A Delphi-based expert judgment method applied to the validation of a mature Agile framework for Web development projects

C. J. Torrecilla-Salinas[1] · O. De Troyer[2] · M. J. Escalona[1] · M. Mejías[1]

Abstract
The validation of any new methodological proposal demands several real-life implementations. However, organizations are reluctant to invest without the firm guarantee that they will be returned the entire expended amount of money. For this purpose, expert judgment techniques are very useful to provide a less-costly initial validation that, when positive, may encourage organizations to use these new proposals. Therefore, the primary goal of the paper will be to assess how expert judgment techniques based on the Delphi method can be applied to Web Engineering field and, more in particular, to assess the validity of the *NDT-Agile* framework. *NDT-Agile* is a framework that combines Agile and Web Engineering techniques to meet Capability Maturity Model Integration development goals. The paper presents a real example of an application of a Delphi-based expert judgment method to assess *NDT-Agile* framework validity, explaining the design as well as the selection and usage of the different techniques it involves. The application of the method will allow assessing benefits and limitations of use in Web Engineering. As a main conclusion, we will state the utility of the proposed methods to obtain a low-resource initial validation of a certain proposal. Finally, we will identify further lines of research related to the analyzed topics.

Keywords Agile · Web Engineering · CMMI · Delphi · Expert judgment · Organizational issues

1 Introduction

It is known that the full validation process of a new framework or methodological proposal can take a long time, as it might need several real-life implementations in order to gather empirical data. In consequence, before starting real-life implementations, organizations will often claim for some guarantees on the results, because both their investment and reputation are at stake in each project. This fact might delay considerably the deployment, implementation, test and introduction of any new proposal, hindering significant innovation and organizational improvement.

The expert judgment methods appear as a way to provide organizations with certain guarantees to assess the expected results of a particular new proposal. They can be considered as a "compromise" solution in which an initial evaluation can be carried out with less investment, anticipating somehow the expected results of the application of these new methods.

These types of approaches, that are based on the combined use of Delphi method [17] and several statistical techniques, are quite common in other fields, such as social science or medicine, but they are not that frequent in the context of Software and Web Engineering.

The main goal of this paper will be the validation of an Agile framework, named *NDT-Agile*, that tries to provide a set of Agile practices to achieve all CMMI-DEV maturity

✉ C. J. Torrecilla-Salinas
carlos.torrecilla@iwt2.org
http://www.iwt2.org

O. De Troyer
Olga.DeTroyer@vub.ac.be

M. J. Escalona
mjescalona@us.es
http://www.iwt2.org

M. Mejías
risoto@us.es
http://www.iwt2.org

[1] ETS Ingeniería Informatica, Av. Reina Mercedes S/N, 41012 Seville, Spain

[2] Department of Computer Science, Vrije Universiteit Brussel, Room: F.10.718, Pleinlaan 2, 1050 Brussels, Belgium

level goals for organizations developing systems in Web Engineering environments.

As it is known, CMMI [10] is a recognized maturity model that is used by organizations to improve their processes, being CMMI-DEV the particular version of CMMI dedicated to software development. Agile [4] is a generic label applied to a certain set of methodologies that tries to promote close collaboration between business and IT teams and early delivery of software, among other benefits.

The relation between Agile and CMMI has been subject of study during last years, both in the field of general Software Engineering [63] and in the field of Web Engineering [69] Along this time, both Agile and CMMI approaches have moved from initial reluctances [65] to initiatives of combined implementations [28, 63, 69]. In particular, in the field of Web Engineering, as Web systems development shared synergies with Agile approaches [2, 45] and CMMI is normally associated with increases in software quality and customer satisfaction [30], the definition of a framework that, at the same time, could support Web specific characteristics, ensure agility and fulfil CMMI specific and generic goals might be considered valuable both for Agile and CMMI-certified Web development companies. The former would gain in Agile institutionalization, removing customers and partners' fears. The later would increase agility and responsiveness without loosing the benefits of a well-established process. In order to fill in this gap, we find *NDT-Agile* [67, 70], as a mature Agile framework that could, simultaneosuly, support Web developments using advanced Web Engineering techniques and meet all required CMMI-DEV specific and generic goals.

This paper will present the design of an expert judgment method, based on the Delphi method, in order to assess the validity of *NDT-Agile* proposal. Based on the foregoing, our work tends to reach the following goals:

- Present an overview of the different elements to take into account when designing an expert validation method, including mechanisms to measure stability and consensus.
- Present the results of a real Delphi-based expert judgment method for *NDT-Agile*, by means of the use of the presented techniques in a combined way in a real-life experience to assess the feasibility of its application to Web Engineering field.
- Obtain initial results of the validity of *NDT-Agile* proposal regarding its different dimensions.
- Present relevant conclusions and suggest further lines of research.

For this purpose, this study is organized as follows: after this introduction, Sect. 2 will introduce the research questions and method. Afterwards, Sect. 3 will present the background related to *NDT-Agile* and Delphi method. Section 4 will describe the design of our Delphi-based expert judgment method. Then, Sect. 5 will summarize the results of the conducted process, and finally, Sect. 6 will draft the main conclusions of the paper and will recommend further lines of research.

2 Research questions and method

This section will present the research questions together with the proposed research approach. These questions are linked to the generic question: "How to design a suitable expert-judgment method to validate a proposal in the Web Engineering field?". In order to be able to answer this generic question we have divided it into the following more detailed and specific research questions:

- *RQ1* What should be taken into account when designing an expert judgment method?
- *RQ2* What are the most suitable techniques to process gathered data during an expert judgment method?
- *RQ3* How to identify when consensus is reached during an expert judgment method?

The answers to RQ1 will help us to identify firstly the right expert judgment method that should be used and secondly, how to design it in detail. The answers to RQ2 will allow us to define the gathered data processing in a proper way and the ones to RQ3 will tackle particularly the issue of consensus measurement. Managing all of them correctly (method, detailed design, statistical tools and consensus measurement) in the context of the *NDT-Agile* proposal will initially respond to our main research question.

When choosing a research approach, the first decision to make is linked to select which type of approach will be more suitable; a quantitative approach or a qualitative one. According to Creswell [12], a qualitative research approach can be acceptable when the research object is still to be understood. That is the reason by which we selected the qualitative approach. We tried to obtain some general conclusions by means of a concrete application of an expert judgement method to the validation of a particular Web Engineering method.

Following this initial approach, and in order to answer the proposed research questions, we started with analyzing the existing literature regarding expert judgment techniques in general and Delphi method as well as available statistical techniques to process data. Most of the conclusions obtained during this phase will be presented in Sect. 3.

After that review, we defined the subject to be studied during our Delphi method and designed the expert panel accordingly. We also selected the statistical tools to process

the compiled data. The results of this process will be shown in Sect. 4.

Next, we ran the designed Delphi method and finally we processed the compiled information and extracted relevant conclusions by means of the selected techniques. Results of this last phase will be discussed in Sects. 5 and 6.

It is worth emphasizing that, although the Delphi method has been widely used in the field of social and medical sciences, and even in the software development field [47, 61], we were not able to find out in literature applications of this method to the field of Web Engineering (e.g. to validate a new Web Engineering proposal or framework). Moreover, the answer of the above presented research questions will also help to describe how the Delphi method can be tailored and combined with several statistical techniques in order to define a suitable expert-based validation process for Web Engineering, being therefore one of the main contributions and novelties of our paper.

3 Background

In this section we will provide the necessary theoretical background both on the assessed framework (*NDT-Agile*) and on expert judgment techniques and Delphi method, in order to allow a better understanding of the proposal. Finally, it will include some basic information on how to measure consensus and stability along Delphi methods.

3.1 Agile and NDT-Agile

As mentioned, different Systematic Literature Reviews (SLRs) have analyzed the relation between Agile and CMMI-DEV, both for generic software environments [63] and for Web Engineering contexts [69]. We have recognized the existing approaches trying to establish the relation among the fields and identifying the existing gaps. These studies point out only a few approaches combine or modify Agile techniques in order to meet the different CMMI-DEV goals. One of them is Model C-S [42], a Scrum modified proposal, including up to 123 practices, which maps the specific ones of CMMI-DEV maturity levels 2 and 3. The model excludes deliberately those CMMI-DEV process areas that are related to organizational issues, like those of levels 4 and 5.

The studies previously mentioned also highlight some works that present analyses on how Scrum [19, 43], XP [52] or Kanban [6] can cover totally or partially the goals for CMMI-DEV maturity levels 2 and 3. All those works conclude that none of these Agile approaches can, by themselves, cover all the goals, although they can be seen as compatible with the maturity model, as they cover a significant amount of them.

Another conclusion that we can extract from [69] is that no Agile proposal, apart from *NDT-Agile,* could cover all CMMI-DEV specific and generic goals for all maturity levels in the particular case of Web development environment. This means that, despite the large amount of different Agile proposals currently existing and being used, and among the few known Agile proposals to map CMMI-DEV maturity levels' goals, *NDT-Agile* seems to be the only one that systematically covers all of them. Therefore, it is the only suitable approach for those organizations developing Web Systems that aim to keep agility and ensure CMMI-DEV compliance.

Based on these reasons, the main goal of the expert judgment process that we will present in this paper is assessing how an expert judgment method can be designed and used in the Web Engineering field. As a secondary goal, we will validate *NDT-Agile*, a mature Agile framework for developing Web development projects. At the same time, this method tries to cover all CMMI-DEV [10] specific and generic goals, in order to keep organization and project's agility as well as to support all Web development specificities.

NDT-Agile can be seen both as an extension of standard Scrum [66] and XP [3] and as an extension of NDT (Navigational Development Techniques) [21], aiming at covering all CMMI-DEV specific and generic goals. It also keeps an Agile approach and supports Web Engineering special characteristics, acting as a mature Agile framework for Web projects. Figure 1 tries to depict it graphically.

A detailed description of *NDT-Agile* is out of the scope of this paper, although deep information on its proposed lifecycle can be found in works like [67, 70]. Nevertheless, we will offer high-level overview of its design and structure.

Regarding the framework definition and design, and in order to analyze the relation among Agile, CMMI-DEV and Web Engineering and identify possible research gaps, a Systematic Literature Review was conducted [69] according to Kitchenham et al. [37]. As a result of the review, all relevant existing related papers were identified and some relevant

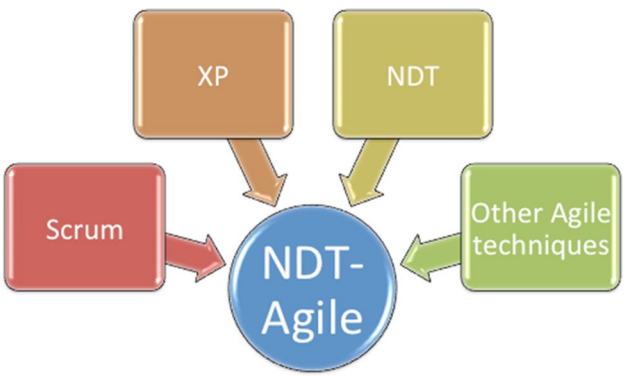


Fig. 1 *NDT-Agile*: a mature Agile framework for Web Engineering

conclusions were drawn. For instance, the review stated that the possibility of proposing an Agile approach to achieve all CMMI-DEV goals in Web environment would be interesting both to Agile and CMMI-certified companies.

As Scrum and XP are the two most popular Agile approaches used either alone or even in combined implementations [53, 73], a gap analysis between Scrum and XP standard practices and the goals of CMMI-DEV maturity levels 2 [72], 3 [71], 4 and 5 [68] was also performed. As a general conclusion of those works, it can be confirmed that, although Scrum and XP by themselves are not able to cover all CMMI-DEV goals, they can be seen as compatible with CMMI-DEV, as they cover a significant percentage of the CMMI-DEV goals, at least for maturity levels 2 and 3 [71, 72]. Based on this finding, other existing Agile approaches were analyzed and other potential Agile practices to fill in the identified gaps between Scrum and XP, on the one hand, and CMMI-DEV specific and generic goals, on the other hand were identified. Those other Agile approaches consisted in other existing well-know Agile techniques or ad-hoc modifications to standard Scrum or XP proposals that aimed to achieve some of the CMMI requirements.

Based on the conclusions of the analyzed works, *NDT-Agile* tries to incorporate in a single, coherent and comprehensive framework all of these Agile approaches, proposing a framework that organizations can use to progress through the different CMMI-DEV maturity levels only by means of Agile techniques.

NDT-Agile consists of three main elements: firstly, *NDT-Agile* lifecycle for projects, basically based on Scrum lifecycle and composed of a set of different phases and techniques; secondly, a set of seven Agile complementary techniques that extend the suggested core lifecycle so as to simultaneously ensure the full coverage of CMMI specific and generic goals, as well as the maintenance of the proposal's full agility; and finally, it includes a governance proposal to assure its correct deployment, adaptation and continuous improvements. Figure 2 displays the different elements of the methodology.

Figure 2 represents *NDT-Agile* iterative and incremental lifecycle, which is encapsulating NDT, together with the different proposed Agile complementary techniques that compose the framework. *NDT-Agile* governance wraps the two other elements to deploy, adapt and improve the organization-wide methodology.

NDT-Agile lifecycle [70] comprises two phases: *Project launching* (the only non-iterative element of the framework), where the initial plan is depicted by means of Agile methods, and *Project development*, where the project is developed and plans are adapted according to a Sprint-based lifecycle. Figure 3 shows the proposed lifecycle.

Figure 3 shows the two framework phases, displaying the main iterative approach of the proposed lifecycle.

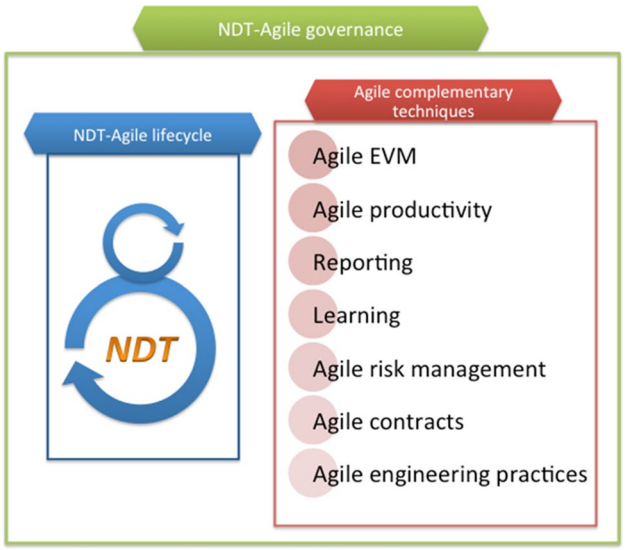


Fig. 2 Elements of *NDT-Agile*

As mentioned, the framework also includes a set of seven Agile complementary techniques that are used to ensure both CMMI-DEV compliance and project agility, as Fig. 4 displays.

To conclude, a governance model, focused on the existence of a governance body that monitors framework deployment, customization and improvement, is proposed to guarantee coherence at organization level.

3.2 Expert judgement techniques and Delphi method

Expert judgement techniques are those who allow a certain number of experts in a particular area of knowledge to express a shared opinion on a particular topic. There are several expert judgment techniques that can be used for diverse purposes such as forecasting, evaluation or policy design, among others. Some of the best-known techniques are:

- *Brainstorming* [20, 51] A group shares ideas associated with an issue with the goal of moving away from the constraints of the more formal problem-solving sessions. The main goal is to produce as many ideas as possible and as much ground-breaking as possible. Besides, the combination or merging of proposed ideas is encouraged.
- *Nominal group technique (NGT)* [8] It is a technique in which an expert group is defined in order to identify elements of a problem, a potential solution and established priorities. In NGT technique, the experts gather together physically and, by means of facilitator guidance and following a structured approach, they discuss, vote and rank the elements of the analyzed problem.

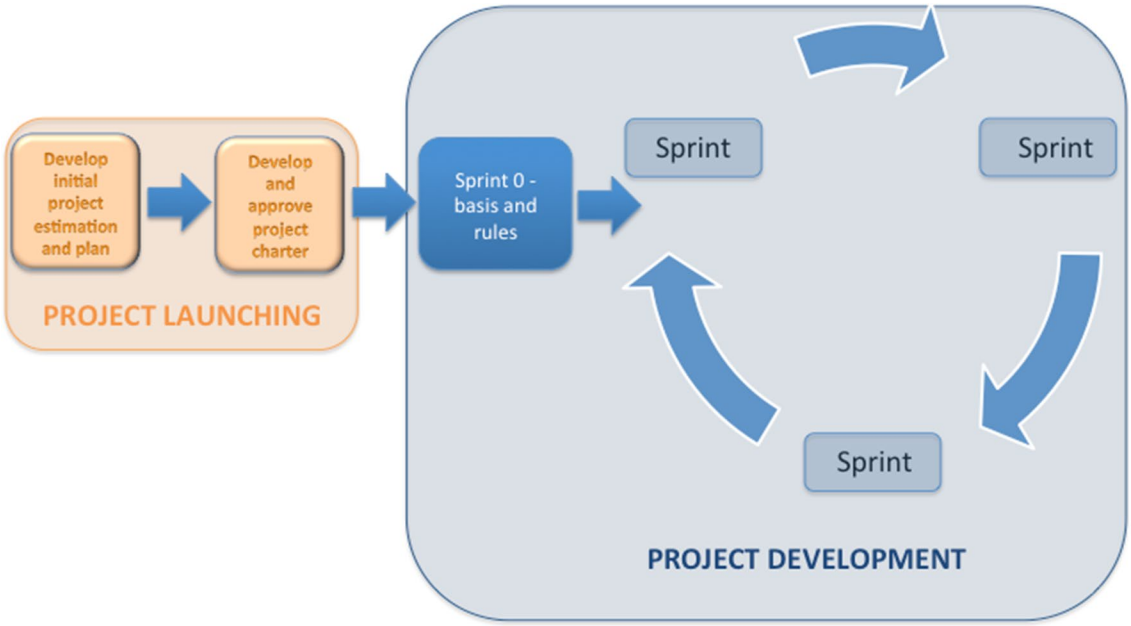


Fig. 3 Lifecycle of *NDT-Agile* [70]

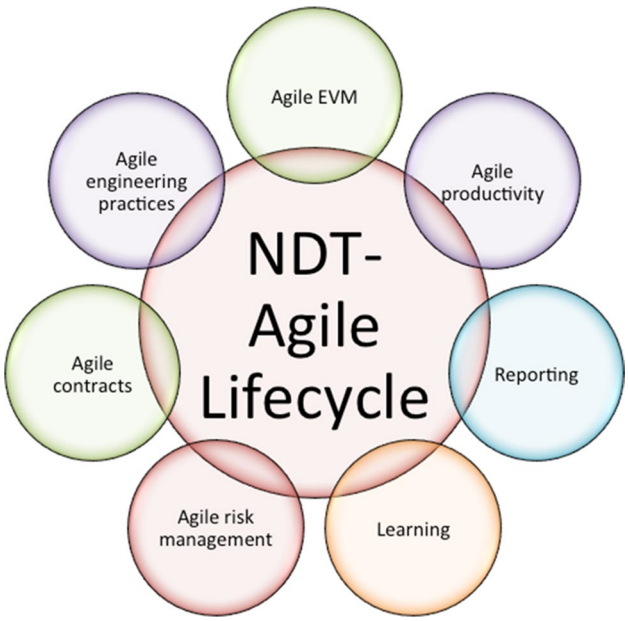


Fig. 4 *NDT-Agile* complementary techniques

- *Delphi method* [32] A group of experts discusses on a selected topic trying to reach consensus, but in this case by means of questionnaires and through successive rounds, preventing experts from knowing each other to avoid group pressures through discussions.
- *Didactic interaction* It is normally used in cases of "yes/no" decisions. It consists in on-site meetings, where the group splits in two subgroups that express different views and discuss about them. Then, they change roles and each group defends the opposite position. This change might help to understand each other's position and move the group through consensus.

As mentioned before, the main goal of the presented research is to assess how an expert judgement method can be used in the Web Engineering field, in particular to evaluate the validity of the proposed *NDT-Agile* approach, before running practical implementations. In order to choose the most suitable expert judgement technique, the above-mentioned ones were qualitatively assessed. As the main objective of brainstorming technique is to bring new and different views on an existing problem, it seems to be less suitable for a general evaluation, as it is the particular case of this research. In light of this, it was discarded as an appropriate approach.

As explained, the didactic interaction is mainly thought for "yes/no" decisions. It is not the case of the current research, which has the goal of identifying the validity of the model from different angles and elements (like CMMI-DEV compliance, Agility or Web Engineering support). This assessment cannot be simplified to a "yes/no" choice, being that the reason to reject the approach for the current research, as well.

The two remaining approaches, NGT and Delphi, seemed to be suitable for the proposed validation process, but NGT requires physical presence of the expert, and, as it will be presented later on, the selected panel was geographically distributed all over 8 different countries. The cost of bringing together the panel for debate, together with the possibility

of avoiding influence of dominant characters in face-to-face discussions led us to choose the Delphi method.

Regarding the Delphi applicability to this particular case, Rowe and Wright [59] suggest that it is applicable when the use of statistical methods is inappropriate, when a relevant number of experts is available and when alternatives are either to average the forecast of several individuals or to use a traditional group. In our case, there is no large amount of data coming, for instance, from a massive practical application of *NDT-Agile*. Thus, the use of statistical techniques to assess its validity is not a possibility. In addition to this lack of data, a proper panel to evaluate the model was identified, as it will be described later on, and alternatives to assess it (via a simple average or a traditional group) were not an option, due to the geographical distribution of the experts. Based on the described elements, Delphi seemed to be a reasonable research method to respond to our generic research question and subsequent specific research questions.

The Delphi method is an expert judgment approach that the RAND Corporation designed along the 1950s. Initially, it was used for its internal research [32] and, due to the special nature of the RAND projects mainly working on the military industry, it was not issued until 1963 by Dalkey and Helmer [17].

This method gained in popularity during the different decades of the XX century, as reported in [58]. Heiko [32] explains that the community is still highly interested in the Delphi method, evidenced by the number of Delphi-related articles published recently.

As mentioned, Delphi is an expert judgment method, in which well-recognized experts on a certain field express their views and opinions in a series of rounds following structured questionnaires, which are processed and returned to the experts as feedback for the subsequent rounds [17, 40]. The main goal of the Delphi method is usually to reach a consensus among the aforementioned expert panel in relation to a certain number of issues. As reported by Heiko [32] and following Rowe et al. [59], four are the main characteristic elements of a Delphi method:

- *Anonymity* [16] During the Delphi method the panel members do not know each other. Anonymity is crucial during the application of the method to avoid group pressure and excessive weight of dominant behaviors, as well as to ensure that opinions are expressed freely regardless of public criticism.
- *Iteration* The method is executed in a series of rounds, with the goal of achieving stability on the results provided by the panel.
- *Controlled feedback* [16] As described previously, Delphi is an expert judgment method with the main goal of gathering feedback from a set of recognized experts on a certain field. Nevertheless, as the method is normally based on a set of pre-defined questionnaires, the received feedback is always structured and "guided" by the organizers of the method in order to avoid unnecessary "noise".
- *Statistical "group response"* [16] As a conclusion of the different Delphi rounds, the organizers provide the experts with a report containing the statistical processing of the panel opinion, normally including mean, median and standard deviation, together with the reviewers' comments. Based on this report, experts can change their opinion during the following rounds.

As Hsu and Sandford [35] describes, a Delphi method usually comprises the following phases:

- *Selection of the subject to be analyzed* This is one of the most important phases of the Delphi method, as a good selection of the subject to be analyzed directly makes an impact on the quality of the gathered results.
- *Panel selection* This phase consists in choosing the members of the panel of experts that will participate in the Delphi method. There are no clear criteria regarding how to select the participants, as reported by Hsu and Sandford [35]. Nevertheless, they state, citing previous authors [50, 54], that elements such as professional background, knowledge of the subject to be analyzed and willingness to participate might be valid criteria to be a member of the panel. Regarding the group size [35] points out that there is no consensus in the literature. In contrast, it has been documented [41] that most of the Delphi studies use a panel that varies from 15 to 20 members. As a main recommendation, the number of panelists should be high enough to be representative and low enough to keep the process manageable.
- *Round 1* It is common to start the Delphi method with an open-ended questionnaire that helps the organizers to better understand the analyzed subject and identify its most relevant aspects. The conclusions of such a questionnaire lead to the development of a deeply structured one that will lay the foundation for the next rounds of the method. It is important to note, as Hsu and Sandford [35] report, that starting round 1 directly with a well-structured questionnaire becomes an acceptable modification of the Delphi method.
- *Further rounds* The results of the initial round are gathered and processed using statistical techniques. A summary of the conclusions, normally including statistical values such as mean, median or standard deviation, together with the anonymized textual reviewers' comments is sent to the participants as an input for the next round. Based on these conclusions, the experts can modify their assessment in the following round, if appropriate. Even if theoretically a Delphi method can run con-

tinuously until consensus is reached, the literature [7, 14] shows that quite often a maximum of three iterations is enough to gather the key information.

- *Conclusions* Once the number of necessary rounds is reached (and, as mentioned before, it highly depends on the analyzed subject, the panel and the received feedback), the gathered information is processed, analyzed and, usually written in a report.

Figure 5 displays graphically this process.

Finally, and to conclude this Delphi method overview, it is worth pointing out the most common limitations and weaknesses this method entails. As Hsu and Sandford [35] explain, criticism of Delphi method is based on the following arguments, among others: potential low response rates, high time consuming and potential to mold opinions.

3.3 Consensus and stability in Delphi method

According to Heiko [32], it is important to differentiate two main concepts when processing Delphi gathered data. One is "consensus", meaning convergence of opinions towards a certain value, and the other is "stability", meaning that values are consistent through different rounds. Some works, such as [15], discuss that "consensus" without "stability" is meaningless and propose a decision tree in order to determine stopping criteria for Delphi methods taking into account both elements. Figure 6 shows the decision tree.

As Fig. 6 shows, stability must be assessed after each round. If stability is reached, then consensus will be examined. If one of the two conditions is not met, a new round will be organized.

Heiko [32] explains that there has been a debate on whether stability should be measured at group or at

Fig. 5 Delphi method overview

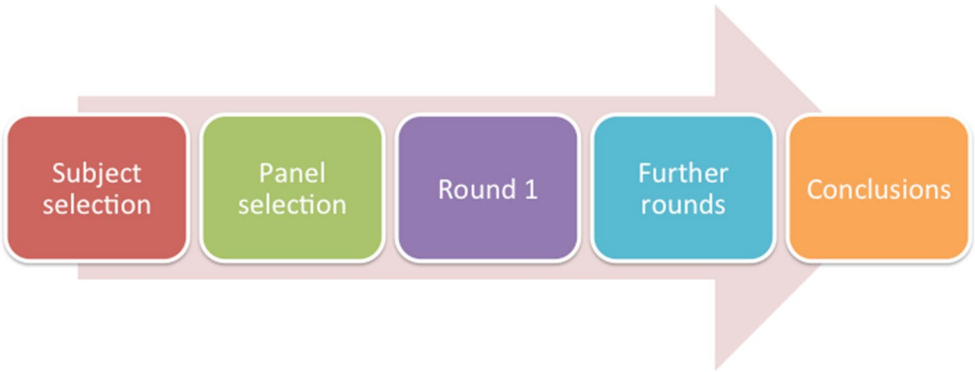
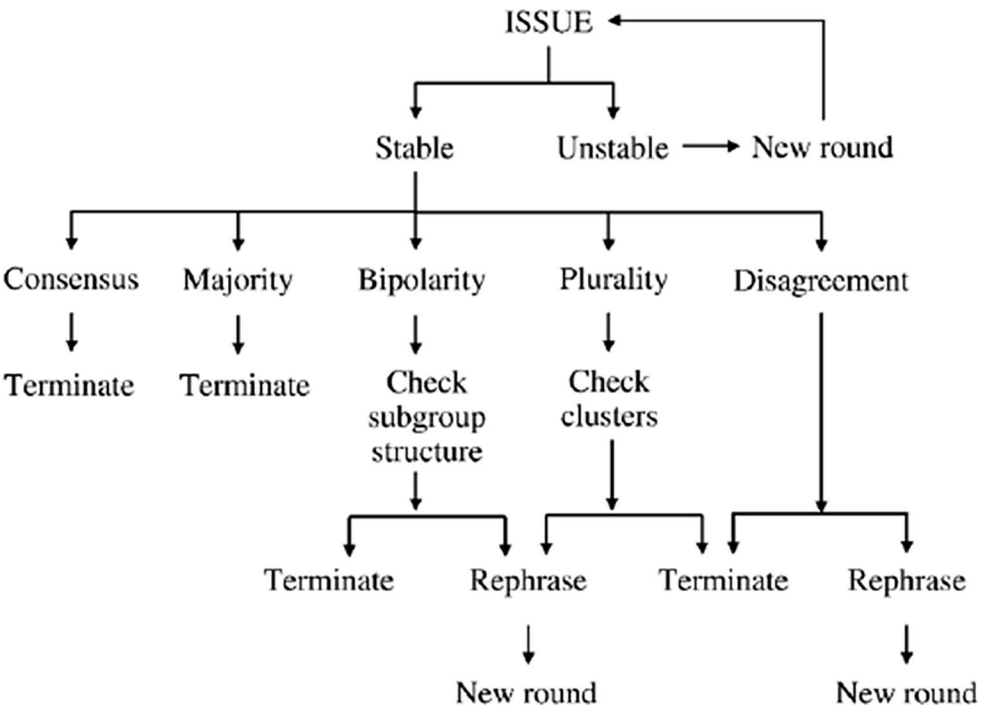


Fig. 6 Hierarchical stoping criteria for Delphi methods [15]



individual level. Some authors claim [9] that it should be measured individually, as high fluctuations might happen for certain panelists and should be compensated by some other panelists' opposite fluctuations. On the contrary, other authors [60] express that as Delphi method is trying to gather a certain group opinion and not individual points of view, stability should be measured at group level.

The issue of consensus is, however, a wider element, as there are several approaches that measure it. Heiko [29] presents a literature review on how consensus is measured in different reported Delphi methods. From the conclusions of this work, we can identify that there are basically two main approaches for consensus measurement in Delphi methods:

- *Qualitative analysis and descriptive statistics* In this group, elements such as pre-defining a established amount of rounds to finish, performing a subjective analysis to detect certain level of consensus, or using elements such as mean, median and standard deviation, are used to evaluate and reach consensus among panelists.
- *Inferential statistics* Among this group we can find different statistics that try to establish several relations among variables. Statistics like Chi square, Cohen's kappa, Fleiss' kappa or Kendall's W can be used to measure the consensus level, depending on the defined scale.

If we analyze in detail the latter, we can find several statistical techniques that can be used for "consensus measurement", or to be more statistically correct, for concordance analysis. The more relevant are listed below:

- *Chronbach's alpha* [13, 62] This statistic is used to measure the level of reliability of a particular test. Normally, when we aim to assess a non-directly observable magnitude by means of a set of n directly observed magnitudes (for instance n answers to a questionnaire), Chronbach's alpha measures the level of reliability of the proposed scale.
- *Cohen's Kappa and Weighted Cohen's Kappa* [11] These statistics are used to analyze the degree of concordance among raters, when the number of raters is two and the rated variables are given in a nominal scale (for instance, a rater is classifying items in two categories: suitable or non-suitable).
- *Fleiss'Kappa* [24] This statistic is an evolution of the aforementioned Kappa. It is used to analyze the degree of concordance among raters when they are more than two and the rated variables are also given in a nominal scale. The advantage of Kappa's statistic is that it corrects random effects and it is relatively easy to calculate.
- *Kendall's coefficient of concordance (W)* [38] This statistic tries to measure the concordance among n raters rating m variables that are given in an ordinal scale. A high value of W statistic might be interpreted as if experts apply the same standards when assessing the items [64].

Besides, some of the previously mentioned statistics can also be used to measure stability through the different rounds, as [34] shows by means of using Kappa to measure stability among the different rounds of a Delphi method.

Together with the two approaches identified by Heiko [32], another statistical technique can be used to measure consensus: *Simple Correspondence Analysis (SCA)*. SCA is an exploratory statistical technique proposed by Hirschfeld [33] and developed by Benzécri [5] during the 70s. It tries to represent graphically, in a two or three-dimension graphic, a large amount of data. It provides a way to measure homogeneity among two or more categories of two variables. Raters' homogeneity (meaning how equally they are behaving) can also be interpreted as a measure of the extent to which they agree on a certain topic.

4 Designing a Delphi-based consensus-making method

This section will describe how our Delphi-based consensus-making method was planned, indicating how the experiment was designed, how the panel was selected, how the different rounds were conducted and which elements were used to assess the level of consensus and stability among the panel members.

4.1 Subject definition

The initial step comprises the subject definition and the questionnaire design. As referred, there are two approaches when designing a questionnaire for a Delphi method: using the first round to clarify the questionnaire or starting the first round with an already well defined one [35], being both acceptable approaches. We chose the second approach as the elements to be validated were already clear at the beginning of the process, although the questionnaire itself is potentially subject to small changes or adaptations based on the received feedback. For this purpose, statistical techniques will be used to ensure that the questionnaire is reliable throughout the process.

As it has been explained, one of our goals was to evaluate *NDT-Agile* as a coherent framework that, at the same time, could guarantee Agility, suit Web development projects specificities and help to achieve all CMMI-DEV goals. Based on the foregoing, there are four dimensions to be assessed by the panel during our proposed Delphi method:

- *Agile*: It assesses the extent to which the framework guarantees project agility, meaning adaptation to changes,

quick value delivery or increased stakeholder communications, among other parameters.

- *CMMI* It evaluates the extent to which the framework achieves the different specific and generic goals of the different CMMI-DEV maturity levels.
- *Web* It analyzes the extent to which the framework supports Web development project specificities, regarding navigation, maintenance, security or user-feedback, among other elements.
- *Framework* It studies the internal coherence of the framework that is to say, whether it is complete or not, or whether it proposes or not contradictory techniques.

We can consider that each of these dimensions is independent of the others, meaning that it is perfectly possible for the framework to be CMMI-compliant, whereas Agile is not or the other way around. Based on that, our approach started both with dividing the questionnaire in four sections, related to each of the identified dimensions and defining a set of statements that could allow assessing each of the dimensions described below:

- *Agile* In this case, the statements are defined following the "Agile manifesto" [4] values and principles. The proposed statements are displayed in Table 1.
- *CMMI* In this case, the statements are defined based on the different CMMI-DEV maturity levels. The proposed statements are displayed in Table 2.
- *Web* In this case, the statements are defined according to the specific Web development characteristics identified

Table 1 Proposed statements to assess Agile dimension

Id	Statement	Rationale
Q1	*NDT-Agile* enhances personal interactions and communication among team members	Based on the first statement of the "Agile manifesto": "Individuals and interactions over processes and tools", this statement evaluates whether or not *NDT-Agile* still promotes communications and interactions
Q2	*NDT-Agile* promotes early delivery of working software to business	Based on the second statement of the "Agile manifesto": "Working software over comprehensive documentation", this statement focuses on *NDT-Agile* delivery of working software
Q3	*NDT-Agile* support better communication between the development team and the business	Based on the third statement of the "Agile manifesto": "Customer collaboration over contract negotiation", this statement evaluates how fluid the communication between technical team and business can be when *NDT-Agile* is used
Q4	*NDT-Agile* allows quick and easy adaptation to changes, even late in the development process	Based on the fourth statement of the "Agile manifesto": "Responding to change over following a plan", this statement assesses how easily changes can be incorporated when *NDT-Agile* is used
Q5	*NDT-Agile* promotes best development practices to ensure quality software	Based on the "Agile manifesto" value: "Continuous attention to technical excellence and good design enhances agility", the statement analyzes whether or not best technical practices are incorporated to *NDT-Agile* framework
Q6	*NDT-Agile* set the grounds for all stakeholders' continuous improvement and learning	Based on the "Agile manifesto" value: "At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly", the statement is based on how *NDT-Agile* can support continuous improvement

Table 2 Proposed statements to assess CMMI dimension

Id	Statement	Rationale
Q7	*NDT-Agile* practices allow meeting all CMMI-DEV level 2 specific goals	This statement focuses on the compliance of CMMI-DEV maturity level 2 specific practices
Q8	*NDT-Agile* practices allow meeting all CMMI-DEV level 3 specific goals	This statement focuses on the compliance of CMMI-DEV maturity level 3 specific practices
Q9	*NDT-Agile* practices allow meeting all CMMI-DEV level 4 specific goals	This statement focuses on the compliance of CMMI-DEV maturity level 4 specific practices
Q10	*NDT-Agile* practices allow meeting all CMMI-DEV level 5 specific goals	This statement focuses on the compliance of CMMI-DEV maturity level 5 specific practices
Q11	*NDT-Agile* practices allow meeting all CMMI-DEV generic goals	This statement focuses on the compliance of CMMI-DEV generic practices

Table 3 Proposed statements to assess Web dimension

Id	Statement	Rationale
Q12	*NDT-Agile* proposed practices help to design navigation patterns on Web applications	This statement assesses Web specific characteristic "Complex navigational structure" [21, 22]
Q13	*NDT-Agile* proposed practices support the design of Web application interfaces	This statement assesses Web specific characteristic "Critical interface requirements (such as unknown users or availability, among others)" [21, 22]
Q14	*NDT-Agile* proposed practices help to meet Web applications security constraints	This statement assesses Web specific characteristic "Security aspects" [36]
Q15	*NDT-Agile* proposed practices help to fulfil Web applications maintenance constraints	This statement assesses Web specific characteristic "Increase on maintenance efficiency, avoiding downtimes" [46]
Q16	*NDT-Agile* proposed practices promote delivery of value on Web systems as soon as possible	This statement assesses Web specific characteristic "Delivery as soon as possible" [44, 55, 57]
Q17	*NDT-Agile* proposed practices enable the reduction of "time-to-market" of Web applications	This statement assesses Web specific characteristic "Reduction of time-to-market" [44, 55, 57]
Q18	*NDT-Agile* proposed practices support the adaptation to quick-changing requirements	This statement assesses Web specific characteristic "Adaptation to quick-changing requirements" [44, 55, 57]

Table 4 Proposed statements to assess framework dimension

Id	Statement	Rationale
Q19	*NDT-Agile* provides a coherent approach to manage Web development projects	This statement controls whether there are any un-coherent or contradictory practices included in the proposed framework
Q20	*NDT-Agile* provides complete support to Web development approaches	This statement controls the completion of the framework, ensuring that the main aspects of Web development are covered
Q21	*NDT-Agile* governance allows an effective customization and deployment of the framework	This statement measures the effectiveness of the proposed framework governance

Table 5 Proposed Likert-scale

Value	Meaning
1	Strongly disagree
2	Disagree
3	Neither agree nor disagree
4	Agree
5	Strongly agree

in literature. The proposed statements are displayed in Table 3.

- *Framework* In this case, the statements aim to evaluate the internal coherence and completeness of the framework. The proposed statements are displayed in Table 4.

Initially, we gave the members of the panel a Likert-scale [39] of 5 values in order to assess each of the statements. Table 5 presents these values.

All statements were compiled in a single questionnaire, which was available to the panel members by means of a Google Forms link [31], together with some other questions that helped to establish the expert characterization. Each panel member was not informed of the panel composition in order to guarantee anonymity. Besides, all communications were addressed keeping the different panelists in blind carbon copy. Results presented in the report of each round were also anonymized, removing any reference to the other participants' identity.

4.2 Proposed statistical processing

Based on the questionnaire definition presented in the previous section, the following appropriate statistical techniques were selected and used during the analysis of the gathered information:

- *Descriptive statistics* This analysis is based on the calculation of mean, median, standard deviation and percentage of agreement and disagreement in the given rates for each of the proposed statements. These values help to identify experts' consensus and raters' agreement on each of the proposed statements, highlighting the panel overall opinion about the proposed framework.
- *Chronbach's alpha* Chronbach's alpha helps to measure the questionnaire reliability. It also assesses how the proposed measurement instrument adapts to the evaluated magnitudes and how the amendments to the questionnaire can affect its reliability through the different rounds of the process.

- *Simple Correspondence Analysis (SCA)* In our case, SCA is used to calculate how homogeneous both experts' ratings and questions rates are, and how this homogeneity evolves through the process rounds. It will help to identify whether consensus is reached or not among experts.
- *Kendall's W* Kendall's W aims to measure the level of consensus among raters when evaluating a certain number of items by means of an ordinal scale. Although in several cases Kendall's W is used to measure intra-rater consensus, in our case it is utilized to evaluate agreement among statements ranking through rounds, as a measure of rating stability through process rounds. It is important to note that in our case, stability is monitored at group level and not individually.

Below, we will offer a deep overview of different statistical techniques so as to better understand how they are calculated and applied.

With regard to Chronbach's alpha [13], since its publication in the 1950s, it has established itself as a "defacto" index to assess the extent to which the items of an instrument are correlated [1]. Typically, Cronbach's alpha is considered a measure of a scale reliability, which can be defined as the degree to which the instrument is assessing the needs to be measured. Chronbach's alpha is commonly used to determine whether the defined scale of a multiple Likert-questions survey is reliable or not. This is the case for the designed questionnaire for our proposed Delphi method.

If we measure a quantity being the sum of K elements (for instance, the number of items inside the instrument being assessed), Chronbach's alpha is calculated as follows:

$$\propto = \frac{K\bar{c}}{(\bar{v} + (K-1)\bar{c})} \tag{1}$$

In this formula, K is the number of items, $\bar{v}$ is the average variance of each item and $\bar{c}$ is the average of all covariances among the components of the current sample of people.

According to Welch and Comer [75], measuring reliability by means of Chronbach's alpha involves that the items are measuring the same construct and that they are highly correlated. The more the coefficient is close to 1, the higher the internal consistency of the analyzed items is. George and Mallery [27] propose the recommendations given in Table 6 in order to interpret Chronbach's alpha values.

Moreover, several authors, such as [29, 49] state that a Chronbach's alpha value higher than 0.8 is a reasonable target.

The second statistical technique to be used is the Simple Correspondence Analysis [5, 33]. As mentioned, it is a dimension reduction technique to visualize a multidimensional cloud of points. It aims to measure homogeneity among different variables of categories. In summary, it allows identifying similarities among categories of two variables and its dependencies.

Table 6 Chronbach's alpha interpretation recommendations

Range	Meaning
alpha > 0.9	Excellent
0.9 > alpha > 0.8	Good
0.8 > alpha > 0.7	Acceptable
0.7 > alpha > 0.6	Questionable
0.6 > alpha > 0.5	Poor
alpha < 0.5	Unacceptable

The Simple Correspondence Analysis, as described by Nenadic and Greenacre [48], tries to reduce the dimensionality of a matrix with the objective to display it in a two or three-dimensional space. The starting point of the Simple Correspondence Analysis, following Yelland [78], is a contingency table. As Yelland explains, this type of table appears when it is possible to classify events in two or more set of categories (as it happens in our case, because the given grades belong both to a given Expert and to a given Statement).

The technique is based on the calculation of two magnitudes, mass and inertia, where the former stands for the relative frequency and the latter represents "the total Pearson Chi square for the two-way divided by the total sum" [18]. The use of this technique enables us to represent in two or three dimensions all the relations, expressed by the given grades, between the proposed 21 statements and each of the panel members. By means of this representation, we can visualize the distance between each of the raters and each of the statements, expressed in terms of Chi square. If the distance among raters or statements is short (meaning all experts' ratings and statements' grades are grouped), it will indicate that experts' rating is homogeneous, that is to say, they are behaving the same way and there is consensus among them. It also helps to identify which raters and statements diverge from the overall assessment.

Finally, Kendall's coefficient of concordance or Kendall's W is a statistic that measures the agreement among several raters who assess a set of n items [38]. This means that, if a number of judges are ranking a set of items according to its relevance, Kendall's coefficient of concordance can be obtained for this set of data. If the obtained value is closer to 1, it will imply that the raters have almost reached consensus. On the contrary, if the obtained value is closer to 0, it will mean that there is no agreement among panelists.

If an item i is given the rank $r_{i,j}$ by the rater number j, being n the total number of items to be assessed and m the number of raters, the rank of this item (R_i) is:

$$R_i = \sum_{j=1}^{m} r_{ij} \tag{2}$$

And the mean value of the ranks is:

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{3}$$

The sum of squared deviations (S) will therefore be:

$$S = \sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2 \tag{4}$$

And finally, Kendall's coefficient of concordance can be obtained as follows:

$$W = \frac{12S}{m^2\left(n^3 - n\right)} \tag{5}$$

With regard to Kendall's W interpretation and according to García-Crespo et al. [26], if W is higher than 0.7 or equal, it can be interpreted as a strong consensus; if W is around 0.5, it indicates a moderate consensus; and finally, if W is lower than 0.3, it means weak consensus.

Summarizing, Kendall's W can be seen as a measure of consensus among different raters when using an ordinal scale. In our case, Kendall's W is used to identify stability among rounds, by measuring consensus between statements' ranks through the different rounds of the method. This means that after each round, statements are ranked by its mean forming an ordinal scale and the final ranking is compared among the different rounds. The goal is to measure the extent to which the ranks vary through rounds. If variations are high, consensus will be low and there will be no stability. On the contrary, if variations are low, consensus will be high and there will be stability among the different rounds.

4.3 Panel selection and characterization

As introduced before, the number of members of a Delphi method expert panel and their profiles are not clearly established in the existing literature, since it depends on the analyzed subject. In our case, as it is known, we have tried to analyze the validity of *NDT-Agile* framework from at least three points of view: Agile, CMMI and Web Engineering Thus, ideally, the panel should be composed of experts in the three areas of knowledge. Although this would be the ideal composition of the panel, finding people that are, at the same time, experts in all three fields (taking into account that, as explained before, Agile and CMMI have been seen as mutually exclusive during several years) is highly complicated. This is the reason why we designed a panel composed of experts in one of the fields with, at least, some theoretical knowledge about the other two fields.

Initially, we invited 27 experts to participate in this exercise, receiving a positive answer from 20 of them, who forwarded us answers during the different rounds of the method. Table 7 displays the Agile experience of the Delphi panel members.

As it can be observed, all panel members have at least theoretical experience in Agile methodologies and a majority of them have practical knowledge of these methodologies. The average years of expertise with Agile methodologies of the panel members is 5.37 years. Table 8 displays the CMMI experience of the Delphi panel members.

As Table 8 shows, at least 80% of panel members have either theoretical or practical knowledge of CMMI, finding 35% with practical experience in CMMI-DEV. The average years of expertise with CMMI methodologies of the panel members is 8.75 years. It is also important to mention that 6 panelists have been directly involved in formal CMMI evaluations, known as SCAMPI assessments (an average of 6.83 assessments per member having this type of experience). Table 9 displays the Web experience of the Delphi panel members.

Table 9 displays that all panel members have practical or theoretical experience in Web projects, among whom, those having practical experience constitute the majority. On average, the panel members have 10.63 years of experience in Web projects.

Table 7 Agile experience of panel members

Experience	Number	%
Practical	13	65
Theoretical	7	35
No knowledge/ no experience	0	0

Table 8 CMMI experience of panel members

Experience	Number	%
Practical	7	35
Theoretical	9	45
No knowledge/ no experience	4	20

Table 9 Web experience of panel members

Experience	Number	%
Practical	17	85
Theoretical	3	15
No knowledge/ no experience	0	0

Fig. 7 Type of organization for which panel members are working

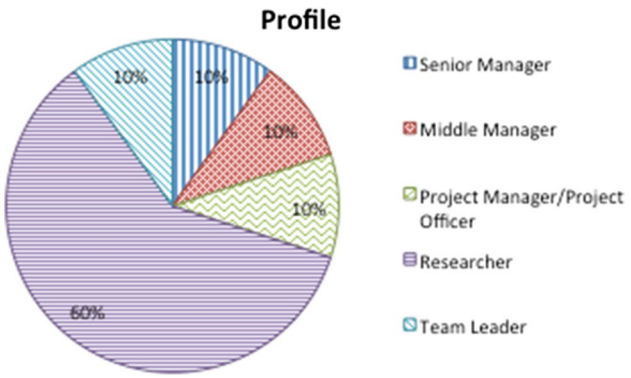


Fig. 8 Panel member work profile

As a general conclusion we can state that the panel is made up of a well-balanced group of people having suitable knowledge of the three fields, including practitioners with "on-the-field" expertise and theoreticians with broad knowledge of the different techniques and methodologies. Figures 7 and 8 display the type of organization for which the different panelists work as well as their work profile.

As these figures show, most of the panel members are researchers working for Universities or Research Institutions (75%), but some of them work either for the private industry or for public institutions. We can find Senior and Middle Manager, Team Leaders or even Project Managers, although 60% of the participants are researchers. This variety of profiles will help us to combine both the project approach, given by project managers working on projects, and the organizational high-level one, given by Middle and Senior Managers more involved in organizational issues.

Finally, it is also interesting to highlight the geographical distribution of the experts, as they come from 8 different countries (Spain, Argentina, Italy, Belgium, France, Ireland, Croatia and Germany). This geographical distribution provides different local interpretations of the methodologies, bringing varied and interesting points of view to the panel.

5 Results

5.1 First round

5.1.1 Descriptive analysis

The first round of the Delphi method was launched on 14th February 2016 and finished on 13th March 2016. As mentioned, 27 invitations were delivered, receiving 20 affirmative responses and evaluations. Table 10 presents the aggregated results obtained during this phase. For each one of the statements, we offer the average assessment of the 20 panelists, the median, the standard deviation, the percentage of agreement (meaning the number of experts providing an assessment as "Strongly agree" or "Agree") and the percentage of disagreement (meaning the number of experts providing an assessment as "Disagree" or "Strongly disagree").

Figure 9 displays the results of the first round in the form of a box and whiskers graph, obtained by means of R [56], using the "ggplot2" [76] and the "reshape2" [77] packages.

As referred, Fig. 9 shows a box-and-whiskers chart. On it, the boxes limits represent the values of the first and third quartile of the given grades to a certain statement. Inside the box, the value of the median is displayed as a point. The graphic also shows as upper and lower whiskers, either the maximum and minimum values obtained during the round or 1.5 times the inter-quartilic range (difference between third and first quartile). If values are going beyond 1.5 times the inter-quartilic range, they are considered atypical values and displayed as simple points.

The previous sections introduced that in the beginning we performed a descriptive analysis of the general gathered data, by means of descriptive statistics. In general, experts do not express strong disagreement with the proposed statements (only one question has 20% of disagreement, ranging the average percentage in most cases from 0 to 10%). For most of the statements (11 out of 21), the median is 4, meaning that "Agree" is the most common value selected by the raters, but it can also be highlighted that there is a significant amount of statements in which "Neither agree nor disagree" is the most selected value (10 out of 21). Regarding the assessment average values, 8 of the statements have a higher value than 3.7, which can point to a clear agreement with the statement content, although in 13 of them the values range from 3.15 to 3.7, which do not indicate a clear agreement with the proposed statement. It can also be observed that some questions include an ambiguous answer (for 9 statements top and down whiskers vary from 5 to 2 and we can even find

Table 10 First round aggregated results

Dimension	Statement	Average	Median	Standard deviation	% Agreement	% Disagreement
AGILE	Q1	4.00	4	0.86	75	5
	Q2	3.85	4	0.93	75	5
	Q3	3.90	4	1.02	70	5
	Q4	3.60	4	1.19	60	20
	Q5	3.55	3.5	1.05	50	10
	Q6	3.70	4	0.66	60	0
CMMI	Q7	3.35	3	0.75	30	5
	Q8	3.35	3	0.75	30	5
	Q9	3.35	3	0.75	30	5
	Q10	3.15	3	0.67	20	10
	Q11	3.20	3	0.77	20	10
WEB	Q12	3.55	4	0.89	55	5
	Q13	3.50	3.5	0.69	50	5
	Q14	3.35	3	0.75	40	10
	Q15	3.45	3	0.83	45	10
	Q16	3.80	4	0.70	75	5
	Q17	3.80	4	0.77	70	5
	Q18	3.70	4	1.08	70	15
FRAMEWORK	Q19	4.15	4	0.67	85	0
	Q20	3.50	4	1.00	60	15
	Q21	3.60	3	0.75	45	0

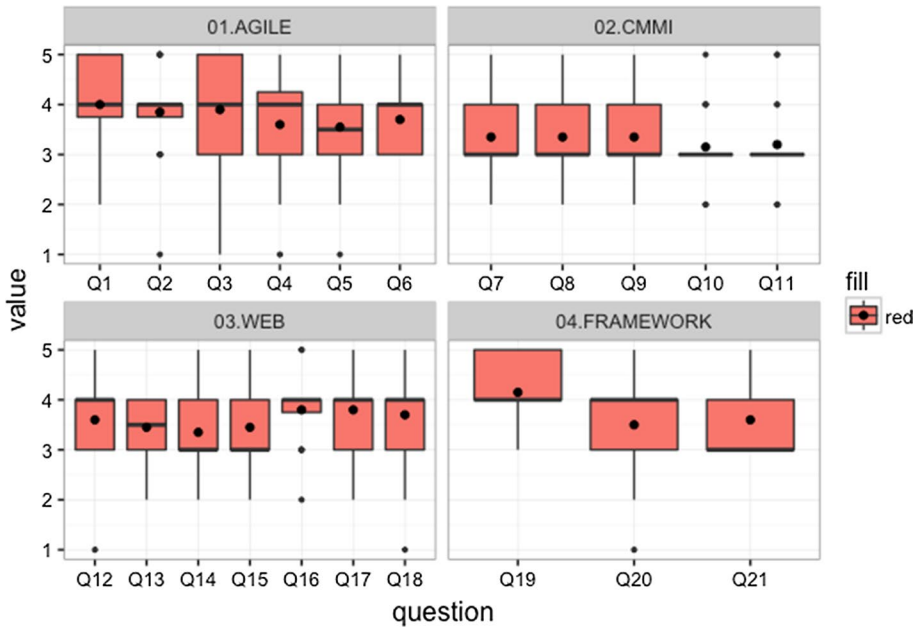


Fig. 9 Results of Delphi first round

extreme atypical values in some of the statements, like Q2, Q4 and Q5), meaning that there are some divergences among experts.

After studying the textual comments provided by the panelists, in order to better understand the initial descriptive analysis, three main concerns are identified:

- Some raters argue that they do not have enough knowledge to assess a particular statement and they choose "Neither agree nor disagree" as a value to express "Don't know/No answer". They suggest that this option should be explicitly included in the questionnaire for future rounds.

- Other raters point out that the provided material, although giving a good overview of the framework, should be complemented with gap analysis and mapping details, so that they can provide a more accurate assessment.
- Finally, several members also claim that some statements are not clear enough and suggest to slightly rephrasing them.

In addition to the overall analysis of the questionnaire, we also carried out a descriptive analysis of each of the proposed dimensions. For each one of the dimensions, we have presented the obtained results as a balloon graph, which was obtained using R, particularly "gplots" package [74]. For this purpose, we have grouped the values representing agreement ("Agree" and "Strongly agree") and those representing disagreement ("Disagree" and "Strongly disagree") per question. Together with the balloon graph, per dimension, we have presented the average means for each one of the statements in the form of a radar chart. Figures 10 and 11 offer the results for the Agile dimension.

Figures 10, 11 and Table 10 show that in this initial round, the experts' agreement with the Agile dimension statements seems to be high for 4 out of 6 statements, since the average is above 3.7, the median is 4 and more than 60% of raters agree with them. The main concerns of the experts at this stage were linked to the ability to incorporate changes, in case of a more structured process is in place, and to how best practices are incorporated in the framework. Figures 12 and 13 display the results concerning the CMMI dimension.

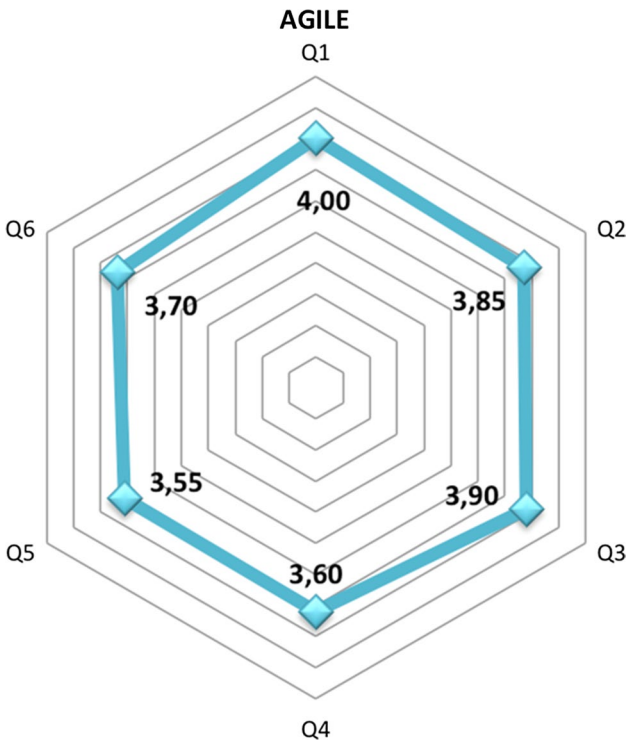


Fig. 11 Agile dimension—first round results (II)

The information provided by Figs. 12 and 13 together with the results in Table 10 confirm that in this initial round,

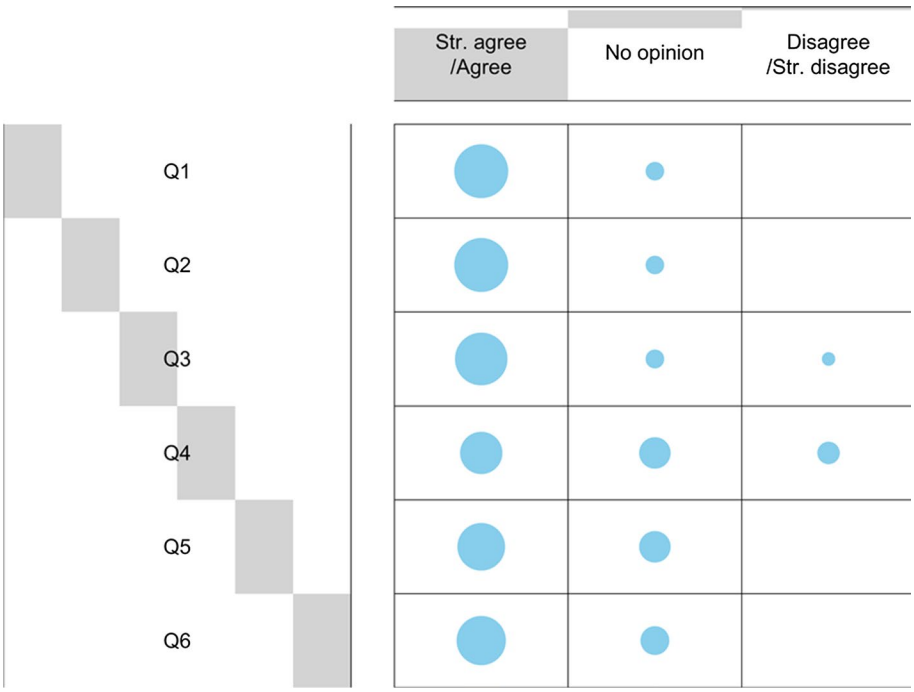


Fig. 10 Agile dimension—first round results (I)

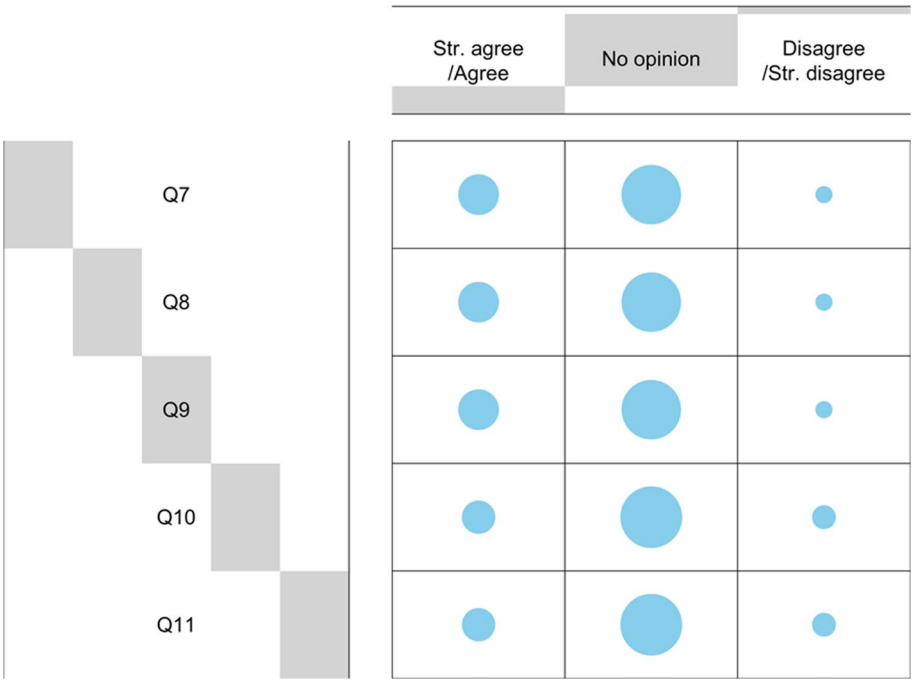


Fig. 12 CMMI dimension—first round results (I)

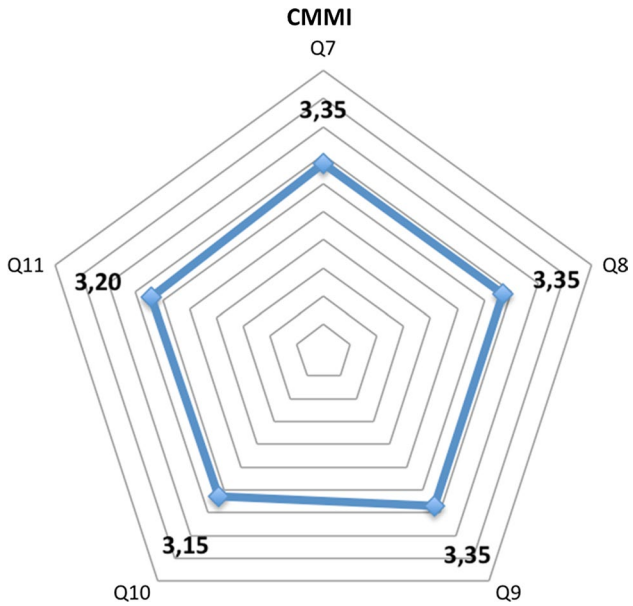


Fig. 13 CMMI dimension—first round results (II)

experts do not express a clear opinion regarding the CMMI dimension, as in most of the cases "Neither agree nor disagree" is the selected option. A detailed analysis of the textual comments indicate that two elements affect this particular result: some of the raters aim to express their lack of knowledge or willingness to answer and some of them need to have more information regarding *NDT-Agile* versus CMMI mapping. Next, we will analyze the Web dimension with the results displayed in Figs. 14 and 15.

From the information Figs. 14 and 15 and Table 10 provide, it can be inferred that in this initial round, experts seem to express a clear agreement for three of the statements (from Q16 to Q18), in which the average is higher than 3.7, the median is 4 and the percentage of agreement is around 70%. There is no clear opinion of the panel in relation to the other four statements. At this stage, main concerns are linked to support navigation patterns and user interfaces, together with security and maintenance aspects. Finally, Figs. 16 and 17 present the results for the framework dimension.

From the presented results, we can notice that in this initial round, experts seem to openly agree on one of the statements (Q19), in which the average is higher than 4, the median is 4 and the percentage of agreement is 85%. The panel does not have a clear opinion about the other two statements. At this stage, main concerns are linked to complete support to Web development projects and proposed governance.

5.1.2 Homogeneity and concordance analysis

The homogeneity and concordance analysis followed the descriptive one by measuring reliability. For this purpose, Chronbach's alpha was calculated and the following value was obtained:

$\propto = 0.8791989$

The calculations of the Chronbach's alpha were computed using the "psy" package [23] of R [56]. The obtained value, following George and Mallery [27], indicated a good reliability of the proposed questionnaire in order to indirectly

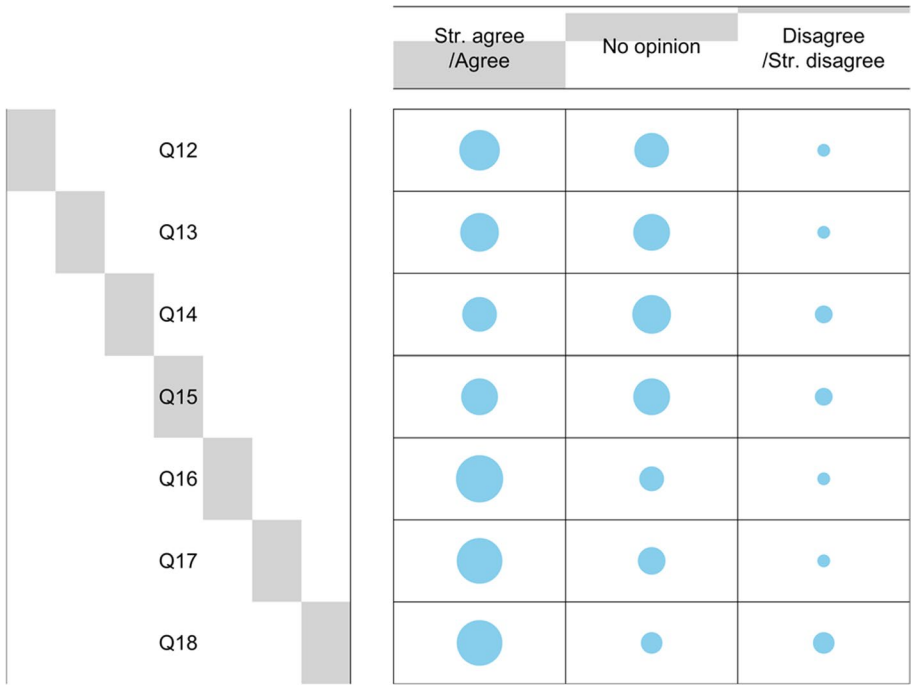


Fig. 14 Web dimension—first round results (I)

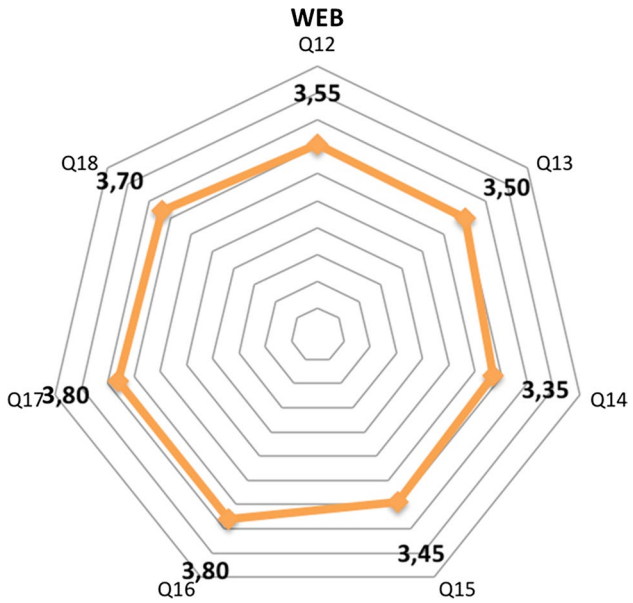


Fig. 15 Web dimension—first round results (II)

measure the desired magnitude. Finally, we performed a Simple Correspondence Analysis for the full questionnaire, obtaining the results presented in Fig. 18. SCA was conducted by means of R package "ca" [48].

As Fig. 18 shows, most experts and statements are grouped around the central point, with the exception of some of them (such as experts 4, 5 and 12 and statements 4, 8 and 14). On the one hand, from the experts' point of view, this result implies that most of them (experts 1, 7, 8, 19 or 20, for instance) are behaving in a homogeneous way (i.e. their assessments are more or less similar). On the other hand, from the viewpoint of the statements, we find the same situation with statements rates grouped around statements 3, 9, 12 or 20. To sum up, this test shows that there is certain homogeneity in the provided assessments, which can confirm a certain level of consensus among raters.

5.1.3 Round conclusions

As a main conclusion of the descriptive, homogeneity and concordance analyses, we can state that the former verifies that there are still elements to improve and clarify, regarding the experts' expressed opinion, despite suitability of the proposed questionnaire and certain homogeneity in the provided answers, as our reliability and homogeneity tests show. Thus, a second round of the Delphi method was proposed to the panel, providing the members with an anonymized version of the comments and statistical data presented in Table 10. For this second round the questionnaire was also modified taking into account the experts' comments, as the next section will describe in detail.

5.2 Second round

5.2.1 Questionnaire redesign

As mentioned above in the previous section, a second round was conducted to clarify and refine the results gathered during the first round of the Delphi method. The first action consisted in modifying the initial questionnaire in two ways:

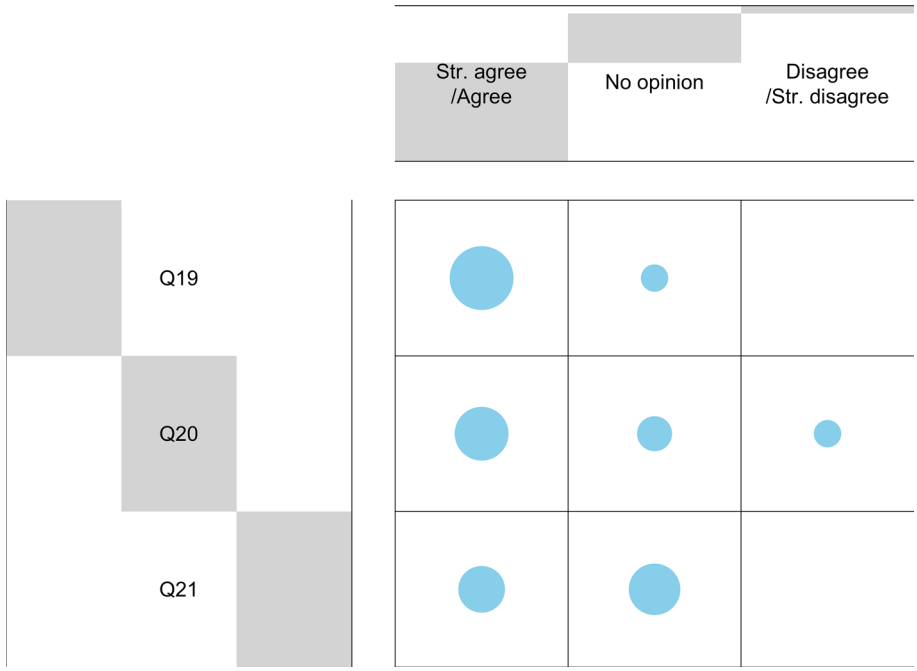


Fig. 16 Framework dimension—first round results (I)

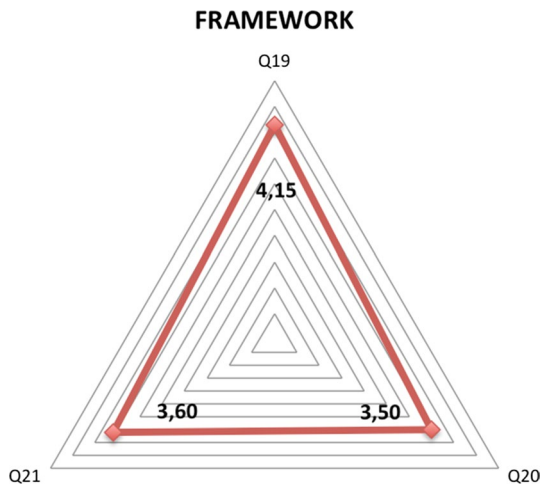


Fig. 17 Framework dimension—first round results (II)

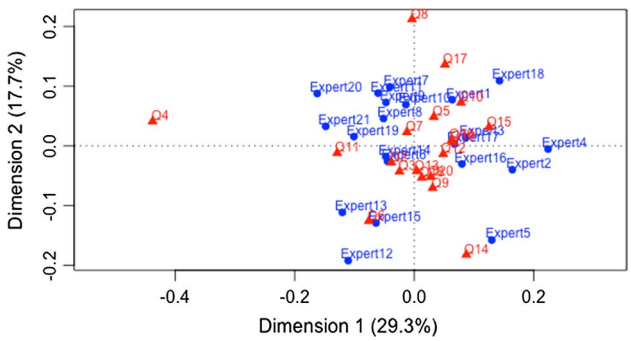


Fig. 18 Simple Correspondence Analysis—first round

- Six of the statements were rephrased, taking into consideration the experts' recommendations (see Table 11 for details).
- An option "Don't know/No answer" was included, to allow experts to opt-out when assessing a question.

Tables 11 and 12 show the initial and the modified questionnaire and the new scale.

5.2.2 Descriptive analysis

The second round was launched on 10th April 2016 and it finished on 24th May 2016. 19 replies out of 20 were received, either modifying or confirming the initial rate. For the remaining one, the assessment of the first round was kept for the second and following rounds. A new message was sent to the experts for this second round, including the detailed gap analysis coming from the previously identified works [68, 71, 72] and the mapping between *NDT-Agile* and CMMI goals, in order to help them to provide a more accurate assessment. Table 13 presents the aggregated results obtained during this phase. For each of the statements, we present the average assessment of the 20 panelists, the median, the standard deviation, the percentage of agreement (meaning the number of experts providing an assessment being "Strongly agree" or "Agree"), the percentage of disagreement (i.e. the number of experts providing an assessment being "Disagree" or "Strongly disagree") and the percentage of experts providing different answers to "Don't know/No answer".

Table 11 Modified questionnaire for round

Id	Initial statement	Modified statement
Q1	*NDT-Agile* enhances personal interactions and communication among team members	No changes
Q2	*NDT-Agile* promotes early delivery of working software to business	No changes
Q3	*NDT-Agile* supports better communication between the development team and the business	No changes
Q4	*NDT-Agile* allows quick and easy adaptation to changes, even late in the development process	No changes
Q5	*NDT-Agile* promotes best development practices to ensure quality software	*NDT-Agile* promotes good-enough software engineering practices to ensure software quality
Q6	*NDT-Agile* sets the grounds for continuous improvement and learning of all stakeholders	*NDT-Agile* sets the grounds for continuous improvement and learning of the key project stakeholders (development team and business representatives)
Q7	*NDT-Agile* practices allow meeting all CMMI-DEV level 2 specific goals	No changes
Q8	*NDT-Agile* practices allow meeting all CMMI-DEV level 3 specific goals	No changes
Q9	*NDT-Agile* practices allow meeting all CMMI-DEV level 4 specific goals	No changes
Q10	*NDT-Agile* practices allow meeting all CMMI-DEV level 5 specific goals	No changes
Q11	*NDT-Agile* practices allow meeting all CMMI-DEV generic goals	No changes
Q12	*NDT-Agile* proposed practices help to design navigation patterns on Web applications	No changes
Q13	*NDT-Agile* proposed practices support the design of Web application interfaces	No changes
Q14	*NDT-Agile* proposed practices help to meet Web applications security constraints	No changes
Q15	*NDT-Agile* proposed practices help to meet Web applications maintenance constraints	No changes
Q16	*NDT-Agile* proposed practices promote delivery of value on Web systems as soon as possible	*NDT-Agile* proposed practices promote early delivery of value on Web systems
Q17	*NDT-Agile* proposed practices enable the reduction of "time-to-market" of Web applications	*NDT-Agile* proposed practices enable the reduction of "time-to-market" of Web applications compared to other Web Engineering approaches (such as standard NDT)
Q18	*NDT-Agile* proposed practices support the adaptation to quick-changing requirements	No changes
Q19	*NDT-Agile* provides a coherent approach to manage Web development projects	No changes
Q20	*NDT-Agile* provides complete support to Web development approach	*NDT-Agile* provides support to some of the most relevant elements of Web development projects
Q21	*NDT-Agile* governance allows an effective, customization and deployment of the framework	*NDT-Agile* governance will help to achieve an effective deployment and customization of the framework

Table 12 Modified Likert-scale

Value	Meaning
1	Strongly disagree
2	Disagree
3	Neither agree nor disagree
4	Agree
5	Strongly agree
N/A	Don't know/no answer

As for the previous round, Fig. 19 displays the results in the form of a box and whiskers graph.

Both Table 13 and Fig. 19 let us know that experts do not disagree in general with the proposed statements (only two questions have around 15% of disagreement, ranging it in most cases from 0 to 10%). The initial low disagreement from the first round and the clarifications provided either by some other expert's comments or by the supplementary provided documentation might explain the reduction of disagreement. The median is 4 for almost all

Table 13 Aggregated results of second round

Dimension	Statement	Average	Median	Standard deviation	% Agreement	% Disagreement	% Answers
Agile	Q1	4.21	4	0.63	89.47	0	95% (19/20)
	Q2	4.05	4	0.60	85	0	100% (20/20)
	Q3	4.00	4	0.79	80	5	100% (20/20)
	Q4	3.55	3.50	1.00	50	15	100% (20/20)
	Q5	4.10	4	0.85	70	0	100% (20/20)
	Q6	3.95	4	0.69	75	0	100% (20/20)
CMMI	Q7	4.07	4	0.83	71.43	0	70% (14/20)
	Q8	4.08	4	0.64	84.62	0	65% (13/20)
	Q9	3.86	4	0.66	71.43	0	70% (14/20)
	Q10	3.54	4	0.78	53.85	7.69	65% (13/20)
	Q11	4.00	4	0.85	66.67	0	60% (12/20)
Web	Q12	3.83	4	0.62	72.22	0	90% (18/20)
	Q13	3.74	4	0.56	63.16	0	95% (19/20)
	Q14	3.50	4	0.79	50	11.11	90% (18/20)
	Q15	3.67	4	0.84	61.11	5.56	90% (18/20)
	Q16	3.94	4	0.64	83.33	0	90% (18/20)
	Q17	4.11	4	0.74	89.47	5.26	95% (19/20)
	Q18	3.74	4	1.05	78.95	15.79	95% (19/20)
Framework	Q19	4.28	4	0.57	94.44	0	90% (18/20)
	Q20	3.75	4	1.06	75	12.50	80% (16/20)
	Q21	3.53	3	0.62	47.06	0	90% (17/20)

of the statements (19 out of 21), meaning that "Agree" is the most common value that raters select. This might be explained either because the experts are convinced of the other panelists' arguments or because they have selected the option "Don't know/No answer" since they feel not having enough background to respond. Regarding the assessment average values, 16 of the statements have a value higher than 3.7, which can point to a clear agreement with the content of the statement. In the remaining 5, the values range from 3.5 to 3.7, which might indicate slight agreement with the proposed statement. It can be deduced that dispersion has been clearly reduced in most of the questions (only in four questions the top and down whiskers vary from 5 to 2, and we find only two questions, Q18 and Q20, in which anomalous values ranging from 1 to 5 appears).

In order to better understand the initial descriptive analysis, we have studied the textual comments provided by the panelists, in which two main elements have been identified:

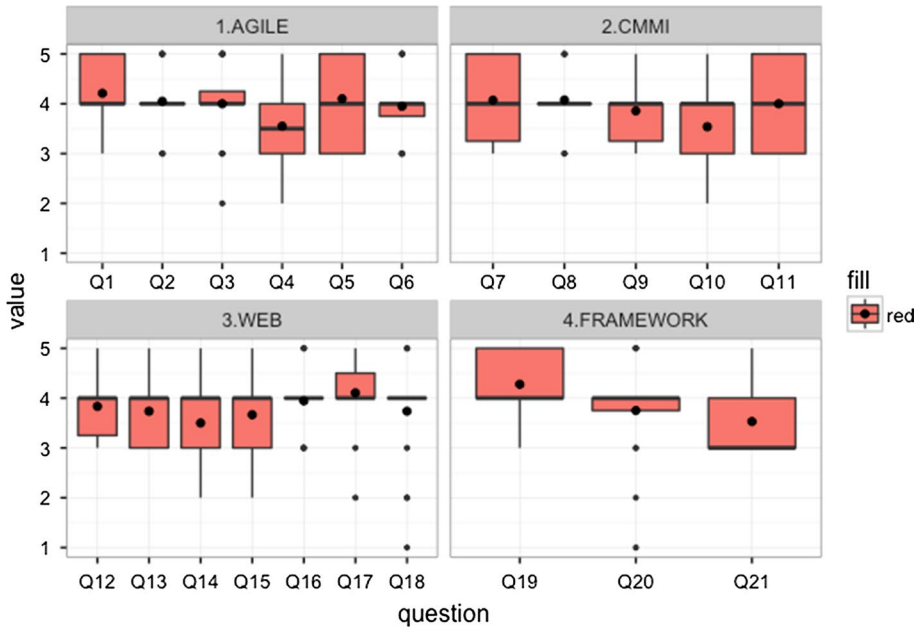


Fig. 19 Results of Delphi second round

- Experts agreed on the modifications performed in the questionnaire.
- Experts expressed that the provided documentation has helped them to get a general idea and give a more accurate grade.

It is also important to mention that the number of comments received during this round is significantly lower than those received during the first round. As in the first round, we will present the performed descriptive analysis per dimension, starting with the Agile one, presented in Figs. 20 and 21.

Figures show that in this second round, the panel agreement with the Agile dimension statements seems to be very high for 5 out of 6 statements. For them, the average assessment is above 3.95, the median is 4 and more than 70% of raters agree with the statement. Besides, the percentage of disagreement is below 5% in 5 of the 6 statements. The rephrase for Q5 and Q6 clearly facilitates the assessment, as the new results show. At this stage, there is no clear

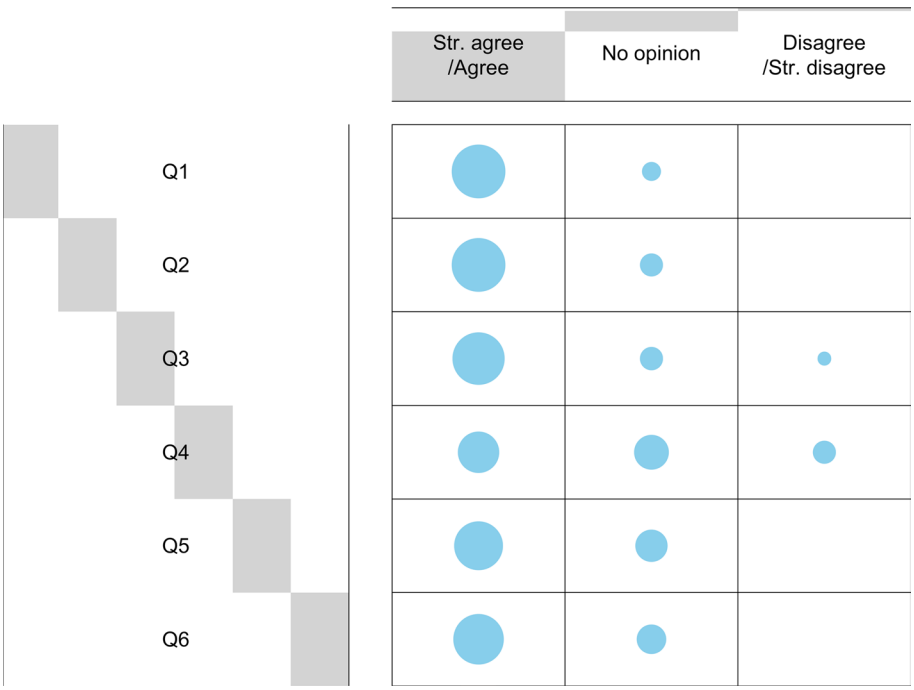


Fig. 20 Agile dimension—second round results (I)

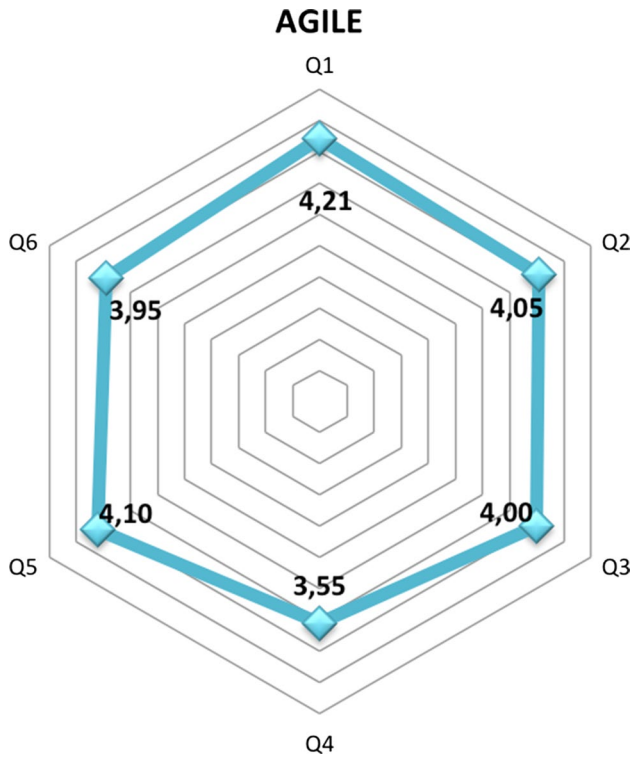


Fig. 21 Agile dimension—second round results (II)

agreement only for statement Q4, regarding the easy adaptation to changes. Nevertheless, the obtained values (average: 3.55 and median: 3.5) might represent slight agreement or no clear validation, but never disagreement. Next, we will present the results of CMMI in Figs. 22 and 23.

We can observe that in this second round, experts have moved from an initial not well-established opinion to a broad consensus on the proposed statements (4 out of the 5 statements have an average assessment higher than 3.8, a median of 4, and a percentage of agreement higher than 65%). The main reasoning behind this change might be, on the one hand, the extra documentation provided to the raters during the second round, which might have clarified some issues and, on the other hand, the possibility for those not having enough knowledge to assess some statements by choosing "Don't know/No answer" option to respond. This has been clearly the case for most of those experts who expressed that they neither have theoretical nor practical knowledge in CMMI. In most of the cases (3 out of the 4 experts without experience) this was the option chosen in the second round, whereas in the first round they selected "No opinion" to express their lack of knowledge, affecting statistical data processing.

The main question still remaining is linked to the feasibility of achieving CMMI-DEV level 5 goals (statement Q10), as they are the most challenging ones. Some experts commented on the fact that without an empirical validation they were not able express a clear agreement with the statement. The following dimension is Web, represented in Figs. 24 and 25.

The presented figures highlight that in this second round experts' agreement have increased, finding 5 out of 7 statements in which the average is higher than 3.7, the median is 4 and the percentage of agreement is higher than 60%. Again, the changes proposed for Q16 and Q17 help to clarify raters' uncertainties. There is slight agreement among

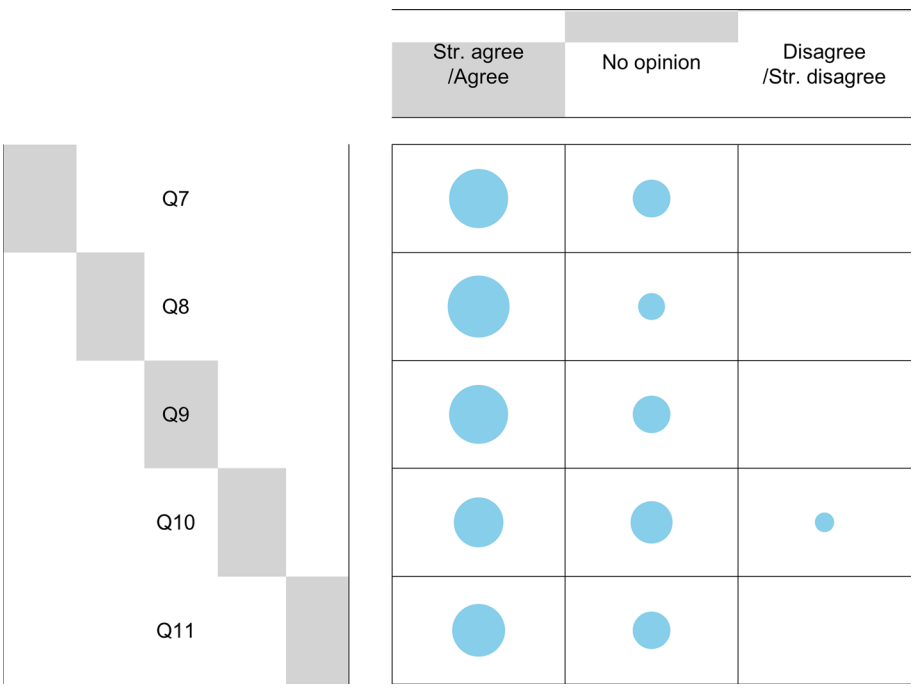


Fig. 22 CMMI dimension—second round results (I)

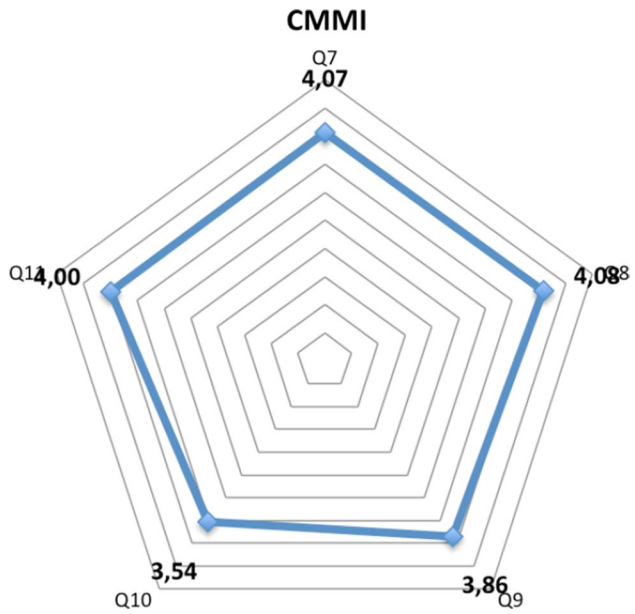


Fig. 23 CMMI dimension—second round results (II)

raters (average between 3.5 and 3.7 and median of 4) on the remaining two statements (Q14 and Q15). At this point, main concerns are linked to how security and maintenance aspects are covered for Web systems. Finally, the Framework dimension is presented in Figs. 26 and 27.

As a main conclusion for this dimension, it is worth highlighting that experts express a clear agreement on two of the statements (Q19 and Q20), in which the average is higher than 3.7, the median is 4 and the percentage of agreement is above 75%. There is no clear opinion from the panel for the remaining statement. Although the rephrase of Q20 seems to have helped raters to understand the idea, there is still a main concern on how the governance will work as well as whether it will enable or not deployment, adaptation and improvement of the framework. Some experts argued that without an empirical validation they were not able to agree on the last statement (Q21).

Finally, we performed a descriptive analysis for stability. Table 14 shows the changes on statements rate between first and second round.

From the presented data, we can notice that more than 20 of the statements have significantly varied their assessment (more than 10%), and more than 20% have slightly changed (between 5 and 10%). This is a clear indicator that stability has not yet been reached.

Finally, it should be put forward that the same analysis was conducted excluding the results provided by the expert who declined to participate in the second round, in order to check if our decision of keeping his replies was affecting significantly the obtained conclusions. After that, it was verified that the results obtained, both general and per dimension would have remained almost the same, with only small variations in mean, median and standard deviation.

5.3 Homogeneity and concordance analysis

Once the descriptive analysis has taken place, the homogeneity and concordance analysis starts with calculating

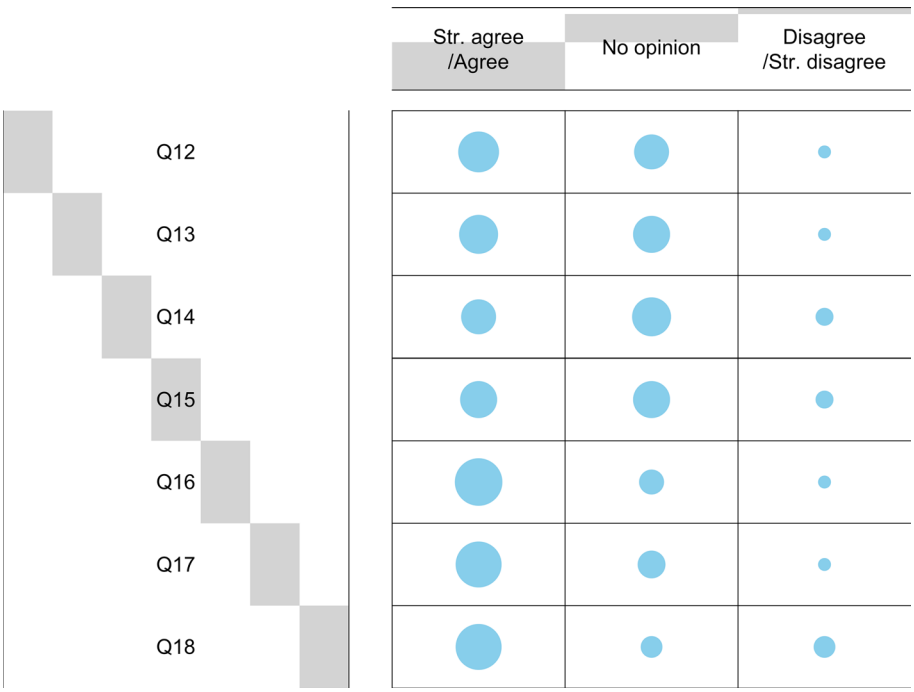


Fig. 24 Web dimension—second round results (I)

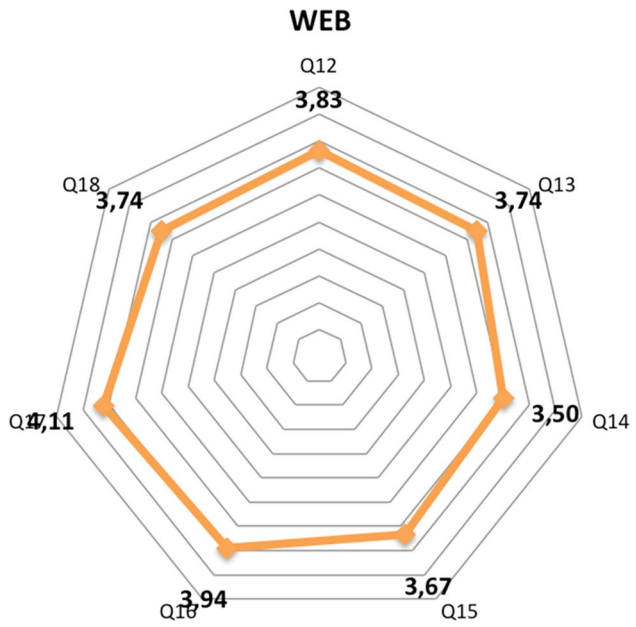


Fig. 25 Web dimension—second round results (II)

Chronbach's alpha with the aim to assess the reliability of the questionnaire, by using R for its computation:

$\propto = 0.8133207$

To calculate Chronbach's alpha, "Don't know/No answer" values provided by the experts are replaced by 0. The obtained value, following [27], indicates a good reliability of the proposed questionnaire in order to indirectly measure the desired magnitude, as it was in the case of the previous round. The next step is measuring experts' and statements' assessment homogeneity, by means of a Simple Correspondence Analysis, whose results are presented in Fig. 28.

As Fig. 28 displays, the distance among points representing experts' and statements' assessments is significantly reduced (it can be clearly seen, for example, in the case of statement 4 or experts 8 and 12). This means that homogeneity among the rates increases. Consequently, they are behaving more similarly than in the previous round. It can be interpreted that consensus among experts has increased.

Finally, we measured stability among the different rounds by means of Kendall's W. As previously explained, Kendall's W is a statistic that measure agreement among raters on ordinal scales. We have categorized (in descending order) the average statements rates in order to be able to use it. This helps us to build an ordinal scale per round. Once the ordinal scale is built both for round 1 and round 2, Kendall's W is calculated with the aim to measure the "degree of agreement" among ranks obtained at the end of each round. If Kendall's W expresses agreement, this will mean that obtained ranks are basically the same, thus we achieve stability. If Kendall's W expresses low or no agreement, stability has not yet been reached and a following round of the Delphi method will be needed. Table 15 shows the ranks obtained for round 1 and round 2.

Kendall's W is calculated by means of "irr" package [25] from R obtaining the following results:

$W = 0.558$

The value of W, between 0.5 and 0.7, indicates a moderate consensus reached between both ranks, which can be interpreted as a "non-stable" situation.

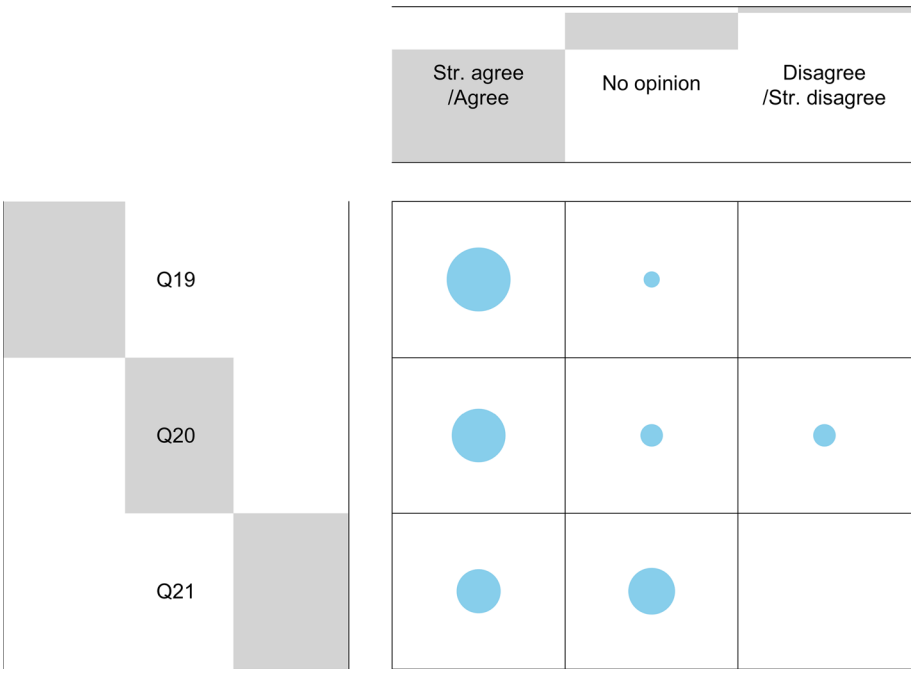


Fig. 26 Framework dimension—second round results (I)

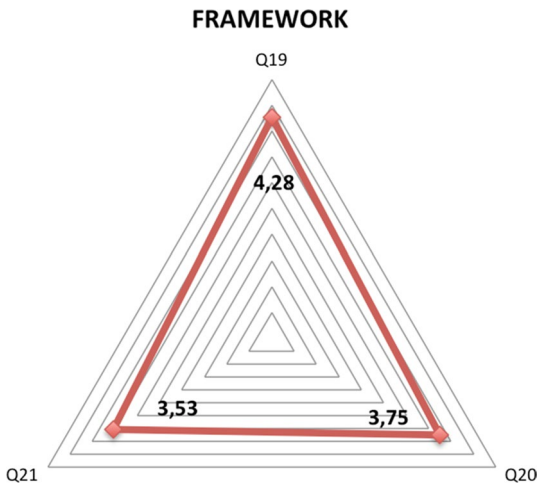


Fig. 27 Framework dimension—second round results (II)

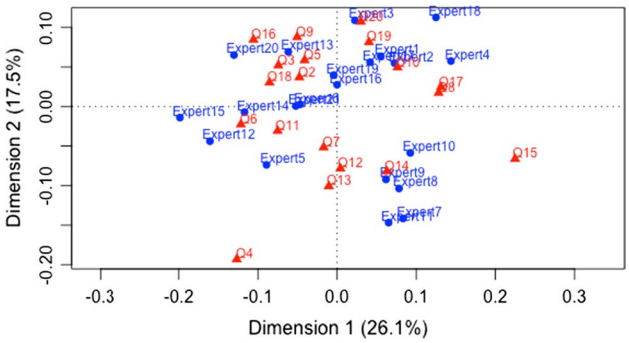


Fig. 28 Simple Correspondence Analysis—second round

Table 15 Average ranks after rounds 1 and 2

Dimension	Statement	Rank (round 1)	Rank (round 2)
AGILE	Q1	4	2
	Q2	16	7
	Q3	17	8
	Q4	8	18
	Q5	9	4
	Q6	21	10
CMMI	Q7	7	6
	Q8	11	5
	Q9	12	12
	Q10	1	19
	Q11	18	9
WEB	Q12	13	13
	Q13	15	15
	Q14	10	21
	Q15	6	17
	Q16	20	11
	Q17	3	3
	Q18	5	16
FRAMEWORK	Q19	2	1
	Q20	19	14
	Q21	14	20

5.3.1 Round conclusions

As a main conclusion of this round, we can state that most of the uncertainties expressed by the raters during the initial round have been solved. This is reflected in the fact that there is agreement on most of the presented statements (16 out of 21), and only some reluctances or doubts in relation to certain aspects of the model. The homogeneity and concordance statistical analysis also shows that the proposed questionnaire is still able to measure the desired magnitudes and that there is a fair level of consensus (shown by the increase in experts and statements homogeneity) among the panel regarding the whole questionnaire.

Nevertheless, and as discussed in previous sections, there are two main dimensions to consider in order to conclude a Delphi method; one is consensus, that apparently is reached, as the descriptive, homogeneity and concordance analysis of the gathered data show; and the other one is stability. Regarding this last dimension, we can clearly state from the conclusions of both analyses that stability is not yet reached, and that a new round is needed.

5.4 Third round

5.4.1 Analysis

The third round of the Delphi method was launched on 31st May 2016 and lasted until 19th June 2106. For this purpose, a message was sent to all panel members, attaching a report that summarized the second round conclusions. That report contained each statements average assessment, mean, standard deviation and percentage of agreement and disagreement. Moreover, all comments received in the previous round were also included in the report. During that round, experts had

Table 14 Changes on statements' assessment between first and second round

Magnitude	Number	%
Statements with average change higher than 10% (included)	5	23.81
Statements with average change between 10% and 5% (included)	4	23.81
Statements with average change lower than 5%	11	53.38

the chance to ratify their previous assessments, if appropriate, or modify their previous reply. Table 16 is summarizing the experts' decision during the third round.

As it can be seen, all but two experts ratified their previous assessment and only two of them did not respond; therefore no significant change happens on the results presented in the previous section. Based on figures presented in Tables 16, 17 and Fig. 29 display the results of the third round. Table 17 highlights the modified results in bold.

Both Table 17 and Fig. 29 show that the main conclusions of the previous round are still valid, as in all modified statements (Q1, Q2, Q3, Q4, Q19 and Q20) the mean and median increase and standard deviation decreases, meaning that there is even more agreement on those statements. As our conclusions regarding agreement do not differ from the ones of previous rounds, we will skip the analysis to avoid unnecessary repetitions, and we will assess stability both by descriptive and concordance means. Table 18 shows the change on statements' assessment between the second and third round.

According to the presented data, almost 30% of questions undergo minor changes, showing that stability might be reached. As in the previous round, we also use Kendall's W to measure if stability is reached, by ranking the average values of the statements for second and third round, as Table 19 shows.

Table 16 Experts' decision during the third round

Expert	Decision	Assessments changed
1	Ratify previous assessment	0
2	Ratify previous assessment	0
3	Ratify previous assessment	0
4	Ratify previous assessment	0
5	Ratify previous assessment	0
6	Ratify previous assessment	0
7	Ratify previous assessment	0
8	Ratify previous assessment	0
9	Ratify previous assessment	0
10	Ratify previous assessment	0
11	No answer	0
12	Ratify previous assessment	0
13	Ratify previous assessment	0
14	Ratify previous assessment	0
15	Ratify previous assessment	0
16	Ratify previous assessment	0
17	Ratify previous assessment	0
18	Change assessment	6
19	Ratify previous assessment	0
20	Ratify previous assessment	0

Kendall's W is calculated using the "irr" package [25] from R obtaining the following results:

$W = 0.996$

The value of W close to 1 involves a high consensus reached between both ranks, which can be interpreted as the ranks after both rounds, being almost the same and representing a "stable" situation. Therefore, we can check that the results of the third round also cope with the second desired condition to finalize the Delphi method: stability. Once consensus and stability have been reached (as described before through the descriptive, homogeneity and concordance analysis of data), the Delphi method can be considered as concluded.

5.4.2 Conclusions

As a main result of using the method, we can state that the panel reached high level of agreement on the overall proposed statements, as Table 20 and Fig. 30 show:

To obtain data presented in Table 20 and Fig. 30, we have considered that strong agreement is reached in those statements having an average assessment equal or higher than 3.7, a median of 4 and at least 60% of the raters' value scoring 4 or 5 ("Agree" or "Strongly agree"), with a minimum of 12 experts giving an opinion. We have also confirmed that slight agreement is reached in those statements having an average assessment equal or higher than 3.5 and lower than 3.7, with a median equal or higher than 3.5 and at least 45% of raters' value scoring 4 or 5 ("Agree" or "Strongly agree"), with a minimum of 12 experts giving an opinion.

As it can be seen, in 16 out of the 21 statements strong agreement is reached, thus the overall opinion of the panel is clearly positive in that regard. In the case of the remaining 5 statements, slight agreement is reached, meaning that, although the panel has not reached a consensus on them, their opinion is not as clearly positive as in the case of the previous ones.

If we analyze the results dimension by dimension, in the Agile dimension the agreement is quite high. There is strong agreement in 5 of the 6 statements, with high values as average statements' assessment. The only statement that poses some doubts is Q4, which relates to ability to quickly adapt to changes. Some experts expressed their concerns regarding the fact of using a Project Charter to formalize the project and models and metamodels to develop Web features might somehow slow the development process. The validation of these aspects and the model improvement will be matter of further work and research. In the case of CMMI dimension, there is strong agreement of the panel on 80% of the statements (4 out of 5). Only Q10, related to the possibility of achieving all CMMI-level 5 goals is not fully agreed. In this case, experts commented on the fact that being these

Table 17 Aggregated results of third round

Dimension	Statement	Average	Median	Standard deviation	% Agreement	% Disagreement	% Answers
Agile	**Q1**	**4.20**	**4**	**0.62**	**90**	**0**	**100% (20/20)**
	Q2	**4.10**	**4**	**0.55**	**90**	**0**	**100% (20/20)**
	Q3	**4.05**	**4**	**0.76**	**85**	**5**	**100% (20/20)**
	Q4	**3.60**	**4**	**0.99**	**55**	**15**	**100% (20/20)**
	Q5	4.10	4	0.85	70	0	100% (20/20)
	Q6	3.95	4	0.69	75	0	100% (20/20)
CMMI	Q7	4.07	4	0.83	71.43	0	70% (14/20)
	Q8	4.08	4	0.64	84.62	0	65% (13/20)
	Q9	3.86	4	0.66	71.43	0	70% (14/20)
	Q10	3.54	4	0.78	53.85	7.69	65% (13/20)
	Q11	4.00	4	0.85	66.67	0	60% (12/20)
Web	Q12	3.83	4	0.62	72.22	0	90% (18/20)
	Q13	3.74	4	0.56	63.16	0	95% (19/20)
	Q14	3.50	4	0.79	50	11.11	90% (18/20)
	Q15	3.67	4	0.84	61.11	5.56	90% (18/20)
	Q16	3.94	4	0.64	83.33	0	90% (18/20)
	Q17	4.11	4	0.74	89.47	5.26	95% (19/20)
	Q18	3.74	4	1.05	78.95	15.79	95% (19/20)
Framework	**Q19**	**4.26**	**4**	**0.56**	**94.74**	**0**	**95% (19/20)**
	Q20	**3.76**	**4**	**1.03**	**76.47**	**12.50**	**85% (17/20)**
	Q21	3.53	3	0.62	47.06	0	90% (17/20)

practices very challenging, they preferred to supervise some examples or case studies. Again, either a self-assessment or a formal SCAMPI might stay as future lines of work.

As Web dimension concerns, we find strong agreement in 5 out of the 7 statements (approximately 71.5%). The two remaining statements are Q14 and Q15, linked to security and maintenance of Web applications. As both aspects of the model rely on NDT, maybe as a future line of work, a revision of the model might include specific Agile practices to tackle both of them. Finally, strong agreement is reached for the Framework dimension on two of the three statements, with doubts only for Q21, related to the effectiveness of the governance model. Experts suggested that to be able to provide a better opinion, they needed some real examples. As mentioned, future case studies might help to clarify this issue.

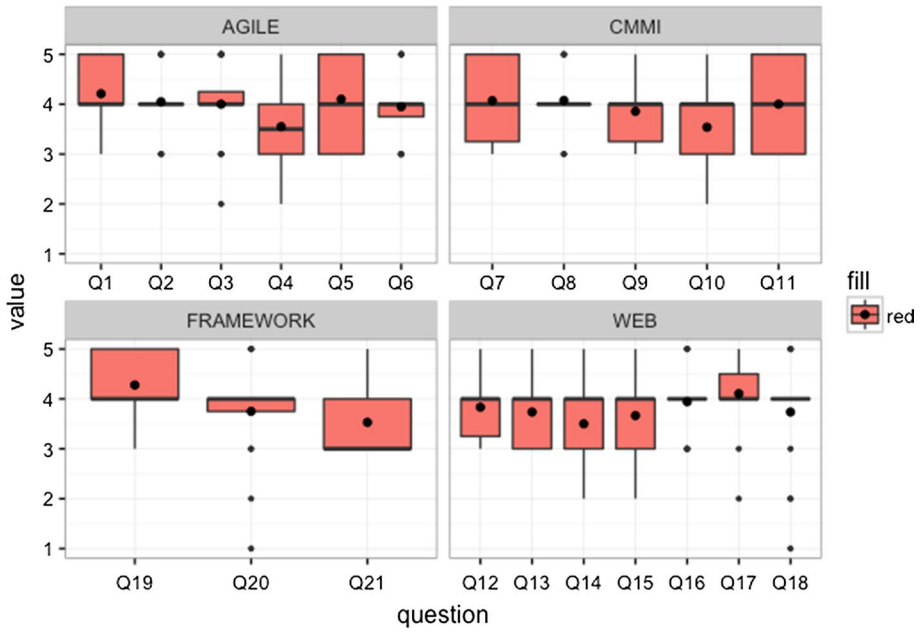


Fig. 29 Results of Delphi third round

Table 18 Changes on statements' assessment between second and third round

Magnitude	Number	%
Statements with average change higher than 10% (included)	0	0
Statements with average change between 10% and 5% (included)	0	0
Statements with average change lower than 5%	6	28.57

6 Conclusions, limitations and future work

The previous section has presented the results of our Delphi method. This one will try to extract the main research conclusions, linking them to the previously exposed research questions and goals: to apply expert judgment methods to Web Engineering-related issues and validate the *NDT-Agile* framework.

As the latter concerns, and as an initial and general conclusion of the experiment, we can state that our proposed Delphi-based consensus-making method was useful to help the panel to reach a shared opinion, in most cases favorable, regarding *NDT-Agile*. Besides, the Delphi method encouraged the panel to learn more on the topic and clarify misunderstandings or uncertainties during the different rounds. By means of the textual comments, the experts were able to communicate any missing element or uncertainty, which would be solved in the next round. Moreover, these comments, shared in an anonymized way with the rest of experts by means of the different rounds' reports, helped everyone to achieve certain level of agreement, as it can be noticed in the convergence of results after the first round (extreme values were reduced, and experts' opinion concentrated mainly on the statement's average value).

Figure 31 shows, in the form of a box and whiskers graph, the evolution of the obtained results.

In Fig. 31, it can be observed how the size of boxes and the length of whiskers tend to reduce, meaning that values tend to converge (distance between first and third quartile and distance between maximum and minimum values are smaller). Besides, the graph shows how the median increases through the different rounds of the method for the majority of the statements.

The expert feedback also helped to improve the questionnaire, suggesting rephrases of statements and the possibility of opting-out whenever the expert's knowledge on the particular field was not sufficient. The analysis of reliability of the questionnaire by means of Chronbach's alpha showed that it was appropriate along the Delphi method, even after changes.

The use of Simple Correspondence Analysis proved to be useful to identify if a certain level of consensus was reached, but it is important to note that it has to be combined with a descriptive analysis to better understand the panel's expressed opinion. If not, experts expressing a common "Don't know" opinion will not be suitably interpreted.

Finally, the use of Kendall's W confirmed its utility in the use of quantitative elements to evaluate assessment

Table 19 Average ranks after rounds 2 and 3

Dimension	Statement	Rank (round 2)	Rank (round 3)
Agile	Q1	2	2
	Q2	7	4
	Q3	8	8
	Q4	18	18
	Q5	4	5
	Q6	10	10
CMMI	Q7	6	7
	Q8	5	6
	Q9	12	12
	Q10	19	19
	Q11	9	9
Web	Q12	13	13
	Q13	15	15
	Q14	21	21
	Q15	17	17
	Q16	11	11
	Q17	3	3
	Q18	16	16
Framework	Q19	1	1
	Q20	14	14
	Q21	20	20

Table 20 Overall level of agreement

Level of agreement	Number of statements	%
Strong agreement	16	76.19
Slight agreement	5	23.81
No agreement	0	0

stability, going further than the analysis of the percentage of changes in the assessment.

As a final summary, we were able to meet our main goal and our primary research question, "How to design a suitable expert-judgment method to validate a proposal in the Web Engineering field?" by conducting our experiment and gathering the conclusions presented before. If we analyze our detailed research questions, we can also link them to our main extracted conclusions:

- *RQ1: What should be taken into account when designing an expert judgment method?* As shown during the literature review, elements like the analyzed topic and previous knowledge of the panelists on the topic were key as well as the willingness to participate, since the process might last in time and might require some time and effort from the panel members. It was also shown that, in order to obtain meaningful results, there are two other important issues to take into account: to combine variety of profiles to avoid biases in the obtained results; and to limit the number of panelists, as a high number of them might make the process unmanageable.
- *RQ2: What are the most suitable techniques to process gathered data during an expert judgment method?* We analyzed different techniques, both quantitative and qualitative, from descriptive statistics to advanced techniques like Simple Correspondence Analysis, concluding that the final analysis should be based on a combination of all of them, better than on a single technique. Textual analysis of comments, mean and median calculation, together with reliability, homogeneity and stability tests should be combined in order to process the compiled results and obtain meaningful conclusions.

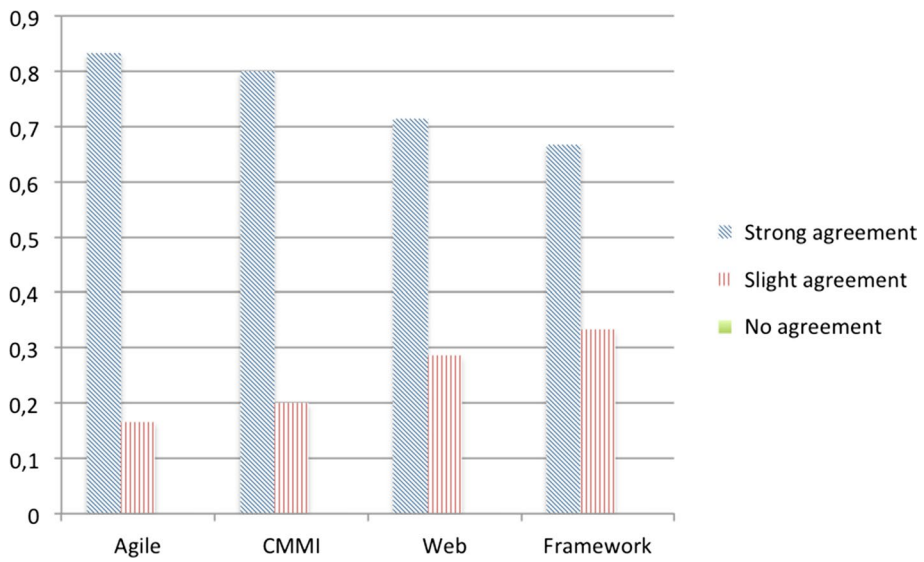


Fig. 30 Overall level of agreement

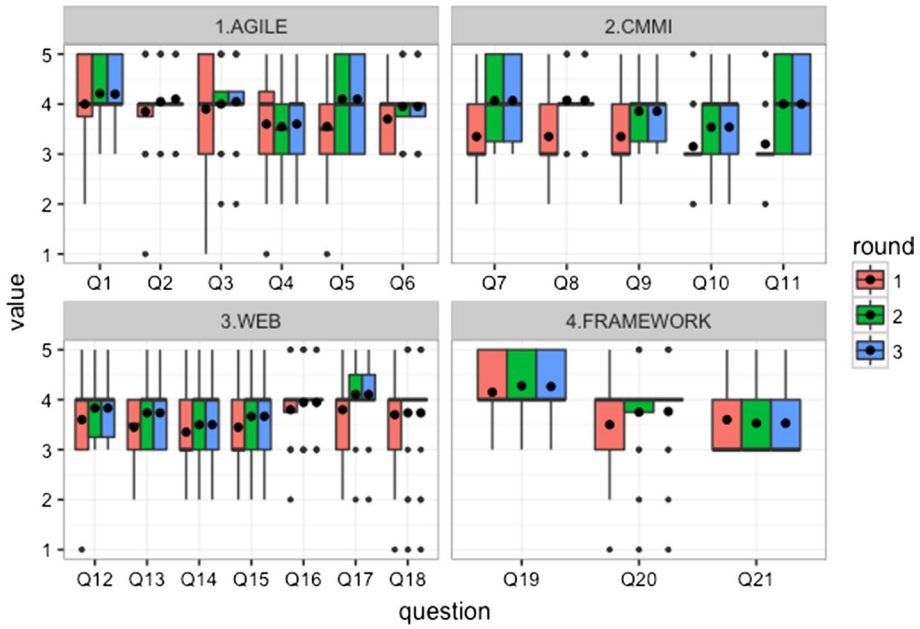


Fig. 31 Evolution of the results through rounds

- *RQ3: How to identify when consensus is reached during an expert judgment method?* As mentioned along previous sections, the literature shows a broad set of techniques in order to evaluate consensus. In our work we made use of the Simple Correspondence Analysis, showing a creative approach of this homogeneity test, that enabled us to measure, in a very visual and graphical way, if consensus was reached among raters on the different questions.

Of course, the generalized application of this design approach using the Delphi method to assess Web Engineering related issues would require further applications, in order to refine the way it should be utilized.

As a final conclusion, we have shown how to apply techniques commonly used in social sciences, like the Delphi method, inferential statistics and homogeneity tests to the field of Software Engineering, which represent by themselves an added value to our work. Nevertheless, and as a limitation of the presented work, it is important to notice that the expert validation method we conducted, even being quite valuable, due to the expertise and the theoretical and practical background of the participants, could never replace real examples and case studies, which constitute the main weakness of our assessment.

As a future line of work and research and linked to our secondary goal, the implementation of *NDT-Agile* framework in a real organization and in several projects needs to be assessed. As mentioned, the expert judgment method might be considered as a preliminary cost-effective step, not replacing empirical data. Additionally, conducting either a self-assessment or a real SCAMPI assessment of *NDT-Agile* implementation will definitively help to improve and validate the model.

Acknowledgements This research has been supported by the Megus project (TIN2013-46928-C3-3-R) and by the SoftPLM Network (TIN2015-71938-REDT) of the Ministerio de Ciencia e Innovación, Spain. We would like to thank Dr. Pedro Antonio García, Dr. Diego Torrecilla de Amo and Dr. Diego Nieto Lugilde, all from the University of Granada, for their useful and helpful comments. Finally, we would like to thank all experts participating in the process for their time, help and useful contribution.

References

1. Alonso JAG, Santacruz MP (2015) Cálculo e interpretación del Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos posibles escalas tipo Likert. Rev Publ 2(2):62–77
2. Ambler SW (2002) Lessons in agility from Internet-based development. IEEE Softw 19:66–73
3. Beck K, Andres C (2004) Extreme programming explained: embrace change, 2nd edn. Addison-Wesley, Boston
4. Beck K et al (2001) Manifesto for Agile software development. http://www.agilemanifesto.org. Accessed 08 2016
5. Benzécri JP (1973) L'Analyse des Données. Volume II, L'Analyse des Correspondances. Dunod, Paris/Bruxelles/Montreal
6. Bougroun Z, Zeaaraoui A, Bouchentouf T (2014) The projection of the specific practices of the third level of CMMI model in agile methods: Scrum, XP and Kanban. In: Third IEEE international colloquium in proceedings of information science and technology (CIST). IEEE, pp 174–179
7. Brooks KW (1979) Delphi technique: expanding applications. North Cent Assoc Q 54(3):377–385
8. Carney O, McIntosh J, Worth A (1996) The use of the nominal group technique in research with community nurses. J Adv Nurs 23(5):1024–1029

9. Chaffin WW, Talley WK (1980) Individual stability in Delphi studies. Technol Forecast Soc Change 16:67–73
10. CMMI Product Team (2010) CMMI for development, version 1.3., Carnegie Mellon University, technical report. http://www.sei.cmu.edu/reports/10tr033.pdf. Accessed 08 2016
11. Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213–220
12. Creswell JW (2003) Research design: qualitative, quantitative, and mixed method approaches, 2nd edn. SAGE, Thousand Oaks
13. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334
14. Cyphert FR, Gant WL (1971) The Delphi technique: a case study. Phi Delta Kappan 52:272–273
15. Dajani JS, Sincoff MZ, Talley WK (1979) Stability and agreement criteria for the termination of Delphi studies. Technol Forecast Soc Change 13(1979):83–90
16. Dalkey NC (1972) The Delphi method: an experimental study of group opinion. In: Dalkey NC, Rourke DL, Lewis R, Snyder D (eds) Studies in the quality of life: Delphi and decision-making. Lexington Books, Lexington, pp 13–54
17. Dalkey NC, Helmer O (1963) An experimental application of the Delphi method to the use of experts. Manag Sci 9:458–467
18. Dell Software (2015) How to analyze simple two-way and multi-way table, correspondence analysis. https://documents.software.dell.com/statistics/textbook/correspondence-analysis#index. Accessed 08 2016
19. Díaz J, Garbajosa J, Calvo-Manzano JA (2009) Mapping CMMI level 2 to Scrum practices: an experience report. SPI, Chennai, pp 93–104
20. Diehl M, Stroebe W (1987) Productivity loss in brainstorming groups: towards the solution of a riddle. J Pers Soc Psychol 53(3):497
21. Escalona MJ, Aragón G (2008) NDT: a model-driven approach for web requirements. IEEE Trans Softw Eng 34(3):370–390
22. Escalona MJ, Mejías M, Torres J (2004) Developing systems with NDT and NDT-Tool. In: 13th International conference on information systems development: methods and tools, theory and practice, Vilna, Lithuania, pp 149–159
23. Falissard B (2012) psy: various procedures used in psychometry. R package version 1.1. https://CRAN.R-project.org/package=psy. Accessed 08 2016
24. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
25. Gamer M, Lemon J, Puspendra Singh IF (2012) irr: various coefficients of interrater reliability and agreement. R package version 0.84. https://CRAN.R-project.org/package=irr. Accessed 11 2016
26. García-Crespo A, Colomo-Palacios R, Soto-Acosta P, Ruano-Mayoral M (2010) A qualitative study of hard decision making in managing global software development teams. Inf Syst Manag 27(3):247–252
27. George D, Mallery P (2003) SPSS for windows step by step: a simple guide and reference. 11.0 update, 4th edn. Allyn & Bacon, Boston
28. Glazer H et al (2008) CMMI or Agile: why not embrace both!, Carnegie Mellon University. http://www.sei.cmu.edu/reports/08tn003.pdf. Accessed 08 2016
29. Gliem RR, Gliem JA (2003) Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In: Midwest research-to-practice conference in adult, continuing, and community education
30. Goldenson DR, Gibson DL, Ferguson RL. Why make the switch? Evidence about the benefits of CMMI. http://www.sei.cmu.edu/library/assets/evidence.pdf. Accessed 08 2016
31. Google. Google Forms. https://www.google.com/intl/es/forms/about. Accessed 08 2016
32. Heiko A (2012) Consensus measurement in Delphi studies: review and implications for future quality assurance. Technol Forecast Soc Change 79(8):1525–1536
33. Hirschfeld HO (1935) A connection between correlation and contingency. Proc Camb Philos Soc 31:520–524
34. Holey EA, Feeley JL, Dixon J, Whittaker VJ (2007) An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. BMC Med Res Methodol 7:52. https://doi.org/10.1186/1471-2288-7-52
35. Hsu CC, Sandford BA (2007) The Delphi technique: making sense of consensus. Pract Assess Res Eval 12(10):1–8
36. Joshi JB, Aref WG, Ghafoor A, Spafford EH (2001) Security models for web-based applications. Commun ACM 44(2):38–44
37. Kitchenham B et al (2009) Systematic literature reviews in software engineering—a systematic literature review. Inf Softw Technol 51(7–15):2009
38. Legendre P (2005) Species associations: the Kendall coefficient of concordance revisited. J Agric Biol Environ Stat 10(2):226–245
39. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 22:1–55
40. Linstone HA, Turoff M (1975) The Delphi method: techniques and applications. Addison-Wesley Publishing Company, Reading, pp 3–12
41. Ludwig B (1997) Predicting the future: have you considered using the Delphi methodology? J Ext 35(5):1–4
42. Lukasiewicz K, Miler J (2012) Improving agility and discipline of software development with the Scrum and CMMI. Softw IET 6(5):416–422
43. Marcal ASC, de Freitas BCC, Furtado Soares FS, Belchior AD (2008) Blending Scrum practices and CMMI project management process areas. ISSE 4:17–29
44. Mendes E, Mosley N (2005) Web cost estimation: an introduction. In: Web engineering: principles and techniques. IGI Global, Hershey, pp 182–202
45. Model Driven Web Engineering Workshop (2012) Satellite workshop of ICWE'2012 conference. http://mdwe2012.pst.ifi.lmu.de/. Accessed 08 2016
46. Murugesan S, Deshpande Y, Hansen S, Ginige A (2001) Web engineering: a new discipline for development of web-based systems. In: Murugesan S, Deshpande Y (eds) Web engineering. Springer, Berlin, pp 3–13
47. Nakatsu RT, Iacovou CL (2009) A comparative study of important risk factors involved in offshore and domestic outsourcing of software development projects: a two-panel Delphi study. Inf Manag 46(1):57–68
48. Nenadic O, Greenacre M (2007) Correspondence analysis in R, with two-and three-dimensional graphics: the ca package. J Stat Softw 20(3)
49. Nunnally JC (1967) Psychometric theory, 1st edn. McGraw-Hill, New York
50. Oh KH (1974) Forecasting through hierarchical Delphi. Unpublished doctoral dissertation, The Ohio State University, Columbus
51. Osborn AF (1957) Applied imagination (rev. ed). Scribner, New York
52. Paulk MC (2001) Extreme programming from a CMM perspective. IEEE Softw 18(6):19–26
53. Pikkarainen M et al (2008) The impact of Agile practices on communication in software development. Empir Softw Eng 13:303–337
54. Pill J (1971) The Delphi method: substance, context, a critique and an annotated bibliography. Socio Econ Plan Sci 5:57–71
55. Pressman RS (2000) What a tangled Web we weave. IEEE Softw 17:18–21
56. R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 08 2016

57. Reifer DJ (2000) Web development: estimating quick-to-market software. IEEE Softw 17:57–64
58. Rieger WG (1986) Directions in Delphi developments: dissertations and their quality. Technol Forecast Soc Change 29(1986):195–204
59. Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Armstrong JS (ed) Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers, Boston, pp 125–144
60. Scheibe M, Skutsch M, Schofer J (1975) Experiments in Delphi methodology. In: The Delphi method—techniques and applications. Addison-Wesley, Reading, pp 262–287
61. Schmidt R, Lyytinen K, Mark Keil PC (2001) Identifying software project risks: an international Delphi study. J Manag Inf Syst 17(4):5–36
62. Schmitt N (1996) Uses and abuses of coefficient alpha. Psychol Assess 8(4):350–353
63. Selleri Silva F, Santana Furtado Soares F, Lima Peres A, Monteiro de Azevedo I, Vasconcelos A, Kenji Kamei F, de Lemos Romero, Meira S (2015) Using CMMI together with agile software development: a systematic review. Inf Softw Technol 58:20–43
64. Siegel S, Castellan NJ (1995) Estadística no paramétrica aplicada a las ciencias de la conducta. Trillas, Mexico City
65. Staples M, Niazi M, Jeffery R, Abrahams A, Byatt P, Murphy R (2007) An exploratory study of why organizations do not adopt CMMI. J Syst Softw 80(6):883–895
66. Sutherland J, Schwaber K (2011) The Scrum guide: the definitive guide to Scrum: the rules of the game. http://www.scrum.org/Scrum-Guides. Accessed 08 2016
67. Torrecilla-Salinas CJ, Guardia T, De Troyer O, Mejías M, Sedeño J (2017) NDT-Agile: an Agile, CMMI-compatible framework for web engineering. In: International conference on software process improvement and capability determination. Springer, pp 3–16
68. Torrecilla Salinas CJ, Sedeño J, Escalona MJ, Mejías M (2014) An Agile approach to CMMI-DEV levels 4 and 5 in Web development projects. In: Information systems development (ISD2016 proceedings). Katowice, Poland
69. Torrecilla Salinas CJ, Sedeño J, Escalona MJ, Mejías M (2016) Agile, web engineering and capability maturity model integration: a systematic literature review. Inf Softw Technol 71(2016):92–107
70. Torrecilla Salinas CJ, Sedeño J, Escalona MJ, Mejías M (2015) Estimating, planning and managing Agile Web development projects under a value-based perspective. Inf Softw Technol 61(2015):124–144
71. Torrecilla Salinas CJ, Sedeño J, Escalona MJ, Mejías M (2014) Mapping Agile practices to CMMI-DEV level 3 in Web development environments. In: Information systems development: transforming organisations and society through information systems (ISD2014 proceedings). Varaždin, Croatia
72. Torrecilla Salinas CJ, Escalona MJ, Mejías M (2012) A Scrum-based approach to CMMI maturity level 2 in Web development environments. In: Proceeding of international conference on information integration and web-based applications and services, Bali, Indonesia, December 3–5 2012. iiWAS, 12. ACM
73. VersionOne (2016) 9th Annual State of Agile survey. http://www.versionone.com/pdf/state-of-agile-development-survey-ninth.pdf. Accessed 08 2016
74. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2016) gplots: various R programming tools for plotting data. R package version 3.0.1. https://CRAN.R-project.org/package=gplots. Accessed 07 2018
75. Welch S, Comer J (1988) Quantitative methods for public administration: techniques and applications. Brooks/Cole Publishing Company, Pacific Grove
76. Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York
77. Wickham H (2007) Reshaping data with the reshape package. J Stat Softw 21(12):1–20. http://www.jstatsoft.org/v21/i12/
78. Yelland PM (2010) An introduction to correspondence analysis. http://www.mathematica-journal.com/2010/09/an-introduction-to-correspondence-analysis/. Accessed 08 2016