PAPER TITLE

Dynamic argumentative Delphi: Lessons learned from two large-scale foresight exercises

Radu Gheorghiu, Institute for World Economy, Romania

Liviu Andreescu,* Faculty of Administration and Business, University of Bucharest, Romania, andreescul@gmail.com

Adrian Curaj, Politehnica University, Bucharest, Romania

Abstract

The Delphi method has been traditionally employed to achieve a degree of consensus within a relatively narrow group of experts on a usually divisive question or small number of questions. Given the typical features of a Delphi – the anonymity of participants, their indirect interaction mediated by a team processing the individual input a.s.o. –, such exercises were not usually designed to engender a community of experts; after the exercise is completed, the group at its core disbands. The advent of online Delphi in the 2000s brought a number of important changes to the Delphi format, among them a larger number of participants, a reduction in the number of rounds, instant access to response statistics etc. However, in light of the expanded participation and the facility of participants' engagement with the process, we argue that online Delphi can also serve to create a broad community of practice, to be regularly consulted on strategic and policy issues. This is particularly the case if the online Delphi preserves its strong qualitative / argumentative component, thus supporting the quantitative dimension with an intersubjectivity-enhancing mechanism. We illustrate this opportunity – as well as other benefits of the online Delphi format – with a series of three, broadly participative exercises on R&D&I, higher education, and public administration, respectively, carried out in Romania between 2011 and 2014.

Keywords: Delphi Method, online Delphi, computer-based Delphi, roundless Delphi, foresight

Introduction

Computer-based Delphi has a notable pedigree. It was adopted relatively early (Turoff 1972) in an attempt to mitigate some of the drawbacks or inconveniences of the Delphi process. These implied, among others, a cumbersome manual collection and processing of data, lengthy response times (due to, for instance, mailed questionnaires), or the resulting limitations in extending the circle of participants. Nevertheless, generally speaking computer-based Delphi exercises have followed rather traditional, tried-and-tested templates – although the diversity and innovativeness of such templates should not be underestimated (Turoff & Hiltz 1996).

With online Delphi, a considerably more recent development – Gordon and Pease published their important article on the topic in 2006 –, more liberties have been taken with the Delphi format. This time, the electronic medium was employed to enable synchronous participation, dispense with multiple rounds, ensure even quicker completion, or make statistics and tutorials available instantly (Gordon & Pease 2006, Gnatzy et al. 2011).

In this paper, we discuss two online Delphi exercises designed to expand on the opportunities provided by the web. We will refrain from introducing the context of these large-scale Delphi consultations in Romania (systemic foresight exercises on the future of higher education and, respectively, a national strategy for research and innovation) for reasons of space. Rather, we focus on the structure of the exercises, we discuss the relevance of the argumentative dimension, as well as the practicalities and a few statistics on the implementation process and the results.

Consensus in Delphi

In essence, the Delphi is a technique designed to approach consensus while minimizing conformity. Ever since it was first introduced in an experiment at Rand Corp. (Dalkey & Helmer 1963), a key goal was to “further reduc[e] the influence of certain psychological factors, such as specious persuasion, the unwillingness to abandon publicly expressed opinions, and the bandwagon effect of majority opinion” (Helmer & Rescher 1959, p.47). As the Delphi method expanded to cover a growing type of problems, the consensus-making objective was occasionally abandoned in some special Delphi designs (such as the Policy Delphi [Linstone & Turoff 2002]). Nevertheless, reaching some form of agreement among experts remains a typical goal of Delphi exercises, and studies have shown consistently that participants’ opinions tend not only to converge, but to do so towards more moderate positions (Spinelli 1983). In their literature review of Delphi studies, Rowe and Wright (1999, p.363) went so far as to claim that “the phenomenon of increased ‘consensus’, per se, no longer appears to be an issue of experimental interest.”

This is not to say, however, that consensus-making in Delphi does not raise a variety of problems, some theoretical, and some more concrete. As it has been noted, it remains hard to practically distinguish consensus from conformity, as both are measured in essentially the same way, as a reduction in the variance of responses after several rounds (Rowe & Wright 1999). The bandwagon effect cannot be completely eliminated, while sometimes the experts holding the more extreme positions give up and drop out of the exercise altogether. Furthermore, the agreement reached or approached in a Delphi may be a “specious consensus” (Strauss & Zeigler 1975) rather than an agreement converging on the ‘best judgment’. Sometimes, consensus on the quantitative variable (say, an estimate or a prediction) does not necessarily reflect consensus on the qualitative dimension (the reasons offered in support of the estimate).

Last but not least, how to best measure consensus among Delphi participants remains a problem which has received a variety of solutions and for which there is, as yet, no gold standard (Rayens & Hahn 2000).

One of the most common practical difficulties involves the mobilization of participants throughout the typical 3-5 rounds of the exercise. As the number of experts increases, defections will often multiply as well. Participation is also limited by the manual processing of expert responses (clustering, cleaning, categorizing, summarizing etc.). As noted previously, the process can also be very time-consuming due to its iterative, rounds-based nature and to the effort of summarizing positions and arguments.

The online Delphi format presented below was developed with these practical difficulties in mind. It is meant to enable synchronous or asynchronous participation in the exercise, increase the number of participants without large attrition penalties, ensure comparatively shorter completion and processing times, and provide other conveniences, such as instant statistics and monitoring (Gordon & Pease 2006; Gnatzy et al. 2011). In a more radical departure from the classical formats, online Delphi sometimes also reduces the number of rounds, occasionally to a single one, taking advantage of the fact that participants’ input can be instantly made visible to the other participants.

While an improvement in the respects mentioned above, the online Delphi may, nevertheless, also intensify some of the challenges listed previously. Without the carefully executed feedback provided by the organizers to the participants after each round, it becomes more difficult to identify or track the reasoning that supports the experts in changing their original opinions. If qualitative arguments are introduced in the questionnaire sheet to compensate for this potential pitfall, there is a danger of information overload, especially with a large number of experts.

In response to these challenges as well as to the opportunities provided by the online format, a new type of Delphi-like instrument was developed and tested in Romania in the summer of 2010 and then, in a somewhat different form, in 2013. One innovation in this roundless Delphi is that, while it invites experts to back their quantitative estimates with qualitative arguments which become instantly visible to the other respondents, it ranks the arguments in real-time according to the number of 'votes' they have gathered up to that point. This technique is used as a means of controlling the potential explosion in qualitative data. Thus, the Delphi dispenses not only with rounds, but also with the typical controlled feedback (as opposed to, e.g., the design presented by Gnatzy et al. [2011, p.1686]).

Methodological approach

One of the uses of Delphi occasionally mentioned in the literature is not so much output- as process-related: enabling structured communication among participants (Linstone & Turoff 2002). This is also one of the key benefits of foresight exercises, which, in light of their typical incarnation in large-scale processes, are frequently tasked with engendering communities of practice and establishing durable policy networks. Both of the online Delphi exercises presented below were undertaken in the context of future-oriented exercises with a strong strategic dimension. The first concerned higher education and took place in the summer of 2010;[1] the second was developed within the process of elaborating a national R&D strategy and was carried out in the late summer of 2013.

The twin key characteristics of the classic Delphi design are its iterative nature and the controlled feedback provided by the facilitators after each round. Together, they ensure anonymity (so as to minimize conformity), focus, and a gradual decrease in the spread of quantitative estimates as respondents rethink their quantitative and qualitative input in light of the feedback summarizing the positions of the group members. As suggested previously, online Delphi has tended to decrease the number of rounds (to the point of giving up the iterative format) and to simplify or automate the feedback (often reducing it to a simple statistic and/or access to other participants' qualitative input 'as is'). What is gained, in exchange, is an economy of time and effort, and broader participation.

In the exercises presented here, we followed a similar path, with some twists along the way. As far as iteration is concerned, we reduced the number of rounds to two (in the first exercise) and, respectively, to just one (the second exercise). We ensured interaction with the other respondents by making available – and, indeed, encouraging – repeated or renewed input by the respondents as the exercise advanced and more responses accumulated. However, we stopped short of providing instant statistics to the participants; in the first exercise, they were provided, in typical fashion, after the first round.

[1] The Romanian Agency for the Funding of Higher Education, Development and Innovation (UEFISCDI) undertook in 2008 a systemic foresight project titled "Quality and Leadership for Romanian Higher Education in 2025" (QLHE), which was completed in 2011. The project aimed to diagnose Romanian HE, chart the challenges ahead, envision a shared desirable future for higher education, and propose a systemic transformation. The project is described in some detail in (Andreecu et al. 2012).

In what regards feedback, we automated the procedure but introduced constraints that, we hoped, would prevent the qualitative content from exploding beyond the (reasonable) limits of manageability. To this end, we also introduced an innovative procedure – the real-time, 'dynamic' ranking of arguments.

It should be noted here that even in online Delphi there has been a reluctance to do away completely with controlled feedback, i.e., with some human processing of the qualitative data generated by respondents. Thus, Gnatzy et al. (2011) employed a human facilitator who reviewed the qualitative data – clustering and cleaning it – before it was speedily made available to the participants. Such a solution seems to have been manageable with their pool of experts (62 participants in the online exercise). However, a substantial increase in the number of respondents beyond this threshold would have likely rendered much more difficult the task of providing feedback *in (quasi-)real time*.

Gordon and Pease (2006) did not provide any facilitator-controlled feedback on the qualitative data, which amounted simply to a list of arguments updated in real time as they were entered into the platform by any of the participants. This worked because participation was limited. Otherwise, the number of arguments might have exploded to the point of becoming impossible to parse in a reasonable manner.

We solved the problem of controlled feedback in the context of a large number of participants by:

(a) inviting the latter to alternatively select reasons previously offered by the other respondents and/or offer new ones;

(b) limiting the size of each argument to a few dozen words so as to keep it simple and, so far as possible, specific and precise;

(c) limiting the number of reasons that could be selected and/or entered by any single respondent; and

(d) dynamically ranking all arguments in real time, so that each new or returning respondent would be confronted with a list of previously entered reasons followed by a figure between brackets indicating the number of other participants who had already selected that reason.

Our key assumptions were that this system would tend to keep the number of arguments sufficiently low to be manageable by the respondents; and that it would focus the latter's attention on the main or most significant arguments. Specifically, we assumed that our design solution for the qualitative part of the questionnaire would:

- induce the respondents to select pre-existing arguments as originally formulated, instead of unnecessarily adding 'new' reasons that mirrored in part or in whole previously entered reasons (thus rendering unnecessary the subsequent clustering of arguments for feedback);
- keep the conversation concentrated around the key / most 'popular' issues and prevent irrelevant digressions (such as, for example, those in which experts devote at least part of their arguments to pet ideas only marginally relevant to the matters at hand);

- make it easier to interpret the ‘meaning’ or significance of the quantitative result, by highlighting what arguments tended to support it or go against it.

This final point deserves a little elaboration. Given the electronic format of the exercise, it is simple for the organizers of the exercise to determine, after completion, which exactly were the reasons adduced by each individual respondent in support of his or her quantitative estimate. Indeed, such a close look at the data would be crucial to any in-depth analysis of the results. However, unless the real-time feedback to the respondents were very complex (and thus needlessly burdensome and possibly distracting), such information would remain inaccessible to the participants themselves. While the simple quantitative metric provided by the ranking of arguments (the number of ‘votes’ accumulated by each argument up to that point) does not by itself paint a clear picture of how the arguments are distributed in relation to the quantitative estimates, it does provide a hint as to what the main points of agreement and contention are likely to be. When the respondents are not provided with a statistic of the group responses before they enter their quantitative variable (e.g., Gordon & Pease 2006), a solution which we considered prone to bias respondents towards conformism, such information becomes particularly relevant.

The design choices made in light of the assumptions above naturally entailed some trade-offs:

- the ranking could increase the bandwagon effect, as participants, especially those logging in late in the exercise, may be tempted to go with the more popular arguments;
- some potentially relevant arguments may be sidelined if they do not manage to attract a sufficiently large number of votes early on, as they might be comparatively disregarded by later respondents.

This being said, in light of the potential advantages of the solution, we considered these trade-offs to be well worth it.

Results, discussion and implications

This section provides some additional, though still sketchy, detail on the two eLPHI exercises. Given the complexity of the processes within which they were embedded, we must leave out many details, some of them rather relevant. We focus therefore on the key design elements, and particularly on pertinent commonalities and differences.

The 2010 Delphi exercise was designed to test the broader acceptability of and to improve on a set of shorter-term policy goals by reference to a vision document (operationalized, for the purposes of the exercise, as three institutional scenarios). The goal of the exercise was thus to improve the fit of these broad policy objectives to the contemplated vision of Romanian higher education in 2025. Since the policies envisioned had been predefined by several expert panels, the online Delphi was used to:

- evaluate quantitatively the match between each of the policy goals proposed and the substance of the vision,
- support this quantitative assessment with brief arguments, and
- advance new policy suggestions, if and where the experts felt this was appropriate.

In the 2013 exercise, of which one key component was defining a limited number of strategic fields of smart specialization, respondents were asked to evaluate some 6-8 research and innovation programs within a specific RDI field with respect to a number of criteria pertinent for defining a 'smart specialization'. Specifically, the Delphi served to:

- narrow down a short list of candidate smart specialization fields (by enabling a ranking of the fields); and
- provide a set of arguments for and/or against the prioritization of several research and innovation programs within these fields.

For the first exercise, we targeted a pool of several thousand respondents, out of which 448 fully completed their questionnaires. In the second online Delphi, an even more complex questionnaire was dispatched to around 44,000 potential respondents – researchers and academics, doctoral students or recent PhDs, members of the scientific diaspora. The online consultation lasted a month, until early September 2013.

In both exercises, respondents were asked to provide a quantitative estimate and to support it with arguments. In the 2010 higher education exercise, participants had to estimate, on an unbalanced scale, the likely impact of a policy at the vision horizon; and to justify this estimate by selecting pre-existing arguments and/or by adding a new reason. In the 2013 RDI exercise, respondents were invited to assess a candidate field of smart specialization in relation to several criteria.[2] For each criterion, they were asked to give a quantitative estimate on a scale (e.g., "In your opinion, to what extent does the candidate field provide promising research and innovation programs that can be accomplished in significant measure by 2020?"); and to justify the response by selecting pre-existing arguments and/or by entering new ones.

Figure a: Quantitative estimates and dynamic ranking of arguments in the 2010 eLPHI

[2] The social, economic, and scientific impact of the program by 2020; the currently available research resources (human resources, research infrastructures, past successful endeavors); the resources needed in the near future in order to meet the program's goals by 2020; the results expected from the program by 2020 (provided the necessary resources are met).

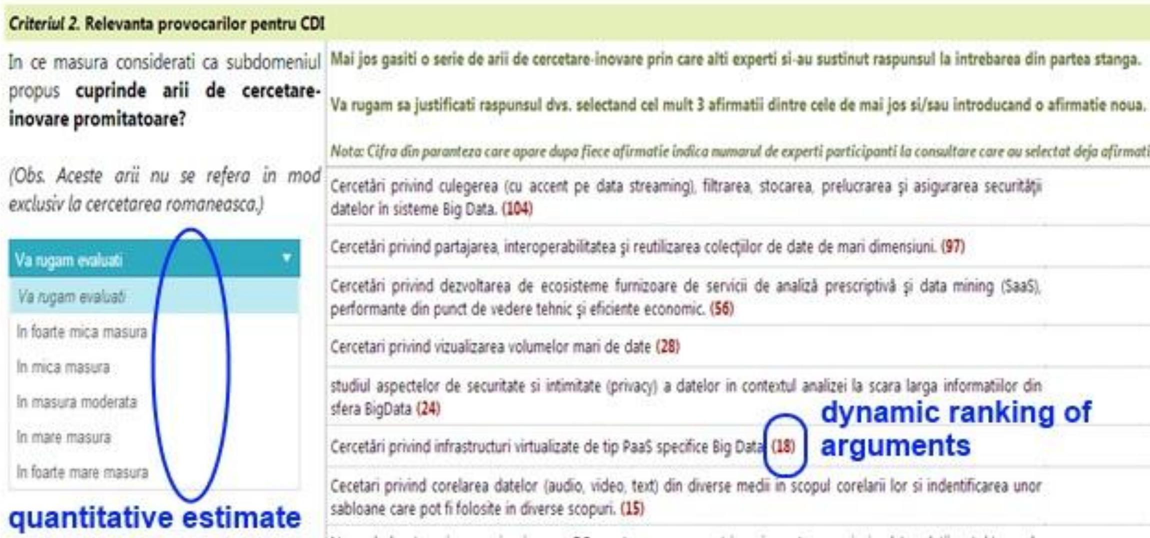

Figure b: Quantitative estimates and dynamic ranking of arguments in the 2013 eLPHI

In both exercises, each respondent already had available, before making the quantitative estimate, the list of all arguments provided by the previous respondents (or, for the first respondents, by default). These arguments were ranked according to the number of 'votes' (selections) received by each argument by that time (the default arguments started at 1); and the number of votes was visible in brackets. There was a slight difference in format, however: in the 2010 exercise the arguments became visible if the respondent pressed on a dropdown list; in the 2013 eLPHI, all arguments were visible in an extended list occupying the right frame of the page (the quantitative variable frame was placed to the left).

As a result, the respondents were able to read the arguments *before* making an estimate. This enabled us to dispense with a second round altogether in the 2013 exercise, without however providing the experts with a group statistic in advance, as Gordon and Pease (2006) did. As noted, in order to ensure that the early respondents would already have some pre-existing reasons to choose from, the facilitators provided some arguments generated after a pilot testing of the Delphi.

Finally, in both exercises participants were allowed to save their progress and return to their questionnaire at a later time. This implied that, on the next visit, the list of arguments they would consult might be larger; that the ranking of arguments might be changed; and that the number of votes cast for the arguments would be greater. To the extent that this information enabled the experts to better justify and/or calibrate their estimate, the option to save and return compensated for the absence of a second round in the 2013 exercise; and for the small number of rounds (just two) in the earlier eLPHI.

In terms of structure (rather than substance), the main difference between the two exercises was that only the 2010 online Delphi provided for a second iteration. In the latter round, the participants whose qualitative estimate diverged from the group mean beyond a pre-established threshold value were invited to re-assess their estimate in light of this information and of all the arguments provided before the end of the first round.

Findings from the two exercises

Over the set of 84 questions in the first round of the first exercise, the number of arguments added varied between 4 and 24 arguments per question. The average number of arguments was 11. (Given the multiple choice of arguments, the number of votes received by arguments was 186% the number of respondents.) The vaguely bell-shaped distribution of the number of arguments per question in the first exercise (Fig. 1a) suggests that the number of arguments depends on the type or nature of question.

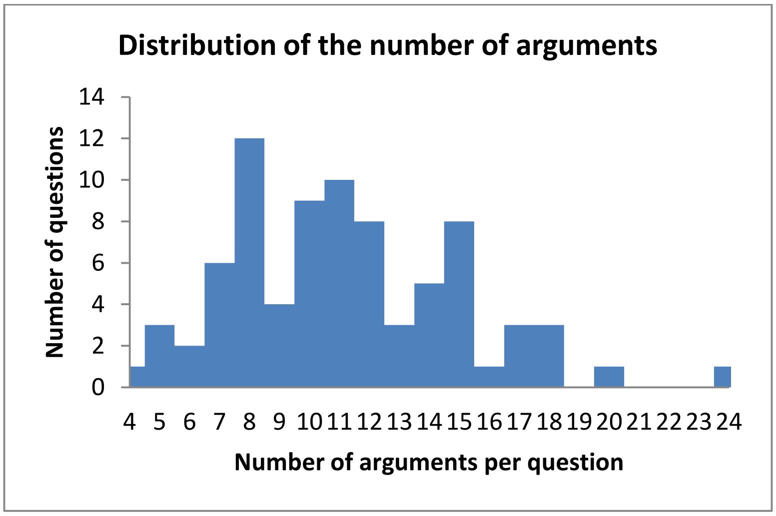

Fig. 1a: The distribution of no. of arguments per question in the first exercise

In the second exercise, 90 distinct questionnaires were administered, totaling 270 items (questions). The number of arguments per item varied between 4 and 50, with an average of 12.5 arguments per question. Each item started with a number of between 2 and 6 default arguments (with an average of 4.5). The figure below (Fig. 1b) shows the distribution of the number of arguments for the second exercise (in six cases, no new arguments were added).

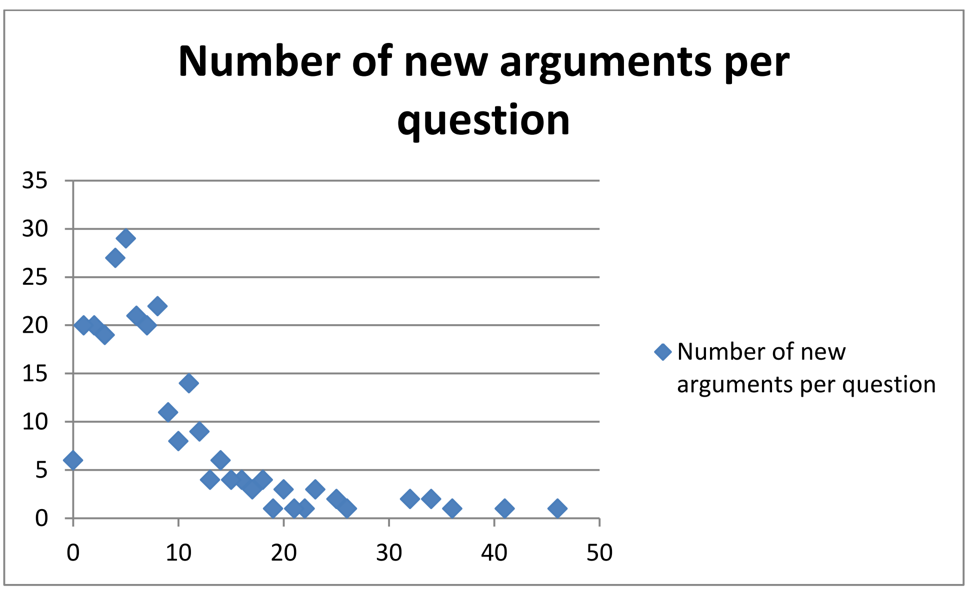

Fig. 1b: The distribution of no. of arguments per question in the second exercise

The distrbution suggests that, beyond a certain number of arguments already on the table, the respondents no longer focus on promoting a particular argument, but rather on making their voice known. As a result, the number of arguments sometimes increases substantially.

We also looked at the percentage of votes received by the *first* 4 arguments entered for each question, as an estimate of how much the order in which the arguments are introduced influences their chance of being selected. In the first exercise (Fig. 2a), the votes were relatively concentrated among the arguments entered first. Specifically, the first 2 arguments gathered between themselves, on average, a bit over 50% of the votes; the first 4 arguments – a bit over three fourths of the votes.

Argument no. (in the order of introduction on the 84 lists)	**Smallest percentage of votes received**	**Highest percentage of votes received**	**Average of votes received**
1st	16.7%	52.8%	29.8%
2nd	12.5%	35.7%	20.9%
3rd	7.2%	22.1%	14.6%
4th	5.7%	16.8%	10.8%

Fig. 2a: The first arguments and the percentage of votes they gathered in the first exercise

For the second exercise, 1,216 default arguments were made available to the first respondents across the 270 questions; an additional 2,158 arguments were introduced by the respondents, for a total of 3,374. Together, the default arguments, accounting for 34% of the total number of arguments, garnered 73,080 votes, i.e., a little over 70%, out of the total 103,745 votes. The table below (Fig. 2b) provides an additional summary for the first 4 *new* arguments entered by the respondents across the 270 items.

Argument no. (in the order of introduction on the 270 items)	**Smallest percentage of votes received**	**Highest percentage of votes received**	**Average of votes received**
1st	0.4%	39.2%	8.8%
2nd	0.3%	24.3%	5.6%
3rd	0.3%	24.3%	4.2%
4th	0.1%	19.2%	3.1%

Fig. 2b: The first new arguments and the percentage of votes they gathered in the second exercise

In other words, the default arguments together with the first 4 new arguments entered (accounting together for 68% of the total arguments) received around 92% of the votes.

We examined the same issue from a different angle by exploring the 'dynamics of arguments', i.e., relationship between the number of votes gathered by an argument and its 'exposure time', which we defined as the number of respondents to the question since the argument was first entered. For the first exercise (Fig. 3a), we found a positive moderate correlation of $R^2 = 0.474$ across the set of 929 arguments.

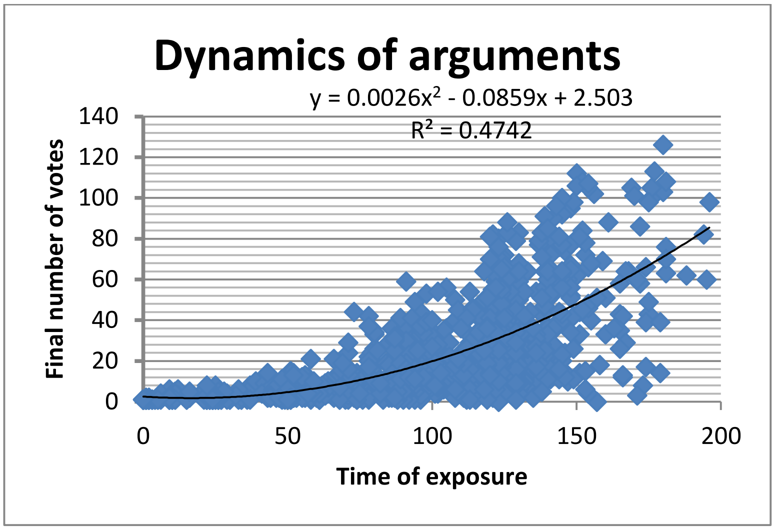

Fig. 3a: The dynamics of arguments in the first exercise

The figure was identical across the 3,374 arguments in the second exercise: R^2=0.476. We interpret this number as showing that while exposure time does have a moderate impact on the chance of an argument of being selected by the respondents, the relationship is not so strong as to impair the logic of the exercise.

We further looked into the 'degree of consensus' – the relationship between the number of arguments per question and the 'rate of convergence', i.e., number of convergent respondents per total number of respondents. This serves as an estimate of the extent to which a lagrger number of arguments increases consensus. For the first exercise, we found a low correlation between 'degree of consensus' on a question and number of arguments per question (R^2 = 0.17).

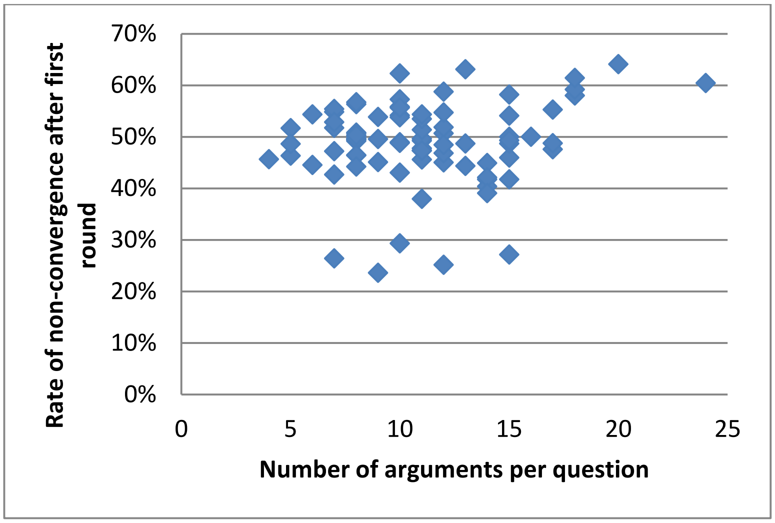

Fig. 4a: The relationship between the number of arguments and consensus in the first exercise

Finally, we looked into the 'evolution of consensus' as additional respondents entered the exercise. Specifically, we wanted to find out whether late respondents were more likely than earlier ones to converge towards the final average estimate for each question. Across the set of 84 questions in the first exercise (Fig. 5a), the correlation between the order of entry and the rate of convergence (same as above) was low ($R^2 = 0.19$), suggesting that late-comers were not significantly more likely than early respondents to give estimates closer to the final group average.

Also, in the first exercise the rate of divergence does not change substantially after the 4 first arguments. Among the first half of the respondents, 53% diverged significantly from the mean; among the second half – 47%. This suggests that the introduction of additional arguments after a certain point has at best a moderate effect on convergence. In other words, the late-coming respondents are not significantly better predictors than the early respondents.

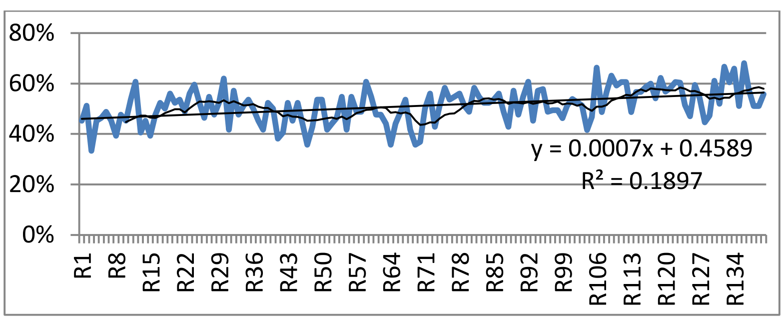

Fig. 5a: The evolution of consensus in the first exercise

Conclusions

Among the many findings and lessons learned after carrying out the two exercises, the following stand out as, in our view, particularly significant.

First, the principles of the organization of the exercise and the interaction with the platform are easily grasped. The fact that we used a video tutorial might have, however, improved the user experience, at least in this respect. Also, in light of the feedback we received directly and obtained through our own monitoring, these principles and the mode of interaction are accepted by the participants.

Secondly, for the most part the number of arguments remains at a reasonable level (questions with at most 15 arguments accounted for 77% of the total). The exception occurs when they reach a certain threshold, after which participants seem to switch from the game of promoting an argument with real chances of being selected to simply expressing their opinions. Questions with 20 arguments or above amounted to a little below 12% of the total number in the second exercise; and those with 30 arguments or more accounted for only 3.7% of the cases.

Thirdly, new arguments do have a moderate chance to enter the top of the most selected. On the other hand, it is indisputable that the order in which arguments are entered does play a significant role in which garners more arguments.

Fourth, the number of arguments is not substantially correlated with convergence in the quantitative estimate. In other words, more arguments associated to a question do not necessarily mean more (dis)agreement on the quantitative variable. This suggests that, perhaps, respondents may differ on reasoning, though not always on the way they assess things, or that they sometimes argue over finer points or formulations.

Fifth, the bandwagon effect seems limited. Late-coming participants are not substantially more conformist than early respondents.

Finally, the results of the exercise are informative for the end-users, both when they are presented as general statistics and when they are disaggregated in, for instance, the arguments provided to support each individual quantitative estimate; or the reasons behind the more extreme assessments.

References

Andreescu, L., Gheorghiu, R., Zulean, M., Curaj, A. (2012). System foresight for Romanian higher education. In: A. Curaj, P. Scott, L. Vlasceanu & L. Wilson (Eds.), European Higher Education at the Crossroads: Between the Bologna Process and National Reforms. Dordrecht: Springer.

Dalkey, N., Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. Management Science 9.

Gnatzy, T., Warth, J., von der Gracht, H., Darkow, I.-L. (2011). Validating an innovative real-time Delphi approach. A methodological comparison between real-time and conventional Delphi studies. Technological Forecasting and Social Change 78.

Gordon, T.J., Pease, A. (2006). RT Delphi: An efficient, ‘round-less’, almost real-time Delphi Method. Technological Forecasting and Social Change 73.

Helmer, O., Rescher, N. (1959). On the epistemology of the inexact sciences. Management Science 6.

Linstone, H.A., Turoff, M. (Eds.) (2002). The Delphi Method: Techniques and Applications. Available at: is.njit.edu/pubs/delphibook/delphibook.pdf.

Rayens, M.K., Hahn, E.J. (2000). Building consensus using the Policy Delphi method. Policy Politics Nursing Practice 1.

Rowe, G., Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. International Journal of Forecasting 15.

Spinelli, T. (1983). The Delphi decision-making process. Journal of Psychology 113.

Strauss, H.J., Zeigler, L.H. (1975). The Delphi technique and its uses in social science research. Journal of Creative Behavior 9.

Turoff, M. (1972). Delphi conferencing. Computer-based conferencing with anonymity. Technological Forecasting and Social Change 3.

Turoff, M. & Hiltz, S.R. (1996). Computer-based Delphi Process. In: M. Adler & E. Ziglio (Eds.), Gazing into the oracle: The Delphi method and its application to social policy and public health. London: Jessica Kingsley.