Reflections on the application of the Delphi method: lessons from a case in public transport research

Hirschhorn, Fabio

**Citation (APA)**
Hirschhorn, F. (2018). Reflections on the application of the Delphi method: lessons from a case in public transport research. *International Journal of Social Research Methodology*. https://doi.org/10.1080/13645579.2018.1543841

**Important note**
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

# Reflections on the application of the Delphi method: lessons from a case in public transport research

Fabio Hirschhorn

Faculty of Technology Policy, Delft University of Technology, Delft, The Netherlands

**ABSTRACT**

The paper outlines the author's experience employing the Delphi method, using as an example a particular application of the Delphi in the field of public transport research. Attention is given to aspects such as the choice of method, selection of experts, design of questionnaires, interaction between survey coordinator and participants, and also the analysis of experts' responses. Some of the challenges encountered during the survey, the way they were dealt with, and risk mitigation strategies used by the Delphi coordinator are highlighted too. The primary objective of this paper is to offer insights that can support other researchers or practitioners preparing to apply the Delphi methodology. Furthermore, the article contributes to the methodological debate by reflecting on the introduction of novel practices that can help overcome some typical pitfalls of the Delphi: a dedicated blog supporting the survey, safety-net questions, and a constant-sum type question.

## 1. Introduction: the public transport question and purpose of this article

One of the main discussions in the field of governance of metropolitan public transport systems (PT) involves the relationship between organisational form and performance. The underlying assumption in these debates is that certain elements of the organisation of PT (such as market entry rules, contracting models, integrated fare policies etc.) can promote or hinder public goals attached to PT (such as sustainability, cost-efficiency, safety etc.) Analysts seek to shed light on the links between organisation and performance to discern potentially better strategies on how to organise PT services.

Chadwick (1859) for instance contrasts competition for the field – namely competition for having access to a market or area where to deliver PT services – and competition within the field – related to the competition between different transport providers operating in the same market. This same theme spurred much research after the British bus deregulation in 1986, and similar deregulation attempts elsewhere (Cowie, 2014; Fernández & Muñoz, 2007). The possible performance impacts of different awarding mechanisms or varied contractual regimes between government authorities and operating companies is also part of this scholarship (Stanley & Hensher, 2008; Vigren, 2016). Some authors examine the issue of ownership structure, both at the market level and at the level of operating companies, and debate possible performance implications (Docherty, Shaw, & Gather, 2004; Scheffler, Hartwig, & Malina, 2013). The role of key stakeholders, the arenas where they interact, and the tier of government responsible for PT policy are

also scrutinised as relevant features of the organisational set-up of PT that might influence performance (Pemberton, 2000).

These analyses offer important insights and help improve the understanding of some key mechanisms linking the governance of PT and performance. However, by only looking at the summation of effects of isolated policy interventions, they may fail to capture a more textured view of PT governance. PT is a complex multifarious sociotechnical system in which different technical elements, actors (with multiple interests), and norms coexist (de Bruijn & Herder, 2009; Schwanen, 2013). These systems are complex and more than the sum of their parts (Macmillen, 2013) and, as such, their analysis can benefit from a configurational perspective that acknowledges that several elements interact and influence each other (Ostrom, 2010; Ragin, 1987). In other words, analyses of the relationship organisation-performance in PT can benefit from the recognition of this systemic character, and of the importance of the interplay between different PT elements.

To operationalize such configurational approach, the first step is to identify and select adequate research variables. Therefore, the problem posed for PT research is to select performance indicators suitable to measure the achievement of strategic goals in PT as well as organisational features that might affect strategic outcomes.

Hirschhorn, Veeneman, and van de Velde (2018) selected a participatory approach, the Delphi method, to tackle this problem. The Delphi relies on a sequence of questionnaires distributed to selected experts in a process managed by a survey coordinator. After the first round of questions, and preceding any new questionnaire, the survey coordinator provides participants with anonymous feedback on answers offered by all panel members. Individual participants can reflect on this feedback and reconsider their opinions when responding to subsequent questionnaires. This process, interspersing questionnaires and controlled opinion feedback in a protected anonymous environment, constitutes a powerful mechanism to expose and articulate diverse views and to help creating knowledge and solving complex policy issues.

By developing a Global Delphi in Public Transport ('GDPT'), Hirschhorn et al. (2018) gathered insights from PT experts across the world. The GDPT produced authoritative inventories and ratings of core performance indicators and organisational features affecting performance in PT. In their paper, Hirschhorn et al. (2018) discuss the GDPT's results, as well as their possible implications for the study of PT under more textured and systemic lens. The current paper, instead, uses the GDPT as an example of the application of the Delphi method to make more generic observations related to the methodology. The paper looks at some of the challenges encountered during the survey, the way challenges were dealt with, some risk mitigation strategies, and lessons learned. The description of the GDPT points to assets and drawbacks of the Delphi, and thus can support researchers or practitioners wishing to apply the methodology. In particular, the GDPT introduces novel practices that can be applied in new Delphi studies to help overcome some typical pitfalls of the method, as well as to leverage some of the method's strengths: a dedicated blog supporting the survey, safety-net questions, and a constant-sum type question.

## 2. Research strategy

### 2.1. Choice of method

As advanced in Section 1, transport research many times fails to recognise the complex and systemic character of PT, and does not address the topic in a comprehensive manner. Macmillen observes that: 'As complex systems, however, modern transport systems are more than the sum of their parts. They cannot be understood as reducible to stable, established and deterministic relationships between variables...' (2013, p. 203). It is thus necessary to find new approaches to analyse PT governance that allow moving beyond an incremental perspective that only considers

isolated policies or variables. A configurational approach thus, appears as an alternative to tackle the question on how performance of PT may be affected by the organisation of the system by allowing an examination of the combined effects produced by multiple relevant factors. For this, however, a necessary first step is to identify these factors that can serve as adequate research variables.

The specialised literature on the debate on the performance repercussions of different features of the organisational form of PT, briefly illustrated above, constitutes a first relevant input to develop this task and define suitable organisational and performance factors to be analysed in combination. However, the GDPT's research aim could gain from also considering the views of other stakeholders beyond academia that also possess relevant knowledge in the field. Schwanen, Banister, and Anable (2011) make a similar observation, however looking at public transport research more broadly. The authors claim that the field has overall focused on a limited range of actors, leaving aside the role played by and the views from important players, such as financial institutions, insurance companies, social movements etc.

The Delphi method emerges as a possible tool to help addressing the research need just highlighted: identifying the critical factors in the discussion organisation-performance that would allow approaching PT governance with more systemic lens and, also, perform this task attending to views of other types of stakeholders beyond academia. Although as initially conceived the Delphi was not meant as an open and inclusive participatory process – and instead as a methodology to achieve consensus among a small and selected group of experts – over time new variants of the method appeared and opened up for inputs from more actors as well. Delphi techniques recognise and seek value in the articulation of varying and contrasting visions as a tool to support the solution of complex policy matters (Turoff, 1970). Kezar and Maxey (2016) corroborate this perception, and emphasise that the Delphi technique is particularly well suited to solve complex and multilayered problems that require the attention of multiple stakeholder groups.

Therefore, the Delphi was chosen, and the GDPT conceived, as a means to find suitable organisational features and appropriate indicators that can help guiding PT policy design and evaluation, but relying on insight beyond a simple literature review. The ambition to amplify the reach of PT research to consider a wider set of actors is understood and applied with caution though. In some cases (the GDPT for instance), not any and all stakeholder involved directly or indirectly with PT will have relevant knowledge for debating questions that are essentially technical and require in-depth expertise of the field. If asked to stakeholders not minimally familiarised with these issues, questions about if/how the adoption of competitive tendering to select bus operators might, as an organisational feature of PT, impact levels of cost-recovery, for instance, might lead to irrelevant responses. The GDPT integrates views of academics and practitioners from varied backgrounds and this search for variety of views will be made clear in the description of GDPT's process to identify and select participants (Section 3.2).

## 2.2. The Delphi

The Delphi method was developed within the RAND Corporation in the 1950s. Back then, it was devised as an organised participatory process for consensus building. By eliciting the opinions of experts, the Delphi method was shaped to build authoritative forecasts in relation to the occurrence of events or trends. Originally, the method was created and used for decision-making regarding military matters, and only years later it was disclosed to the general public (Dalkey & Helmer, 1963). Nowadays its use is widespread in a variety of domains, such as technology forecasting, engineering, and nursing sector, as well as in different social science fields (Gupta & Clarke, 1996; Landeta, 2006).

Procedure-wise, the Delphi relies on a sequence of questionnaires distributed to selected experts in a process managed by a survey coordinator. After the first round of questions, and

preceding any new questionnaire, the survey coordinator provides participants with anonymous feedback on answers offered by all panel members, ensuring that opinions are not assigned to particular individuals. Participants can reflect on this feedback and reconsider their opinions when responding to subsequent questionnaires. This process, interspersing questionnaires and controlled opinion feedback, continues until a desired level of consensus is reached among respondents or until opinions are stable across survey rounds (Dajani, Sincoff, & Talley, 1979; von der Gracht, 2012).

Rowe and Wright (1999) identify four core elements in a Delphi survey:

(a) Anonymity: providing opinions anonymously and free of direct interaction with other respondents should allow participants to express themselves freely, under no influence of potential dominant figures or group conflicts. Opinions and arguments can thus be evaluated on their merit only.

(b) Iteration: the multiple rounds in a Delphi allow participants to reassess their own judgements and, given the anonymity of the process, reconsider earlier responses.

(c) Controlled feedback: after each round participants are confronted with the group's opinions and encouraged to re-evaluate their own responses. This feedback is normally presented through statistics based on aggregated responses. It is also possible to include in this feedback some anonymised textual arguments offered by participants in support of certain opinions.

(d) Statistical aggregation of group responses: at the end of the survey, the group's opinion is taken as the statistic average (mean/median) of overall opinions of panellists in the final round.

While the use of the Delphi follows some central features, the method is flexible in its application and the researcher has the possibility to customise the process to the particular characteristics of the problem in discussion, or to his/her specific objectives. As a result, a number of variants of the method emerged and continue to be developed (Kezar & Maxey, 2016; Schmidt, 1997; Steinert, 2009; Turoff, 1970). These variants adapt the method in different ways, such as techniques to select participants, types of questions employed, tools used for the analysis of responses, and type of outcome sought (see, for instance, discussion in de Loë, Melnychuk, et al., 2016; Paré, Cameron, Poba-Nzaou, & Templier, 2013).

## 3. The Global Delphi in Public Transport (GDPT)

### 3.1. Survey structure

The GDPT was structured in three different stages: (i) brainstorming (respondents could freely propose all relevant elements in connection to the issues at stake), (ii) narrowing-down (respondents shortlisted most relevant elements from previous stage) and (iii) rating (respondents rated shortlisted elements). In each of these stages one questionnaire was used. This design was mainly inspired by the ranking-type Delphi, although it does not strictly follow the structure and steps proposed by Schmidt and others who have employed this variant (Paré et al., 2013; Schmidt, Lyytinen, Keil, & Cule, 2001). Differences are highlighted in the remainder of this section. Figure 1 summarises the GDPT structure and steps.

The GDPT was entirely carried out with online questionnaires using a survey platform to collect and aggregate responses. In addition to direct email interaction, a dedicated blog was created to support the GDPT. The blog served as a platform for the publication of GDPT's results, as well as to provide more detailed information on the survey's motivation and aims. By creating this separate communication channel, the GDPT managed to make additional information
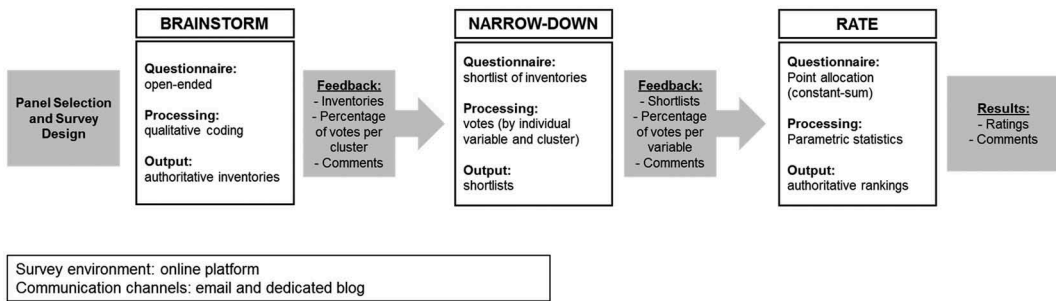
Figure 1. Schematic representation of the GDPT.

available for those participants interested in learning more about the survey, while avoiding very lengthy emails that could be overwhelming to participants with limited time availability.

## 3.2. Choice of experts

The choice of experts is a crucial step to ensure that breadth of knowledge is represented among panellists (Delbecq, Van de Ven, & Gustafson, 1975). Panel-building in a Delphi comprises two moments: (i) defining the relevant expertise and (ii) identifying individuals with the desired knowledge.

Concerning relevant expertise, the GDPT attempted to cover knowledge on (i) PT performance monitoring and/or evaluation and (ii) PT governance – particularly the design and functioning of PT organisational structures worldwide. Importantly, the GDPT recognises that this expertise is also outside academic debates and tries to gather the views of varied types of practitioners as well.

For the identification of experts, the GDPT combined two conventional approaches, namely sampling based on actor types and snowball sampling. The first approach seeks representativeness in terms of perspectives by sampling actors from diverse affiliations. In snowball sampling, the researcher starts off by picking a small number of stakeholders, and then asks them to recommend other potential participants.

Firstly, in terms of actor types, the following criteria were used to find academics and also practitioners that could contribute to the survey:

(a) Variety of roles: different types of stakeholders: (i) academics, (ii) government officials, (iii) transport operating companies; (iv) users' associations, (v) multilateral financing institutions, (vi) consultants. In the case of academics, two more aspects are considered: (i) works published in relevant journals and retrieved on Google Scholar, and (ii) variety of academic discipline in PT: geography; economics; engineering; public administration and policy; and urban planning.
(b) Knowledge in a variety of organisational settings: experts based in, and/or with expertise on different geographical locations.
(c) Prominence in the field: affiliation to eminent organisations, as well as involvement in major international fora, major universities, government entities responsible for PT, PT providers, participation in editorial boards of prominent journals.

A matrix for the identification of experts was developed to support this selection process (Table 1). After a first attempt to populate the matrix, some of the initially identified experts were asked to provide recommendations for other specialists that could participate (snowball sampling) and the suggested names that had not been identified previously were also included in the panel. All experts

**Table 1.** Matrix for the identification of experts.

|  |  | Africa | Oceania | USA and Canada | Latin America | Europe | Asia |
|---|---|---|---|---|---|---|---|
| Practitioners | Consultants | 3 | 2 | 3 | 2 | 5 | 2 |
|  | Multilaterals | 2 | – | – | 3 | – | 3 |
|  | Operators | 3 | 6 | 4 | 4 | 11 | 2 |
| Governments | Authorities | 5 | 4 | 5 | 4 | 9 | 3 |
| Passenger Association |  | – | – | – | – | 2 |  |
| Academics | Public administration | 1 | 2 | 2 | 2 | 6 | 1 |
|  | Engineering | 3 | 2 | 1 | 2 | 2 | 3 |
|  | Traveler behavior | 1 | 2 | 2 | 2 | 1 | 2 |
|  | Economics | 1 | 1 | 2 | 1 | 4 | 2 |
|  | Urban planning | 1 | 1 | 2 | 1 | 2 | 6 |

were then assessed more closely to confirm the direct relevance of their work for the topics being surveyed, i.e. that their work in PT was indeed connected to PT governance on its various dimensions.

The matrix illustrates the central objective of the GDPT to encompass a broader set of views on the debate organisation-performance. The survey includes an ample set of practitioners from several backgrounds. It also gives room to relevant actors that tend to be 'less heard', such as users' associations, multilateral financing institutions and private consultants. At the same time, the important views from academics – with varied technical and geographical experience – are still considered.

A minor imbalance in favour of European participation is also visible through the matrix. This was ultimately the result of the greater ease in identifying experts across contacts from personal networks of GDPT's authors or through these contacts' recommendations. Faced with this initial outcome in the construction of the matrix, a trade-off became evident: reduce the potential panel to avoid that the survey could be skewed towards a European view or stick to as many participants and views as possible as a way to prevent the risk of a very low rate of responses. Hirschhorn et al. (2018) chose the latter path, mainly because the association of experts to a region in the initial matrix took into account their regional base (in the case of the experts that work in multilateral funding institutions, the region indicates their current affiliation in terms of geographical area of work), which does not correspond necessarily to their regional knowledge. Most (if not all) the experts that were screened have knowledge on diverse PT systems and the fact that they are based in, or currently work with, a given region does not curtail the panel's overall geographical breadth of knowledge. This trade-off is nonetheless relevant because it is related to a potential pitfall of the Delphi method: the survey coordinator has no control over response rates, and there is a real risk of having just a few participants reacting, and/or a high rate of withdrawals throughout survey rounds, what can compromise the entire Delphi process.

The first communication with potential participants described the overall goals of the survey, indicated the planned number of questionnaires and the expected duration of the entire process. The link for the first questionnaire was also included. Due to the large number of invitees, the GDPT's authors chose not to approach and consult individuals' willingness to participate in advance to only after acceptance provide access to the questionnaire. This relates to another trade-off linked to the number of participants that one is willing to involve in the survey. The larger the set of experts, the harder it is to interact more directly with each of them. The GDPT's final participation numbers were very positive nonetheless, and response rate was high and drop-out numbers very low compared to other Delphi studies: 96 experts accessed the first online questionnaire. The first questionnaire was responded by 54 experts. Of these, 48 responded to the second questionnaire, and finally, 46 experts from 18 different countries concluded the third and last questionnaire. The profile of respondents shows the intended diversity, including approximately 60% practitioners and 40% academics, varied technical and regional knowledge, without any group being overrepresented (Figure 2).
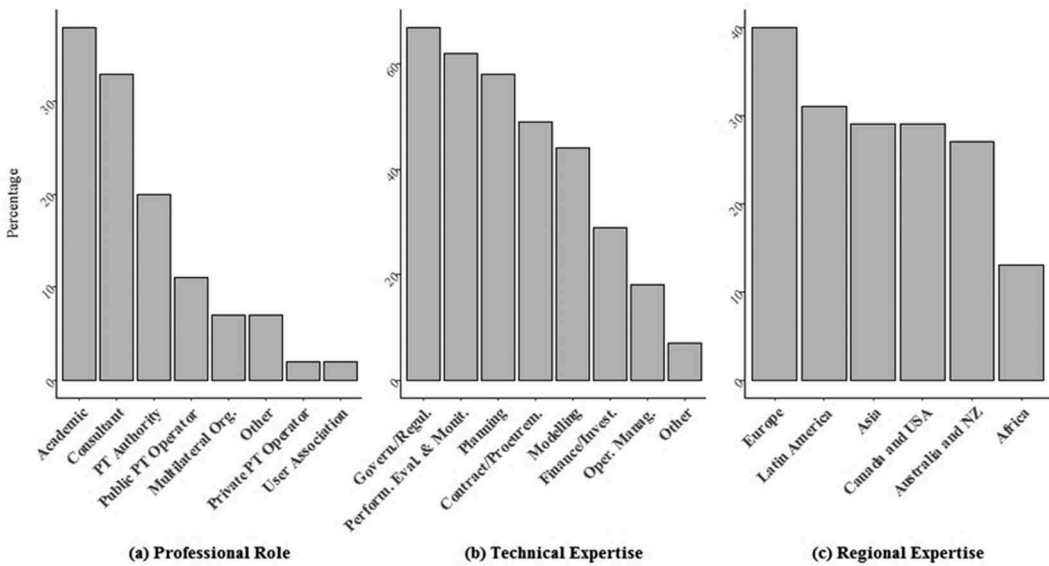
**Figure 2.** Profile of experts that concluded the GDPT (Hirschhorn et al., 2018). Permission granted by Elsevier. License Number: 4460780239521

## 3.3. Round 1: brainstorming

### 3.3.1. The questionnaire

The first questionnaire of the Delphi used two sets of questions. The first set contained open-ended questions and the second contained Likert-scale questions.

In open-ended questions, experts were presented with a decision situation and asked to assume they would be hiring a consultant to analyse a PT system. They had to list and briefly describe (in one or two sentences) at least five performance indicators that they would want examined by the consultant. Experts were instructed to prioritise indicators better able to provide relevant information on what, in their views, would be critical aspects of PT. Experts were also asked to associate each selected indicator to a broad performance aim. The question did not set any limit to the number of indicators that could be suggested, and also avoided defining any specific performance goal to be measured. Similarly, a second question asked experts to list and briefly describe (in one or two sentences) at least five organisational features considered important drivers of PT performance.

The second set of questions in the brainstorming exercise presented experts firstly with some 'pairs' of performance aims and indicators normally found in PT literature analysis (e.g. Environmental Sustainability; per capita emissions of NOx) and secondly a list of organisational features normally discussed in PT literature (e.g. Allocation of Ownership of Long-Life Assets). The performance pairs and the features in the list had to be rated using a Likert-scale from 1 to 5. The Likert-scale questions were deliberately included as the second set of questions to avoid introducing any bias before open-ended questions. They were only visible to participants in a second page of the questionnaire (accessible only after completion of the first set of questions).

The Likert-scale questions were introduced in the survey as a risk mitigation strategy: in case the answers to the open-ended questions were not really suited to the research goal (due to misinterpretations for instance) and had to be discarded (a risk in Delphi), the responses to the rating questions could be used to inform the following rounds of the

survey. As such, the Likert questions offered a way to reduce possible loss of expert input. This risk did not materialise in the GDPT, and the open-ended questions generated nearly 700 lines of content fitting the research aim. Responses to the Likert-scale question did not have to be used in the remaining of the survey.

### 3.3.2. Analysis of responses and feedback material

Answers to open-ended questions were qualitatively analysed and coded. The coding identified major themes, eliminated redundancies and produced long inventories of performance indicators and organisational features based on expert opinion. Sensitising concepts (Bowen, 2006) based on relevant PT literature were used as guidelines for the interpretation and organisation of input received from experts, and no supporting software was employed although resorting to these tools is also possible (Bailey et al., 2012; de Loë, Murray, & Brisbois, 2016). The three authors of the GDPT worked in parallel in the coding task: the main author of the GDPT performed an initial coding of the entire set of answers. As a reliability test, the co-authors independently coded randomly selected samples of the answers received. The three independent analyses produced consistent conclusions, with minor exceptions. These differences were discussed and the main author reconciled them to generate the final output of this round. This triangulation procedure is similar to the one proposed by Schmidt et al. (2001).

The coding process was the most challenging and time-consuming step of the entire GDPT. It involved a constant trade-off between two conflicting tasks: consolidation, on one hand – to produce a reasonably-sized list to be used in following rounds by experts (that have limited time available for participation) – and, on the other hand, no excessive generalisation of answers – that could defeat the purpose of the Delphi by impoverishing experts' inputs. Having asked experts to include brief descriptions of their answers, a measure proposed by Schmidt (1997) and Okoli (2004), was essential for this step. It helped clarifying opinions and enriching the material.

Finally, two inventories were produced listing core performance indicators and core organisational features. The feedback material sent to experts after this stage included

Table 2. Sample from inventory of performance indicators (Hirschhorn et al., 2018).

| ACCESSIBILITY IMPACTS | | |
|---|---|---|
| Cluster | Mentioned by | Indicators |
| Access to Destinations | 9% | a) Number of opportunities and services that can be reached by public transport within a given time or distance. |
| Access to PT | 33% | a) Average walking time or distance to access selected routes. b) Percentage of inhabitants (or users) who live within walking distance of frequent transport service. c) Number of stations or bus stops per square km. d) Distance between PT stops. |
| Fairness and Affordability | 28% | a) 'Access to Destinations' measure for low income population. b) 'Access to PT' measure for the bottom 40% *'increase social inclusion and reduce inequality.'* c) Percentage of income or household budget (of low income or lowest quartile) spent on transport. d) Percentage of immobile. e) Percentage of poor served by subsidies. f) Average fare per passenger km. g) Average fare relative to petrol costs for medium-size car for short, medium and long trips (to be defined). |
| Universal Design | 9% | a) Percentage of stations/stops or terminals with facilities such as on level boarding/escalators and lifts. |

these two inventories, indicating also the percentage of respondents that mentioned elements in each cluster, and also all additional comments made by experts. Table 2 exemplifies the coding structure with a sample from the inventory of performance indicators from Hirschhorn et al. (2018).

### 3.4. Round 2: narrowing-down

#### 3.4.1. The questionnaire
In Round 2, experts were asked to shortlist seven performance indicators and seven organisational features among all those listed in the inventories produced in Round 1. The wording used in this questionnaire was consistent with the previous round, and requested experts to prioritise indicators better able to provide insights on most critical aspects of PT and organisational features with greater impact on performance.

Establishing a limit number for the shortlist items was a way to emphasise the need for prioritisation (also in view of the very long inventories from round 1), and to reduce the workload of participants (another pre-emptive measure against possible experts fatigue). The suggested number of seven items was chosen because it was the average number of performance indicators listed by each respondent in Round 1. Literature does not offer a recommendation in relation to what should be this limit number, and there are authors that suggest that no limit should be established (Schmidt, 1997).

#### 3.4.2. Analysis of responses
In similar Delphi designs authors suggest building the final shortlist based on simple majority of votes (Schmidt, 1997). The GDPT followed this criterion, however after a two-fold counting process. A first analysis considered votes at the level of individual variables, i.e. votes that each performance indicator and each organisational feature received. In addition, a second examination was done for cross-checking results and responses were counted at the cluster level, i.e. the sum of votes given to each of the clusters (those clusters defined in the coding of round 1 as illustrated in Table 1). For instance, the performance cluster 'Total (and operating) costs ratios' defined in round 1 comprised nine different indicators, so all votes given to these nine indicators were added to assess the total votes of the respective cluster.

The analysis at the cluster level revealed that defining a final shortlist of 7 items for the subsequent round of the GDPT would not adequately reflect experts' priorities. The cluster 'Total (and operating) costs ratios' is again an example: overall, it was the third most voted cluster in the inventory of performance indicators, but because these votes were dispersed among the nine alternative individual indicators, none of these nine would, individually, be in the final shortlist of 7. Therefore, whilst clearly relevant to experts, no cost ratio would have made the cut. To circumvent this possible shortcoming, Hirschhorn et al. (2018) expanded the final shortlists, each to include the 10 most voted variables.

Four experts voted for more options than requested in the questionnaire and thus they could skew the results. This is a potential limitation of Delphi techniques because experts might not fit their responses to the format defined by the researcher. The survey coordinator must be flexible to identify the issue and act to find a solution 'on the go'. The GDPT's coordinator reached out to these four experts to consult them about the possibility of repeating the exercise selecting only the stipulated number of items. Experts that could not repeat the exercise had their votes considered based on weighted values.

The feedback material sent to respondents described the two shortlists including the ratio of votes each variable received and also any comments from experts.

### 3.5. Round 3: rating (constant-sum)

### 3.5.1. The questionnaire

The final questionnaire of the GDPT introduced a constant-sum (or fixed-sum) question. Experts were presented with a practical decision situation: they were asked to consider the elements shortlisted in Round 2 as possible variables to be employed in a comparative study of PT in 15 metropolitan areas and that would be aimed to unveil how different ways of organising PT influence performance. Experts had to allocate 100 points among (all or part of) the variables in each shortlist to indicate the relevance of the variables for the hypothetical study. To avoid inducing any bias, the online platform where the survey was conducted listed variables in random order across the questionnaires sent to participants. Furthermore, the platform would not allow the exercise to be concluded unless exactly 100 points had been allocated.

### 3.5.2. Analysis of responses

In Round 3, answers were compiled and measured in multiple ways: (i) the average points received by each variable, (ii) the standard deviation in points received by each variable, (iii) the highest single score attributed to each variable, (iv) the percentage of experts attributing zero point to a variable and (v) the ranking of variables based on the amount of points they received. Results also included information on how preferences moved between rounds, by indicating the ranking of variables (based on votes) in Round 2. Table 3 presents a sample of the results of this round.

Although very common as a survey method and of simple execution, the constant-sum format has not been employed as part of a Delphi survey to the best of the author's knowledge. The main advantage envisaged in adopting this format was the possibility to analyse results with simple parametric statistics (average points, standard deviation etc.). If a simple ranking question were to be used instead, in which experts were asked to order variables according to their relevance from 1 to 10, the range of analytical tools available for the examination of responses would be more restricted: most literature on Delphi agrees that only nonparametric statistics can be used to measure level of consensus in these cases (e.g. de Loë, 1995; Schmidt, 1997; but see Norman (2010) for a different opinion). The constant-sum question prevents this possible drawback while still allowing rankings to be built based on the amount of points allocated to each variable. Furthermore, it also allows analysing the level of consensus or divergence between experts through the measures of standard deviation, maximum single score or zero scores attributed to each item.

**Table 3.** Sample of results round 3 (Hirschhorn et al., 2018).

| Performance Indicators | Avg. Points | Std. Dev. | High | Mode | Zeros | Rank | Rank 2 |
|---|---|---|---|---|---|---|---|
| User Satisfaction (overall index) | 15.91 | 11.03 | 50 | 15 | 11% | 1 | 3 |
| Cost-Recovery Ratio | 15.24 | 9.07 | 30 | 20 | 14% | 2 | 1 |
| Modal Split | 13.20 | 9.68 | 40 | 20 | 20% | 3 | 2 |
| % of Inhabitants (or users) living within walking distance to Frequent PT Service | 9.78 | 7.71 | 30 | 10 | 23% | 4 | 6 |
| Ridership per Capita | 9.57 | 8.70 | 30 | 5 | 25% | 5 | 7 |
| Ratio between Travel time in PT and Car | 8.22 | 6.36 | 25 | 10 | 25% | 6 | 5 |
| On-time Performance according to Timetable | 8.04 | 6.95 | 20 | 0 | 32% | 7 | 4 |
| Total revenue and total cost | 7.28 | 8.39 | 30 | 0 | 45% | 8 | 9 |
| Cost per Passenger Km | 6.98 | 6.77 | 26 | 0 | 36% | 9 | 8 |
| Comfort | 5.78 | 5.88 | 20 | 0 | 41% | 10 | 10 |

## 4. The GDPT's findings in brief

The results of the GDPT highlight a couple of broader central elements in the discussion organisation-performance in PT. In relation to performance, answers manifest a preference for a high-level system-wide assessment of PT systems, using multipart indicators – overall user satisfaction, cost-recovery and modal split. In the GDPT's final results these three core metrics are separated from the rest by a clear gap in points. Concerning features of organisational form driving performance outcomes, integration emerged as the central dimension: policy integration between public transport and other sectors, single integrated planning authority, as well as ticket and fare integration were highly rated by Delphi experts.

In addition to these broad conclusions, the GDPT's results also provide a 'menu' of core performance indicators and organisational features. These 20 elements (10 and 10, respectively) define what, in the view of the consulted experts, are the most important performance metrics suitable to measure strategic PT outcomes and what organisational features might affect strategic outcomes. This output, more broadly discussed in Hirschhorn et al. (2018), enables proceeding with future research efforts on the dependencies and interactions among PT organisational and performance elements – the need emphasised in Section 1.

## 5. Impressions and lessons learned

The experience with the GDPT highlights some important strengths of the Delphi method. It first confirms that the Delphi can be a powerful research tool to increase access to the valuable, and many times difficult to reach, opinion of experts. The input produced in the GDPT comes from a wide variety of technical and geographic perspectives that other research methods would hardly be able to gather. Additionally, the international reach of the survey, facilitated by the use of online questionnaires, is also a positive asset of the Delphi: engaging experts globally is extremely hard if conventional face-to-face interviews are to be used for instance. As a result, and representing another of the method's advantages, the Delphi can produce a breadth of views that makes it almost unparalleled as a building block for continued and more in-depth analysis – for example based on workshops, interviews or case studies for instance (de Loë, 1995; Van Dijk, 1990).

Also noteworthy is that the Delphi allows tailoring the survey according to the researcher's needs. This adaptability appears as one of its greater strengths, evidenced by the continuous and increased use the method witnessed since its inception (Gupta & Clarke, 1996; Landeta, 2006). Although the GDPT is mostly inspired by the ranking-type Delphi, it includes a series of changes to the original design of this type of survey.

The first important change introduced in the GDPT refers to the use of a dedicated blog as a repository of information where survey's details and updates were posted. Having this additional channel was helpful to keep participants engaged. Those participants interested in learning more about the survey had this information available in a separate 'venue', while the core communication for the GDPT process (questionnaires and feedback) could be conveyed in objective and simple emails.

Another novelty of the GDPT was the use of the safety-net questions in round 1. Open-ended questions in Delphi may lead to a wide range of responses, varying substantially both in format and content. If some responses do not fit the purpose of the survey, the coordinator might be forced to discard substantial amount of material. This is not only frustrating given the difficulty in obtaining such type of qualified input, but also jeopardises the continuity of the survey. The input received with the Likert-type questions in round 1 of the GDPT was a guarantee that the survey would not have to be dropped in case of problems with the open-ended questions.

Finally, the point allocation methodology in round 3 is one further new practice developed in the GDPT. This was particularly relevant to enhance the analytical tools available to assess experts'

responses. Simple but powerful parametric statistics are transparent, easy to understand and to replicate and, at the same time, provide multiple perspectives through which answers can be analysed: ranking based on average points, consensus or dissent based on standard deviation, maximum single scores and zero scores etc.

On the other hand, some limitations are inevitable in any research effort. Some difficult trade-offs are involved in developing a Delphi. Selecting experts for the panel requires choosing between engaging either a large and more diverse set of respondents, making close interaction a more difficult task for the Delphi coordinator or, alternatively, a smaller group of individuals that may be easier to follow closely and contact, possibly creating higher commitment by respondents. While the first route was chosen for the GDPT, a substantial effort was also made to keep participants engaged (e.g. with the use of the dedicated blog). Moreover, communication with experts had to ensure adequate provision of information while avoiding unnecessary long messages or questionnaires that could discourage participation of experts with limited time to contribute. Finally, coding in round 1, the most time consuming step of the GDPT, involved two conflicting tasks: consolidation of responses for the feedback material in a clear and concise manner but, at the same time, preventing excessive generalisation of answers – that would defeat the purpose of the Delphi.

These trade-offs underscore the crucial role played by the survey coordinator in articulating opinions and structuring the dialogue between experts, indicating a possible source of weakness in the Delphi. This process is not immune to subjective judgements and this is particularly relevant in the qualitative coding process (thus the importance of using a reliability test as a mitigation measure). On this same note, the praised adaptability of the Delphi may come at a cost: many times authors do not take into account minimum study design and reporting expectations in Delphi studies, producing less rigorous research (Brown, 2007; de Loë, Melnychuk, Murray, 2016).

In addition, it is not possible to ensure that the expert selection procedure employed in the GDPT guarantees that all relevant individuals were included in the panel, or that the final list did not suffer from any bias. This may be the case for different reasons: databases used may be incomplete, experts may not always publish their work, conferences, journals and other major fora may not be geographically pluralistic, experts may not have been recommended due to personal reasons etc. However, the criteria and procedure followed are consistent with best practices in Delphi studies (Delbecq et al., 1975; Marchau & van de Linde, 2016).

Finally, difficulties with the questionnaires or simply respondent fatigue may have prevented further participation in the GDPT. If the Delphi process appears too complex or time-consuming experts may not join or may later drop-out during the survey (Brown, 2007). One possible source of difficulty in the GDPT was language: questionnaires were prepared in English, which may have shied away non-native speakers. This route was chosen though to ensure that all experts were answering to the exact same questionnaires. Professional jargon may also be a source of mis-interpretations, both by the Delphi participants and by the survey coordinator.

## 6. Conclusion

This paper set-off to describe a particular application of the Delphi method – the GDPT – and, based on this experience, reflect on the method's strengths and weaknesses. The Delphi proved its merits by enabling the GDPT to move beyond some of the limitations that are common to participatory approaches and conventional face-to-face interviews: it offers participants anonymity, which allows free expression of opinions, and, enabled by ICT technologies, it also permits the consultation of experts from different parts of the world acting in different roles, circumventing the impossibility of gathering all participants in a single place at the same time.

Furthermore, by resorting to the Delphi, the GDPT authors were able to engage with actors that are not often part of public transport research more broadly, or of the specific debate on the relationship organisation-performance. The survey proved effective in bringing together

and articulating diverse views from professionals of different affiliations, as well as from varied technical and geographic backgrounds (Figure 2). As a result, the GDPT promoted a qualitative exercise – by design with no statistically significant results – with an output that could hardly be achieved differently. The expert insight provided by the Delphi can be a powerful component of a mixed-method research design if findings are complemented with more in-depth research.

Tough trade-offs are faced by the Delphi coordinator while preparing and conducting the survey, such as deciding the number of experts to approach and the way to engage them to attract and maintain participation, dealing with experts' responses so that the Delphi the survey is not compromised and overall objectives are achieved, and balancing information conciseness and richness in the qualitative coding analysis. These trade-offs highlight the critical role played by the survey coordinator, which is a possible source of weakness of the Delphi process. While high-lighting these challenges though, the paper also provided suggestions on how to tackle them. Thanks to the adaptability of the Delphi, the GDPT introduced positive practices that not only suited its particular research objectives, but also expand the Delphi's toolkit and thus can be used in future applications. The dedicated blog, the safety-net questions, and the constant-sum question offer researchers ways to circumvent or mitigate limitations of the Delphi, while still taking advantage of the method's potential.

## Disclosure statement

No potential conflict of interest was reported by the author.

## Notes on contributor

*Fabio Hirschhorn* is a PhD candidate at Delft University of Technology in the Netherlands. Fabio's research focuses on urban mobility, in particular the organisation and governance of public transport in metropolitan areas. Prior to the PhD, Fabio worked at the World Bank in Washington DC, as part of the Urban Transport team for Latin America and as a corporate lawyer in Brazil. Fabio holds a degree in Law (LLB) from the University of São Paulo and a Masters Degree (MA) in Economics and International Relations from the School of Advanced International Studies (SAIS) at Johns Hopkins University.

## References

Bailey, R., Longhurst, J. W. S., Hayes, E. T., Hudson, L., Ragnarsdottir, K. V., & Thumim, J. (2012). Exploring a city's potential low carbon futures using Delphi methods: Some preliminary findings. *Journal of Environmental Planning and Management*, 55(8), 1022–1046.

Bowen, G. A. (2006). Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods*, 5(3), 12–23.

Brown, C. A. (2007). The opt-in/opt-out feature in a multi-stage delphi method study. *International Journal of Social Research Methodology*, 10(2), 135–144.

Chadwick, E. (1859). Results of different principles of legislation and administration in Europe; of competition for the field, as Compared with competition within the field, of service. *Journal of the Statistical Society of London*, 22(3), 381–420.

Cowie, J. (2014). Performance, profit and consumer sovereignty in the English deregulated bus market. *Research in Transportation Economics*, 48, 255–262.

Dajani, J. S., Sincoff, M. Z., & Talley, W. K. (1979). Stability and agreement criteria for the termination of Delphi studies. *Technological Forecasting and Social Change*, 13(1), 83–90.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3), 458–467.

de Bruijn, H., & Herder, P. M. (2009). System and actor perspectives on sociotechnical systems. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, 39(5), 981–992.

de Loë, R. C. (1995). Exploring complex policy questions using the policy Delphi. A multi-round, interactive survey method. *Applied Geography*, 15(1), 53–68.

de Loë, R. C., Melnychuk, N., Murray, D., & Plummer, R. (2016). Advancing the state of policy Delphi practice: A systematic review evaluating methodological evolution, innovation, and opportunities. *Technological Forecasting and Social Change*, 104, 78–88.

de Loë, R. C., Murray, D., & Brisbois, M. C. (2016). Perspectives of natural resource sector firms on collaborative approaches to governance for water. *Journal of Cleaner Production*, 135, 1117–1128.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning: A guide to nominal group and delphi processes*. Scott: Foresman and Company.

Docherty, I., Shaw, J., & Gather, M. (2004). State intervention in contemporary transport. *Journal of Transport Geography*, 12, 257–264.

Fernández, J. E., & Muñoz, J. C. (2007). Privatisation and deregulation of urban bus services: An analysis of fare evolution mechanisms. *Journal of Transport Economics and Policy, 41*(1), 25-49.

Gupta, U. G., & Clarke, R. E. (1996). Theory and applications of the Delphi technique: A bibliography (1975–1994). *Technological Forecasting and Social Change*, 53(2), 185–211.

Hirschhorn, F., Veeneman, W., & van de Velde, D. M. (2018). Inventory and rating of performance indicators and organisational features in metropolitan public transport: A worldwide Delphi survey. *Research in Transportation Economics*. doi:10.1016/j.retrec.2018.02.003

Kezar, A., & Maxey, D. (2016). The Delphi technique: An untapped approach of participatory research. *International Journal of Social Research Methodology*, 19(2), 143–160.

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, 73, 467–482.

Macmillen, J. (2013). Mobility as a complex system: Key elements and interactions. In M. Givoni & D. Banister (Eds.), *Moving towards low carbon mobility* (pp. 190–205). Northampton, MA, USA: Edward Elgar Publishing.

Marchau, V. A. W. J., & van de Linde, E. (2016). The Delphi method. In P. van der Duin (Ed.), *Foresight in organizations: Methods and tools* (pp. 59–79). New York, NY and Milton Park, Abingdon, Oxon: Routledge.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5), 625–632.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, 42, 15–29.

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *The American Economic Review*, 100(3), 641–672.

Paré, G., Cameron, A. F., Poba-Nzaou, P., & Templier, M. (2013). A systematic assessment of rigor in information systems ranking-type Delphi studies. *Information and Management*, 50(5), 207–217.

Pemberton, S. (2000). Institutional governance, scale and transport policy – lessons from Tyne and Wear. *Journal of Transport Geography*, 8(4), 295–308.

Ragin, C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkley: University of California Press.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, 15(4), 353–375.

Scheffler, R., Hartwig, K. H., & Malina, R. (2013). The effects of ownership structure, competition, and cross-subsidisation on the efficiency of public bus transport: Empirical evidence from Germany. *Journal of Transport Economics and Policy*, 47(Part 3), 371–386.

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences*, 28(3), 763–774.

Schmidt, R. C., Lyytinen, K., Keil, M., & Cule, P. (2001). Identifying software project risks: An international Delphi study. *Journal of Management Information Systems*.

Schwanen, T. (2013). Sociotechnical transition in the transport system. In M. Givoni & D. Banister (Eds.), *Moving towards low carbon mobility* (pp. 231–254). Northampton, MA, USA: Edward Elgar Publishing.

Schwanen, T., Banister, D., & Anable, J. (2011). Scientific research about climate change mitigation in transport: A critical review. *Transportation Research Part A: Policy and Practice*, 45(10), 993–1006.

Stanley, J., & Hensher, D. A. (2008). Delivering trusting partnerships for route bus services: A Melbourne case study. *Transportation Research Part A: Policy and Practice*, 42(10), 1295–1301.

Steinert, M. (2009). A dissensus based online Delphi approach: An explorative research tool. *Technological Forecasting and Social Change*, 76(3), 291–300.

Turoff, M. (1970). The design of a policy Delphi. *Technological Forecasting and Social Change*, 2(2), 149–171.

Van Dijk, J. A. (1990). Delphi questionnaires versus individual and group interviews. *Technological Forecasting and Social Change*, 37(3), 293–304.

Vigren, A. (2016). Cost efficiency in Swedish public transport. *Research in Transportation Economics*, 59, 123–132.

von der Gracht, H. (2012). Consensus measurement in Delphi studies. Review and implications for future quality assurance. *Technological Forecasting and Social Change*, 79(8), 1525–1536.