which is fundamentally arbitrary, sometimes not met or revisited, or either resort to post hoc (Boulkedid et al., 2011; Diamond et al., 2014; Jünger et al., 2017). Defining an adequate procedure for consensus building is particularly challenging when applied to a subject matter that is complex and multidimensional, applied to very heterogeneous populations and contexts, and involving multiple perspectives (van der Schaaf & Stokking, 2011).

In the initial list of 57 indicators, we found plenty of dimensions and factors relevant to children's socio-educational integration in our case. This selection included children of different ages, backgrounds, legal status, and migration history in 6 different countries, as well as different stakeholders such as educators and policymakers. Finally, we also included different perspectives based on discipline and specific expertise relevant to the children's integration. The main goal of the Delphi was to reduce this complexity while ensuring that all relevant dimensions were represented in the selected set of indicators. However, we found that the traditional procedure of predefining a benchmark for consensus was too limited for this goal. Instead, we introduced an alternative multi-input procedure that helped us ensure our numerical target of adequate and relevant selected indicators, their empirical soundness, and the necessary representation of the multiple dimensions involved.

In the following sections, we first present the Delphi study. Then we describe the results obtained from applying both the more traditional "benchmark-based" approach and the alternative procedure we introduce for the definition of consensus. Finally, we discuss the implications of this approach and conclude with a proposal of using a multi-input and mixed-method strategy for the definition of consensus in Delphi consultations.

Materials and Methods

A design aiming at two milestones was agreed to build the system of indicators: to obtain a first pool of indicators and then select from these a dashboard of 30 indicators. This threshold of 30 indicators was based on a parsimony principle and set in accordance with the central limit theorem which states that a minimum of 30 indicators provides sufficient heterogeneity while still allowing the dashboard to be manageable for monitoring (Kwak & Kim, 2017).

The first preselection of indicators was based on: a) a literature review that pointed to critical factors in socio-educational inclusion and migrant integration; b) qualitative research to collect the priorities, perspectives, and views of children and other relevant stakeholders. This research included workshops with children and parents, focus groups and a world café with professionals of intervention, and in-depth interviews with public servants and governance representatives;[2] and c) a compilation and assessment of previously developed indicators from secondary sources. As a result of this process, a pool of 57 indicators was selected based on: a) the theoretical research pointing to the need for gathering consensual integration aspects that are relevant and adequate across different contexts and approaches, and b) the technical approach, including those factors with documented empirical robustness (meaning proved efficiency and feasibility), and those that allowed to balance statistical measurement and sufficient observation of all ecological levels (micro, meso, macro) (Heink & Kowarik, 2010).

The Delphi methodology was then implemented to reduce the dashboard to 30 + 5 key performance indicators (KPIs). An additional target of 5 indicators was included to ensure that a minimum of 30 would remain in case some might be dropped in the final ecological validation of the set conducted after the Delphi process.

Participants

Each partner member of the IMMERSE consortium provided the contacts of at least two international experts in education or child migration. A group of 61 top international experts was contacted via email. Of these, 20 did not reply, and 14 alleged difficulties to participate (due to the dates of the study, the work overload, and the need for approval within their organization). Of the 27 experts that agreed to participate, 3 of them finally did not participate once the consultation rounds started. The Ethics Committee of the Comillas Pontifical University approved the Delphi consultation, and the experts provided consent through the software used to conduct the Delphi Study, where they were previously informed about the methodology.

Finally, 24 top international experts in education and migration, academia, NGOs, and public administration participated in the content validation of the pre-selected set of 57 indicators of migrant and refugee children's integration.[3] Following the criteria of social diversity, the researchers sought to recruit experts balancing the gender and age composition of the sample. Additionally, looking to accomplish the criteria of an apparent consilience of evidence and social calibration, the recruitment included a balanced group of experts on migration and/or education with more than ten years of interdisciplinary professional experience that would guarantee that knowledge was to be built upon heterogeneous theoretical and methodological backgrounds (Miller, 2013).

The final profile composition of this group included 12 males and 12 females; 10 were specialists in migration, nine experts in education, and five mixed profiles specialized in both socio-educational inclusion and migrant youth. The current affiliation of the experts included 13 researchers, four consultants with previous experience in migration policy incidence, two high-level representatives of NGOs, three advisors in education, and two advisors with expertise in governance and socio-educational inclusion of migrant children. The average years of professional experience of the experts were 21.6 and all participants had been working in the field for at least 10 years. The large sample of participants, their level of expertise, reputation and specialization, and the

Table 1. Accumulated frequencies of the experts by field of expertise.

Fields of Expertise (Multiple Options)	N
Education	17
Migration	12
Public Policy	5
Childhood	5
Mental Health	4
Refugee studies	4
Education in crisis and post-crisis	1

Table 2. Accumulated frequencies of number of experts by country of origin.

Experts' Countries of Origin	N
Spain	10
Greece	3
United States	3
Italy	2
Australia	2
Belgium	1
Ireland	1
Colombia	1
United Kingdom/France (mixed origin)	1

heterogeneity of their profiles helped to ensure the robustness of the process (Tables 1 and 2).

Procedure

The Delphi consultation was conducted online using the Calibrum software, which allowed the participants to remain anonymous to each other while still being able to see the other participant's contributions. The process to reach the consensus consisted of two consecutive rounds of consultation.[4] The procedure followed in both rounds consisted in presenting each indicator on one screen with the name of the factor, the description of the empirical measurement, and details about the source. The experts were asked to provide for each indicator a score on four dimensions, corresponding to the CARA model developed by Hernández Franco et al. (2009):

- *Clarity* of the indicator: whether the indicator is drafted in a concrete and non-ambiguous way and has a single possibility of interpretation. The experts must rate each indicator on a four-value scale from 1 (very low) to 4 (very high clarity).
- *Adequacy* of the indicator: whether the indicator is appropriate and it refers to the key or highly influential factors to achieve the socio-educational integration of migrant children. The experts must rate each indicator on a four-value scale from 1 (completely disagree) to 4 (completely agree) that the indicator is adequate.
- *Relevance* of the indicator: whether the indicator is essential regarding public policies or educational centers to accomplish their mission of socio-educational integration of migrant children. The experts must rate each indicator on a four-value scale from 1 (not important at all) to 4 (very important).
- *Accessibility* of the indicator: whether there are sources of accessible information that could let us obtain the necessary data to make a reliable indicator measurement. The experts must indicate for each indicator 'Yes, it is available' or 'No, it is not available.'

The experts could add voluntary comments for each indicator and overall comments to the complete inventory. In the second round, the experts had access to the other participants' comments and ratings for each indicator, and they could change their assessments and comments. Finally, at the end of each round, the experts were asked to select the five indicators that, according to their expert judgment, best represented the key factors that are most important to measure the socio-educational integration of migrant children. Each consultation round involved a time commitment of one to two hours depending on the time the expert dedicated to each indicator's response (although it was suggested to spend no more than 2–3 minutes per indicator).

The high profile of the international experts participating in the Delphi was matched with a highly committed response by most of them in providing substantial comments (optional and not explicitly requested from them) and intensively engaging with the first and second round mechanics. Most experts not only participated in the second round but also modified their inputs based on the reflections of other experts in the first round, engaging in a feedback dialogue through comments (Figure 1).

Results

In this section, we present the results obtained by applying two different approaches to consensus building: a more traditional "benchmark-based" consensus and a multi-input mixed-method.

We initially implemented a more traditional approach, consisting in pre-defining a quantitative benchmark for consensus (Linstone & Turoff, 1975). We established this benchmark at 60% of the experts picking the same (positive) value on the CARA dimensions.[5] Following this criterion, the first consultation round resulted in 11 indicators reaching a positive consensus among experts (see Table 3).

In the second round, the 11 indicators that had reached positive consensus were excluded, and the experts were asked to re-assess the remaining 46 indicators. This second round resulted in 5 more indicators reaching consensus (see Table 4). In total, only 16 indicators reached a consensus between the first

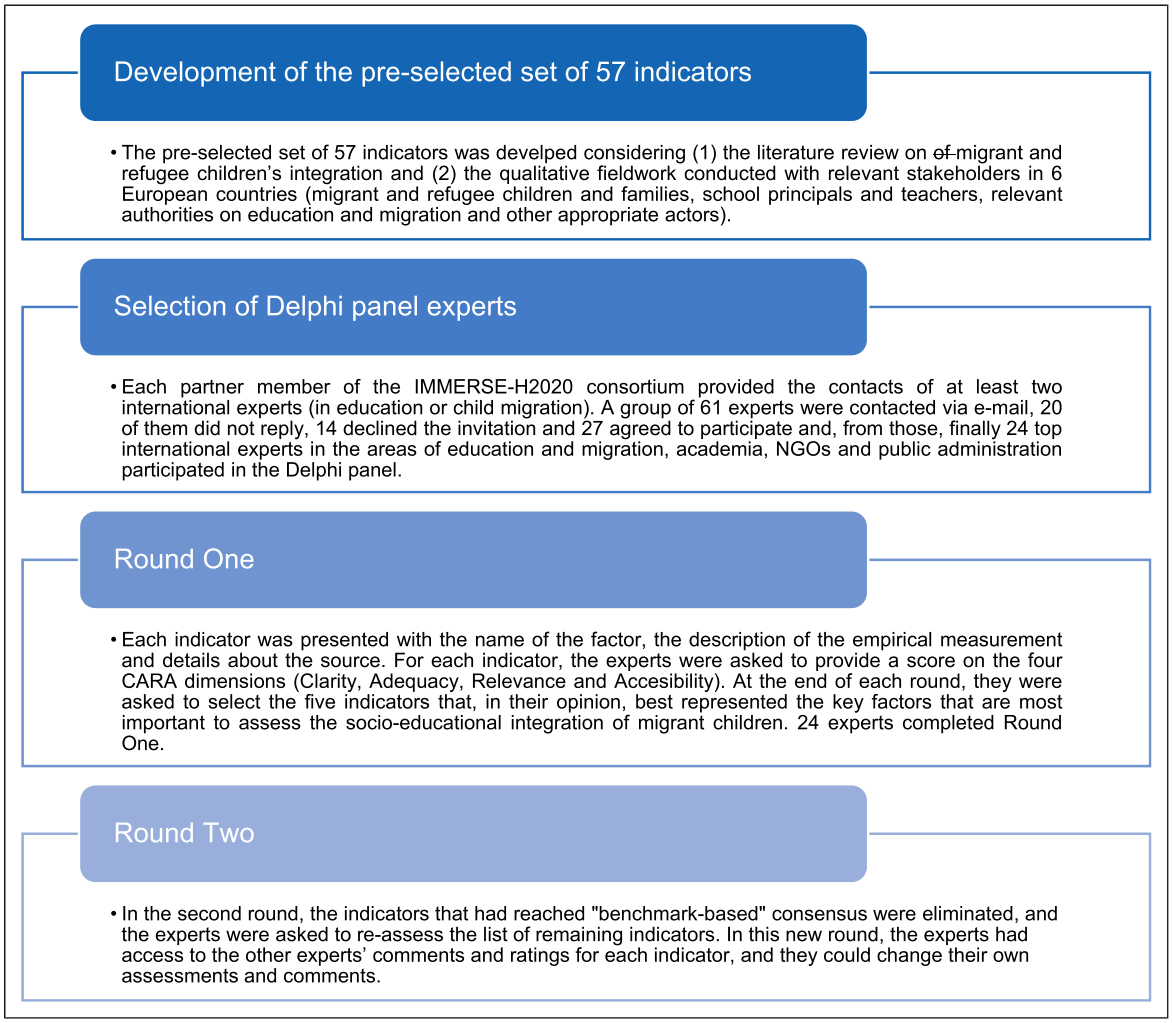

Figure 1. Flowchart of the Delphi panel process.

and second rounds following this approach ("benchmark-based consensus"). This was 19 indicators short for the final intended selection of 30 + 5 and insufficient to reflect all the relevant dimensions of integration. The implementation of a third round of the Delphi was discarded based on the dropout risk and the limited likelihood that the numerical target would be achieved.

So next, we turned from the exclusive "benchmark-based" consensus to an alternative strategy involving multiple quantitative and qualitative inputs in the consensus-building process. In this process, and considering experts' final responses on each indicator, we combined the following:

1. **Rankings of indicators based on the CARA criteria**: First, a primary ranking of indicators was built based on the average score (across experts in the second round) in the Adequacy and Relevance categories. These are the CARA categories that determine the relative importance for inclusion in the final set of indicators. Whereas (lack of) Clarity and Accessibility are practical criteria that affect the strategic exclusion of a given indicator.[7] A total of 31 indicators (at the top of this ranking) had a score above the average. Another primary ranking was built based on the average score across all four CARA categories, including Clarity and Accessibility. A total of 32 indicators had a score above 3 (out of 4). The overlap between both rankings was substantial: 31 indicators met both benchmarks. Numerically, this would have been enough for our intended goals.
2. **Additional quality criteria per indicator to improve the robustness of the results**: Two additional (and demanding) quality criteria were considered that denoted experts' consensus and overall prioritization: (1) indicators picked by more than a quarter of the experts as part of their top-5 selection; (2) indicators that received the maximum value in Adequacy or Relevance from 60% of the experts or more.
3. **Qualitative analysis of experts' comments per indicator**: The experts' qualitative comments on every single indicator were analyzed from different points of

Table 3. Ranking by the relevance of indicators reaching "benchmark-based" consensus in the first round on all four CARA dimensions.

Category	Dimension	Consensual Response[1]	Score	Stability[2]
Legislation recommendation and resources (LRR) devoted to preparatory classes dedicated/with focus on language acquisition: Whether there are (state/regional level) provisions of preparatory classes for newly arrived students with a focus on language/curriculum acquisition	Clarity	Very High	4	65.00%
	Adequacy	Completely agree	4	75.00%
	Relevance	Very important	4	85.00%
	Accessibility	Yes	—	100.00%
Children's sense of belonging: Average score in school belonging among migrant-background children	Clarity	Very High	4	76.19%
	Adequacy	Completely agree	4	71.43%
	Relevance	Very important	4	76.19%
	Accessibility	Yes	—	92.31%
Training and support resources on intercultural competences: Score in MIPEX[6] policy indicator on: Teacher and principals training to reflect diversity	Clarity	Very High	4	60.00%
	Adequacy	Completely agree	4	70.00%
	Relevance	Very important	4	75.00%
	Accessibility	Yes	—	100.00%
Intercultural competence as part of syllabus or/and transversally: Share of schools that declare teaching students intercultural competences	Clarity	Very High	4	75.00%
	Adequacy	Completely agree	4	75.00%
	Relevance	Very important	4	75.00%
	Accessibility	Yes	—	100.00%
Low expectations/stereotypes among teachers towards minority/migrant/low socio-economic background children: Average score per school on perceptions about prevalence among teachers at school of low expectations and negative attitudes towards some cultural groups	Clarity	Very High	4	60.00%
	Adequacy	Completely agree	4	70.00%
	Relevance	Very important	4	75.00%
	Accessibility	Yes	—	85.71%
Children's access to compulsory education: Foreign children at school as a share of foreign children in compulsory ages	Clarity	Very High	4	60.87%
	Adequacy	Completely agree	4	72.73%
	Relevance	Very important	4	73.91%
	Accessibility	Yes	—	89.47%
Experience of harassment and/or physical violence (incl. Bullying) outside the family: Share of migrant-background children who have been bullied in the last couple of months	Clarity	Very High	4	70.00%
	Adequacy	Completely agree	4	65.00%
	Relevance	Very important	4	73.68%
	Accessibility	Yes	—	88.89%
Children complete compulsory education: Share of persons with compulsory education completed among the foreign-born population aged 16–20 who arrived in the host country before age 15	Clarity	Very High	4	61.90%
	Adequacy	Completely agree	4	71.43%
	Relevance	Very important	4	71.43%
	Accessibility	Yes	—	94.44%
Clear leadership and school identity around intercultural values against xenophobia, prejudice and stereotypes: Share of schools implementing policies and practices to teach students how to deal with ethnic and cultural discrimination	Clarity	Very High	4	60.00%
	Adequacy	Completely agree	4	65.00%
	Relevance	Very important	4	65.00%
	Accessibility	Yes	—	100.00%
Share of migrant-background children who declare avoiding certain places for fear of being treated badly because of your cultural or ethnic background? Y/N	Clarity	Very High	4	70.00%
	Adequacy	Completely agree	4	60.00%
	Relevance	Very important	4	65.00%
	Accessibility	Yes	—	90.91%
Share of migrant-background children who have bullied someone in the last couple in the last couple of months[1]	Clarity	Very High	4	60.00%
	Adequacy	Completely agree	4	60.00%
	Relevance	Very important	4	63.16%
	Accessibility	Yes	—	80.00%

Note. This indicator was finally dropped due to the reasons explained below in the Results section.
[1]The "Consensual response" column represents the mode, i.e., the experts' most chosen value.
[2]"Stability" represents the proportion of experts out of the total that have chosen the Consensus Value. Consensus was reached when 60% or more experts selected the same (positive) value.

view, such as the reliability of answers, the potential for misinterpretation, and empirical evidence on the indicator's behavior. These comments allowed distinguishing between the two levels of selection involved in the consultation, that is, the variable we aimed to measure, on the one hand, and the empirical measurement that was proposed with that aim, on the other hand, introducing more refined assessments and considerations. For each indicator, we identified the aspects (positive or negative) where clear or wide consensus emerged in the comments and those where divergences emerged. And we singled out specific

Table 4. Ranking by relevance of indicators reaching "benchmark-based" consensus in the second round on all four CARA dimensions.

		FIRST ROUND			SECOND ROUND		
Category	Dimension	Consensual response[1]	Score	Stability[2]	Consensual response[1]	Score	Stability[2]
Legislation recommendation and resources (LRR) devoted to intercultural competence as part of syllabus or/and transversally: Average score in MIPEX policy indicators on school curriculum to reflect diversity; adapting curriculum to reflect the diversity	Clarity	High	3	50.00%	High	3	62.50%
	Adequacy	Completely agree	4	55.00%	Completely agree	4	62.50%
	Relevance	Very important	4	60.00%	Very important	4	70.83%
	Accessibility	Yes	—	100.00%	Yes	—	93.33%
Interconnectedness/Friends and peers: A. Average score in friends support among migrant-background children	Clarity	Very High	4	61.90%	Very High	4	62.50%
	Adequacy	Completely agree	4	50.00%	Completely agree	4	60.87%
	Relevance	Very important	4	60.00%	Very important	4	69.57%
	Accessibility	Yes	—	87.50%	Yes	—	82.35%
Types and levels of (formal) non-compulsory education attended: Share of persons who have completed or are currently attending upper secondary or tertiary studies in the host country, among the foreign-born population aged 16–24 and with studies completed or currently studying in the host country	Clarity	Very High	4	66.67%	Very High	4	66.67%
	Adequacy	Completely agree	4	60.00%	Completely agree	4	60.87%
	Relevance	Very important	4	52.63%	Very important	4	63.64%
	Accessibility	Yes	—	94.44%	Yes	—	95.24%
Criteria for incorporation to educational levels upon arrival: Adequacy rate: % Of migrant-background students enrolled in the educational level that theoretically corresponds to their age	Clarity	Very High	4	65.00%	Very High	4	62.50%
	Adequacy	Agree	3	45.00%	Agree	3	66.67%
	Relevance	Very important	4	40.00%	Important	3	62.50%
	Accessibility	Yes	—	93.33%	Yes	—	94.74%
Share of migrant-background children spending no time with their friends out of school	Clarity	Very High	4	71.43%	Very High	4	70.83%
	Adequacy	Agree	3	55.00%	Agree	3	65.22%
	Relevance	Important	3	55.00%	Important	3	60.87%
	Accessibility	Yes	—	92.31%	Yes	—	80.00%

[1]The "Consual response" represents the mode, i.e., the experts' most chosen value.
[2]"Stability" represents the proportion of experts out of the total that have chosen the Consensus Value. Consensus was reached when 60% or more experts selected the same (positive) value.

Table 5. Groups of indicators and criteria for their selection.

Group	Basis for Inclusion		Number of Indicators (Total: 57)
A	At the top positions of the CARA-based primary rankings (i.e., indicators with a score above average or above 3, respectively, in the two primary rankings)	With additional quality criteria	12
B		No additional quality criteria	19
C	Next in the primary rankings	With additional quality criteria	9
D	At the bottom of the primary rankings	No additional quality criteria	17

Note. see full detail in Annex 1.

comments pointing at severe limitations or positive qualities of any particular indicator and founded in qualified expertise (i.e., on the particular topic, measurement, or source, as in the case of MIPEX).

The 57 indicators were classified into four groups that helped prioritize their selection according to their position in the primary rankings and the presence or not of additional quality criteria (see Table 5):

- GROUP A included 12 indicators that were located at the top positions of the primary rankings (i.e., those indicators with a score above the average or above 3 in each of the rankings) and that also displayed some of the

additional quality criteria. These were automatically selected for inclusion in the dashboard.

- GROUP B included 19 indicators that were located at the top positions of the primary rankings but did not display any additional quality criteria. These indicators were second-prioritized for inclusion in the dashboard. Together with Group A, they added up 31 indicators.
- GROUP C included the nine indicators that followed those in groups A and B in the primary rankings, most of which also displayed additional quality criteria, adding up a pool of forty indicators for the final selection.
- GROUP D included the remaining 17 indicators at the bottom of the primary rankings, which furthermore did not display any additional quality criteria.

Group by group, the experts' qualitative comments for each indicator were considered. Based on the insight gained from their analysis, all indicators were improved by introducing further points of clarification and information. In some cases, additional refinements and improvements were introduced following the experts' suggestions, such as duplicating survey items for different populations (e.g., principals and teachers) to increase their robustness; widening or adjusting the definition of different groups of reference; resorting to a higher-order indicator (i.e., from an existing indicators system used as reference), etc. All of these changes were meant to increase the robustness and validity of the indicators. The suggestions sometimes led to a change in the survey items used as a basis for the indicator, for which we resorted to other well-established survey items where available or developed new adjusted ones following the experts' qualified suggestions.

Finally, indicators for which the experts raised significant concerns were dropped in those cases where no clear solutions or alternatives were suggested or found. Two indicators were dropped from Group B in this manner. As an illustration, one of these indicators referred to the *exercise* of bullying (complementing another indicator on the *experience* of bullying). While the variable we intended to measure was considered highly interesting and relevant by most experts, a number of them raised specific concerns (based on their specialized knowledge) about the low reliability of the answers on that particular topic using the innovative perspective proposed,[8] and that this was further matched with potential misinterpretation leading to severe social impacts.[9] Once these indicators were dropped, nine more still had to be selected. Group C contained nine indicators, but five were also dropped based on the experts' concerns. These indicators, with lower average scores generally, also received largely negative comments among those experts providing them. As an illustration, the indicator of bilingualism received exclusively negative comments (specifically building on expertise in education).[10] Finally, the remaining indicators were selected among Group D: proceeding by order of score in the quantitative results, we considered whether the qualitative comments allowed for refinements in the indicators that would further support their reconsideration or not. In this way, the final four indicators were selected.

In short, the combination of the quantitative rankings, additional quality criteria, and qualitative analysis ensured not only the selection of a sufficient number of indicators but also their empirical soundness and potential social impact. Following this nuanced procedure for consensus-building, the final selection of 30 + 5 indicators did not include two of the 16 indicators that reached consensus through the more traditional "benchmark-based" approach (see Annex 1): one did not make it to the top of the (average-based) primary rankings (and was included in Group D);[11] the other indicator (i.e., the exercise of bullying) was discarded from Group B based on the experts' comments.

Discussion

The CARA criteria applied in our Delphi study allow experts to assess the list of indicators along several dimensions, providing a much richer base for consensus building and final decision-making. The CARA criteria include (1) key criteria for inclusion in the set of indicators, namely, which indicators are necessary and sufficient to monitor integration levels (Adequacy, Relevance); and (2) key criteria for exclusion from the set of indicators (namely, lack of Clarity and Accessibility). Nonetheless, the application of a traditional "benchmark-based consensus" combining all four criteria was insufficient for reaching the established goals of our study. The four criteria impose a more considerable variability and difficulty in reaching consensus on all four across the board. Rather than adjusting the benchmark in an *ad-hoc* fashion that is more vulnerable to bias and arbitrariness (Jünger et al., 2017), we opted for an alternative strategy combining inputs and methodological procedures (both quantitative and qualitative), which helps increase the robustness and validity of the selected indicators.

First, we considered rankings that included all the initial indicators, thus helping establish a baseline prioritization for selection. One ranking considered only Adequacy and Relevance since these determine the relative importance for inclusion in the final set of indicators. A second ranking also considered Clarity and Accessibility, which affected the strategic exclusion of a given indicator. The combination of the two CARA-based rankings provided an initial list of well-positioned indicators that met Adequacy and Relevance criteria for inclusion while not being overly dragged by Clarity and Accessibility issues (which could furthermore be improved for individual indicators using the experts' comments). This provided an overall more robust picture, considering all four CARA dimensions but in a more responsive manner. While the indicators at the top of these rankings largely overlap with the results of the more traditional "benchmark-based"

consensus approach, these rankings also helped ensure reaching the numerical targets since they included all initial indicators.

However, top overall scores might overlook particular issues in particular dimensions or from specific points of view, particularly those beyond Adequacy and Relevance criteria. The consideration of additional and demanding quality criteria, which required a significant part of the experts selecting that particular indicator among their top five or giving the highest scores in Relevance and Adequacy, helped double-check the robustness of the rankings. It also provided the basis for refining the ranking of indicators by fragmenting the (combined) ranking into groups of indicators based on *both* (1) position in the ranking and (2) the presence or not of these additional (consensus-based) quality criteria. This provided a refined prioritization list for consideration.

Finally, the evaluation of the experts' comments allowed for the nuanced identification of consensuses and remaining divergences and, most importantly, the consideration of high-qualified expertise for specific indicators. This implies moving beyond a mechanical logic of consensus and including an expertise-reflective approach to the definition of consensus in line with findings in the literature that consensus does *not* necessarily imply the 'correct' answer (Jünger et al., 2017). Since the comments pointed out positive qualities and severe limitations of each indicator, these were fully incorporated into the selection considerations: in the case of limitations, solutions were frequently provided in the comments involving improvements or changes in the indicators. However, when no solution was provided, the indicators were dropped (even if they were well-placed in the rankings).

It is important to notice that, numerically, the quantitative inputs of the alternative approach would have been sufficient to achieve the targeted selection. However, the contribution of the qualitative analysis is manifold. First, the comments and suggestions helped improve all indicators by, as a minimum, introducing additional clarifying information. In the case of up to 15 indicators (see Annex 1), this included further improvements, such as redefining groups of reference, including robustness checks, using higher-order indicators within existing alternatives, etc. All of these changes helped ensure and increase the indicators' robustness, validity, and usability. Finally, the comments helped us identify significant concerns, some of them with the potential for negative social impact, that required consideration. In this manner, we were able to refine the selection, fully reflecting the pool of knowledge brought together by the participating experts, while ensuring a sufficient representation of the key dimensions of this complex social process.

Based on these considerations, we propose this multi-input mixed-method strategy as an alternative to the traditional "benchmark-based" approach for the definition of consensus in Delphi consultations, specifically to select indicators. Incorporating these multiple inputs, we ensured not only the selection of a sufficient number of indicators (thirty plus five) but also their empirical soundness and sufficient representation of all relevant dimensions. The use of the CARA model enables the obtainment of standardized scores across four relevant criteria, and this is systematized through the building of the primary rankings. Considering multiple inputs (CARA scores, additional predefined quality indicators, and qualitative comments by experts) allows building a consensus based on the structured discussion, robust and systematized, but that incorporates an expertise-reflective approach for the refinement of selection. The qualitative analysis of comments reflects both the emerging negative and positive consensuses (and remaining divergences) and the specific high-qualified comments that complement the consensus approach. It is essential in this case that a good engagement from the experts is ensured, particularly in providing comments and engaging in the two-round dynamic.

Conclusion

Building cohesive societies in the framework of sustainable development require addressing the demographic challenge of migration from the ground in educational settings. Integration involves understanding the diversity of migrant and refugee children, and for that, new meaningful measures must be developed to capture the nuances of such a holistic and dynamic process. An innovative Delphi strategy has been presented to capture the strategic factors that improve social equality, quality education, and safeguard minors' rights. The methodology described tackles the classic methodological parameters that provide content validity to a new set of indicators. And it considers relevant qualitative inputs based on the experience of a multidisciplinary group of experts in a planned and structured manner that allows managing the complexity of holistic concepts. As a result, a final set of thirty-five indicators has been obtained to explore the state of migrant and refugee children in schools and plan future actions to improve it: twenty-one of the resulting indicators point to political, social, and educational factors enabling and hampering integration, while 14 indicators pointed to outcomes amongst the core components of the integration process. This methodology has revealed the applicability of qualitative strategies to provide empirical soundness to quantitative measures and to permeate society beyond social research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 research and innovation programme.

ORCID iDs

Eva Bajo Marcos https://orcid.org/0000-0002-0618-1805
Ángela Ordóñez-Carabaño https://orcid.org/0000-0001-7552-1300

Elena Rodríguez-Ventosa Herrera https://orcid.org/0000-0003-2441-9159
Inmaculada Serrano https://orcid.org/0000-0002-1451-290X

Supplemental Material

Supplemental material for this article is available online.

Notes

1. In the context of the IMMERSE project, a literature review and a broad consultation with children and relevant stakeholders (parents, educators, policymakers) to provide these 57 indicators. The results of the Delphi were then also ecologically validated by children and relevant stakeholders.
2. For the qualitative research, the Ethics Committee of the Comillas Pontifical University evaluated and approved the proposed methodology, and stakeholders were free to participate or decline the invitation. All of them explicitly provided consent through pen and paper consent forms after being broadly informed by the researchers of the methodology to be followed. In the case of underage participants, consent was provided by their parents, and the children themselves decided on their own participation through explicit assent.
3. As a reference, Boulkedid and colleagues found that the median of Delphi participants is 17 (Boulkedid et al., 2011).
4. As indicated by Steurer (2011), from the second round the positioning of the group is sufficiently consistent and the possibility of incorporating a third round and, therefore, increasing the involvement requested to the experts, should be weighed considering the dropout tendency.
5. As indicated by Foth et al. (2016) in their review of Delphi consensus methods, in 39.3% of the studies they reviewed, consensus was described as percentage of agreement and, usually, that agreement was stablished as 60% or higher.
6. The Migrant Integration Policy Index (MIPEX) is an assessment tool that provides indicator measures of integration policies across eight policy areas (labour market and mobility; family reunion; education; political participation; permanent residence; access to nationality; anti-discrimination; health) in fifty-six countries (all EU member states and other countries in Asia, North and South America, and Oceania).
7. Furthermore, the scores in these two dimensions were not set on stone, since: Clarity of some indicators could be improved with the qualitative suggestions of the experts and Accessibility of some could also be improved during the lifetime of the project by generating some of the data, or, beyond that, through policy recommendations.
8. First, as a matter of honesty among respondents; and second, as a matter of lack of clarity on how bullying is understood from the perspective of participants or potential perpetrators. Significant measurement issues exist for the experience of bullying, which is more commonly researched (and also included in the selection of indicators). The experts understood that these issues could only be exacerbated when measuring the exercise of bullying, leading to unequal self-assessments and answering dynamics, which might require further contextualization.
9. Using these data as an indicator risked a potential criminalizing and stigmatising effect on whichever categories (encompassing both or either migrant-background or native children) that might display a higher rate of positive answers, particularly in the absence of contextualizing information. Having into account that the traditional question on experience of bullying was already selected, it was decided to drop this more problematic indicator.
10. Some pointed out to the diverse situations and contexts that the indicator covers – from regional dialects to multiple national languages and additional foreign languages. Others pointed out that, in a majority of countries and regions, bilingual education is not available for an immense majority of the migrant-children population, making it a less relevant indicator.
11. Although these two indicators reached consensus in the traditional "benchmark-based" approach, they did so based on a relatively low percentage of experts (barely above 60%, in contrast with other indicators reaching consensus most frequently on percentages above 70%). Additionally, for one of them, the "benchmark-based" consensus was reached around relatively low positive values (3 in several criteria, in contrast with other indicators reaching consensus most frequently around top value 4).

References

Ager, A., & Strang, A. (2004). Indicators of integration: Final report. *Immigration and asylum social cohesion and civil renewal* (Vol. *28*). http://www.homeoffice.gov.uk/rds

Asis, M., Bilak, A., Carammia, M., Geddes, A., Ibrahim, Y. A. A., Iturralde, D., Orozco, M., Pizarro, J. M., Plaza, S., Simmons, J., & Singleton, A. (2018). *Informing the implementation of the global compact for migration (issue 10)*. IOM.

Bajo Marcos, E., Fernández, M., & Serrano, I. (2022). Happy to belong: Exploring the embeddedness of well-being in the integration of migrant and refugee minors. *Current Psychology*, *1*(1), 0123456789. https://doi.org/10.1007/s12144-022-03341-2

Booth, T., & Ainscow, M. (2002). *Index for inclusion: Developing learning and participation in schools*. https://www.eenet.org.uk/resources/docs/Index.English.pdf

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *Plos One*, *6*(6), Article e20476. https://doi.org/10.1371/JOURNAL.PONE.0020476

Brady, S. R. (2015). Utilizing and adapting the Delphi method for use in qualitative research. *International Journal of Qualitative Methods*, *14*(5), 1–5. https://doi.org/10.1177/1609406915621381

Council of the European Union (2010). *European ministerial conference on integration*. http://ir.obihiro.ac.jp/dspace/handle/10322/3933

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi Method to the use of experts. *Management Science*, *9*(3), 458–467. https://doi.org/10.1287/mnsc.9.3.458

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, *67*(4), 401–409. https://doi.org/10.1016/J.JCLINEPI.2013.12.002

European Commission/EACEA/Eurydice (2019). *Integrating students from migrant backgrounds into schools in Europe: National policies and measures*. Publications Office of the European Union. https://doi.org/10.2797/222073

European Migration Network. (2022). *Annual Report on Migration and Asylum 2021*. EMN. https://ec.europa.eu/migrant-integration/system/files/2022-06/EMN_Annual-report_Migration_report_final.pdf

Eurostat (2021a). *Migration and migrant population statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Migration_and_migrant_population_statistics#Migrant_population:_23_million_non-EU_citizens_living_in_the_EU_on_1_January_2020

Eurostat (2021b). *Asylum applicants considered to be unaccompanied minors - annual data*. https://ec.europa.eu/eurostat/databrowser/view/tps00194/default/table?lang=en

Eurostat (2021c). *Asylum statistics - statistics explained*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Asylum_statistics&oldid=526224#Age_and_gender_of_first-time_applicants

Foth, T., Efstathiou, N., Vanderspank-Wright, B., Ufholz, L.-A., Dütthorn, N., Zimansky, M., & Humphrey-Murto, S. (2016). The use of Delphi and nominal group technique in nursing education: A review. *International Journal of Nursing Studies*, *60*, 112–120. https://doi.org/10.1016/j.ijnurstu.2016.04.015

Heink, U., & Kowarik, I. (2010). What are indicators ? On the definition of indicators in ecology and environmental planning. *Ecological Indicators*, *10*(3), 584–593. https://doi.org/10.1016/j.ecolind.2009.09.009

Hernández Franco, V., García Suárez, I., Jabonero, M., De la Torre González, B., Hernández Izquierdo, L., Gomariz Moreno, M., Blanco, M. R., & Gonzalo Misol, I., & Universidad de Huelva & AIDIPE (2009). Proyecto Plutarco: Sistema de indicadores para la evaluación del área de intervención del plan estratégico de ciudadanía e integración 2007-2010. *Actas del XIV Congreso Nacional de Modelos de Investigación Educativo* (pp. 239–281). AIDIPE. https://www.researchgate.net/publication/260553356

Huddleston, T., Bilgili, O., Joki, A.-L., Vankova, Z., Bilgili, Ö., Joki, A.-L., & Vankova, Z. (2015). *Migrant integration policy Index 2015*. http://mipex.eu/sites/default/files/downloads/files/mipex-2015-book-a5.pdf

Huddleston, T., Niessen, J., & Dag Tjaden, J. (2013). *Using EU indicators of immigrant integration final report for directorate-general for home affairs using EU indicators of immigrant integration*. European Comission. https://ec.europa.eu/migrant-integration/sites/default/files/2013-08/docl_37216_243039941.pdf

Jünger, S., Payne, S. A., Brine, J., Radbruch, L., & Brearley, S. G. (2017). Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliative Medicine*, *31*(8), 684–706. https://doi.org/10.1177/0269216317690685

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology*, *70*(2), 144–156. https://doi.org/10.4097/kjae.2017.70.2.144

Linstone, H. A., & Turoff, M. (Eds.), (1975). *The Delphi method: Techniques and applications*. Addison-Wesley.

MiCreate. (2019). *Child centred approach across disciplines (Deliverable 2.6)*. MiCreate. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5c7dd0909&appId=PPGMS

Miller, B. (2013). When is consensus knowledge based? Distinguishing shared knowledge from mere agreement. *Synthese*, *190*(7), 1293–1316. https://doi.org/10.1007/s11229-012-0225-5

Miller, B. (2019). *The social epistemology of consensus and dissent* (pp. 230–239). The Routledge Handbook of Social Epistemology. https://doi.org/10.4324/9781315717937-23

OECD (2008). *Handbook on constructing composite indicators: Methodology and UserGuide*. OECD. https://doi.org/10.1787/9789264043466-en

OECD (2018). *The resilience of students with an immigrant background*. OECD Publishing. https://doi.org/10.1787/9789264292093-en

OECD/EU (2018). *Settling in 2018: Indicators of immigrant integration*. https://doi.org/10.1787/9789264307216-en

Steurer, J. (2011). The Delphi method: An efficient procedure to generate knowledge. *Skeletal Radiology*, *40*(8), 959–961. https://doi.org/10.1007/s00256-011-1145-z

Thoreau, C., & Liebig, T. (2018). *Settling in 2018: Indicators of immigrant integration*. Paris: OECD Publishing. https://www.oecd.org/publications/indicators-of-immigrant-integration-2018-9789264307216-en.htm

UNECE. (2019). *Guidance on data integration for measuring migration*. Geneva: UNECE. https://unece.org/info/Statistics/pub/21850.

UNICEF (2022). *Refugee and migrant children in Europe*. UNICEF. https://www.unicef.org/eca/emergencies/refugee-and-migrant-children-europe

van der Schaaf, M. F., & Stokking, K. M. (2011). Construct validation of content standards for teaching. *Scandinavian Journal of Educational Research*, *55*(3), 273–289. https://doi.org/10.1080/00313831.2011.576878

Van Vooren, E., & Lembrechts, S. (2021). Involving children and young people in policymaking: A children's rights-based approach to co-creative practice in REFLECTOR. In L. Van Praag (Ed.), *Co-creation in migration studies. The use of co-creative methods to study migrant integration across European societies* (pp. 247–278). Leuven University Press.

You, D., Lindt, N., Allen, R., Hansen, C., Beise, J., & Blume, S. (2020). Migrant and displaced children in the age of COVID-19: How the pandemic is impacting them and what can we do to help. *Migration Policy Practice*, *2*(2), 32–39.