Relating Verification and Validation Methods to Software Product Quality Characteristics: Results of an Expert Survey

Isela Mendoza[1], Marcos Kalinowski[2], Uéverton Souza[1], Michael Felderer[3]

[1] Fluminense Federal University, Brazil
{imendoza, ueverton}@ic.uff.br
[2] Pontifical Catholic University of Rio de Janeiro, Brazil
kalinowski@inf.puc-rio.br
[3] University of Innsbruck, Austria
michael.felderer@uibk.ac.at

Abstract. [Context] Employing appropriate verification and validation (V&V) methods is essential to improve software product quality. However, while several V&V methods have been documented, little is known about how these methods relate to specific product quality characteristics. **[Goal]** The goal of this paper is to provide an initial understanding on the suitability of selected V&V methods to address ISO 25010 software product quality characteristics. **[Method]** Therefore, we compiled a list of V&V methods and conducted a survey with V&V experts, asking them to evaluate how well each V&V method allows addressing the ISO 25010 characteristics. **[Results]** We received 19 answers from experts of 7 different countries. Our results express the aggregated expert opinion. It is noteworthy that the experts mostly agreed in their opinions, indicating consistency in the results. **[Conclusions]** To the best of our knowledge this is the first result on the relationship between V&V methods and quality characteristics. We believe that the aggregated opinion of 19 experts can serve as a starting point for further investigations by other researchers and to provide an initial understanding to practitioners.

Keywords: verification and validation methods, software product quality.

1 Introduction

Software quality can be defined as the degree to which a system meets the specified requirements and expectations of a customer or user [9]. For the industry, quality assurance in software development projects has become a high-cost activity. An adequate selection of verification and validation methods (V&V) to ensure that the product is correctly implemented and meets its specifications, is essential for reducing these costs [2][4][7] .

To guarantee the quality of a software product there are standards, such as ISO 25010 [5], specifying the main product quality characteristics. V&V methods, on the other hand, are employed in order to assure software product quality. Unfortunately, the selection of different V&V methods as well as the interdependencies among them are still not well understood. Hence, the software industry faces the problem of choosing specific V&V methods to assure the quality of the software, since an inadequate selection of these methods may generate significant effort throughout the software development process and consequently high costs [2][4][7] .

Taking into account the quality characteristics of the ISO 25010 standard [5], a series of V&V methods compiled mainly from the SWEBOK [1], and other sources [9][10], the main goal of this work is to obtain an initial understanding on which V&V methods are the most appropriate ones to address each of the ISO 25010 characteristics, from the point of view of the software engineering experts. Therefore, we conducted a survey with a sample of 145 experts, all PhDs in software engineering, with relevant publications in the V&V area, and active in at least one of the following software engineering and V&V program committees: ICSE, ICST, ESEM, SEAA-SPPI, and SWQD.

At all, 19 experts from 7 different countries responded to the survey. The results provide an initial characterization of V&V methods against ISO 25010 quality characteristics. While our results still represent an initial understanding, the overall agreement among the experts reinforces our confidence that they represent a meaningful starting point for other researchers and that they can be used as an initial reference on the topic by the software industry. In response to one of the survey questions, experts also recommend new methods to be evaluated in future survey trials.

The document is organized as follows. In Section 2, the background on the chosen V&V methods is presented. In Section 3, the ISO 25010 quality characteristics are described. In Section 4, the survey plan is outlined. In Section 5, we describe the survey operation, i.e., how the survey was conducted. In Section 6, the survey results are presented and analyzed. Finally, Section 7 contains the concluding remarks.

2 Software Verification and Validation Methods

Several V&V methods have been proposed throughout the years. Hereafter we detail a selection of such methods, representing an aggregated compilation of the methods presented in the SWEBOK [1] and two books focused respectively on software product quality and peer reviews [9][10]. Table 1 shows the classification of the selected V&V methods and a very short description of each of them based on the descriptions provided in [1][9][10].

Table 1. Description and Classification of the Selected V&V Methods

Classification	*Methods*	*Short Description*
Based on Intuition & Experience	Ad hoc Testing	Tests are derived relying on the software engineer's skill, intuition, and experience with similar programs.
	Exploratory Testing	Is defined as simultaneous learning, test design, and test execution, that is, the tests are not defined in advance in an established test plan, are dynamically designed, executed, and modified.
Input Domain-Based	Equivalence Partitioning	Involves partitioning the input domain into a collection of subsets (or equivalent classes) based on a pacified criterion or relation.
	Pair wise Testing	Test cases are derived by combining interesting values for every pair of a set of input variables instead of considering all possible combinations.
	Boundary-Value Analysis	Test cases are chosen on or near the boundaries of the input domain of variables, with the underlying rationale that many faults tend to concentrate near the extreme values of inputs.
	Random Testing	Tests are generated purely at random. This form of testing falls under the heading of input domain testing since the input domain must be known to be able to pick random points within it.
	Cause-Effect Graphing	Represent the logical relationships between conditions (roughly, inputs) and actions (roughly, outputs). Test cases are systematically derived by considering combinations of conditions and their corresponding resultant actions.
Code-Based	Control Flow-Based Criteria	Are aimed to covering all the statements, blocks of statements, or specified combinations of statements in a program.
	Data Flow-Based Criteria	In data flow-based testing, the control flow graph is annotated with information about how the program variables are defined, used, and killed (undefined).
Fault-Based	Error Guessing	In error guessing, test cases are specifically designed by software engineers who try to anticipate the most plausible faults in a given program.
	Mutation Testing	A mutant is a slightly modified version of the program under test, differing from it by a small syntactic change.
Usage-Based	Operational Profile	In testing for reliability evaluation (also called operational testing), the test environment reproduces the operational environment of the software, or the operational profile, as closely as possible. The goal is to infer from the observed test results the future reliability of the software when in actual use.
	Usability Inspection Methods	Usability principles can provide guidelines for discovering problems in the design of the user interface. Are also called usability inspection methods, including: Heuristic evaluation or User Observation Heuristics, Heuristic estimation, Cognitive walkthrough, Pluralistic walkthrough, Feature inspection, Consistency inspection, Standards inspection and Formal usability inspection.
Model-Based Testing	Finite-State Machines	By modeling a program as a finite state machine, tests can be selected in order to cover the states and transitions.
	Workflow Models	Workflow models specify a sequence of activities performed by humans and/or software applications, usually represented through graphical notations.
Reviews	Walkthrough	The purpose of a systematic walk-through is to evaluate a software product. A walkthrough may be conducted for educating an audience regarding a software product.
	Peer Review or Desk Checking	The authors do not explain the artifact. They give it to one or more colleagues who read it and give feedback. The aim is to find defects and get comments on the style.

	Technical Review	Further formalizes the review process. They are often also management reviews or project status reviews with the aim to make decisions about the project progress. In general, a group discusses the artefacts and decides about the content.
	Inspection	The purpose of an inspection is to detect and identify software product anomalies. Some important differentiators of inspections as compared to other types of technical reviews are the roles (author, inspection leader, inspector, and scribe) and the inspection process which consists of the steps planning, kick off, individual checking, logging meeting and edit and follow-up. Some examples of inspections are: Checklist-based reading, Usage-based reading, Defect-based reading, and Perspective-based reading.

3 Software Quality Characteristics

The quality of the processes is important to deliver high quality software products. However, many factors influence the quality of the product itself, so it is necessary to evaluate and monitor the quality directly in the product, and improve the processes that create them [9].

In this paper we focus on software product quality. In this context, the ISO 25010 standard defines the quality model that is considered the cornerstone of a product quality evaluation system. The product quality model defined by ISO 25010 is composed of eight quality characteristics, which determine the properties of a software product for its evaluation [5]. Figure 1 shows the eight quality characteristics of the ISO 25010 standard. A short description of these quality characteristics and a listing of the corresponding sub-characteristics, based on the definitions contained in [5] is follows in the Table 2.

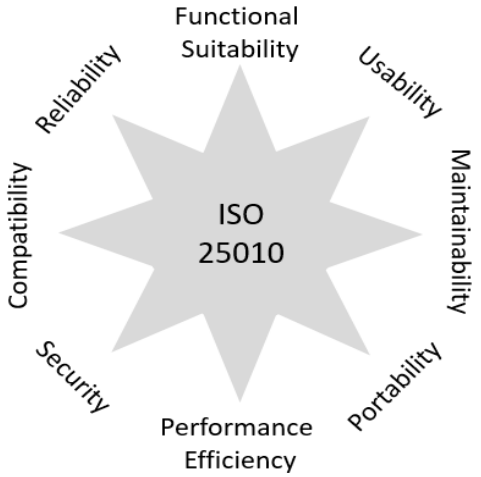

Fig.1 Quality characteristics of ISO 25010.

Table 2. Description of ISO 25010 Quality Characteristics

Characteristic	*Short Description*	*Sub-characteristics*
Function Suitability	Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.	Functional Completeness, Functional Correctness and Functional Appropriateness.
Performance Efficiency	Represents the performance relative to the amount of resources used under stated conditions.	Time Behavior, Resource Utilization and Capacity.
Compatibility	Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.	Co-existence and Interoperability.
Usability	Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.	Appropriateness, Recognizability, Learnability, Operability, User Error Protection, User Interface Aesthetics and Accessibility.
Reliability	Degree to which a system, product or component performs specified functions under specified conditions for a specified period.	Maturity, Availability, Fault Tolerance and Recoverability.
Security	Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.	Confidentiality, Integrity, Non-repudiation, Accountability and Authenticity.
Maintainability	Degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements.	Modularity, Reusability, Analyzability, Modifiability and Testability.
Portability	Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.	Adaptability, Installability and Replaceability.

4 Survey Plan

4.1 Main Goal and Scope

The main goal of our survey is to gather initial evidence, through expert opinion, about the suitability of V&V methods to address ISO 25010 software product quality characteristics. Using the GQM (Goal Question Metric) definition template [11] this goal can be stated as: **Analyze** V&V methods **for the purpose of** characterization **with respect to** their suitability for addressing ISO 25010 software quality characteristics **from the point of view of** experts in the area of V&V **in the context** of the software engineering research community.

4.2 Population

Following the advice of deciding upon the target population based on whether they are the most appropriate to provide accurate answers instead of focusing on hopes to get high response rates [8], our population of V&V experts will be sampled by selecting PhDs in software engineering that are active in at least one of the following software engineering and V&V program committees: ICSE, ICST, ESEM, SEAA-SPPI, and SWQD. Additionally, each survey participant should have at least one publication within the last 5 years directly related to V&V methods. We believe that this strategy allows effectively reaching a sample of V&V experts from the software engineering research community.

4.3 Survey Questions

The survey was designed with only two very direct questions. The intent of keeping the design simple was to allow the experts to answer within a reasonable timeframe.

Q1: To what extent do you agree that the following V&V methods can be applied to address the listed quality attributes?

This question was structured as a table crossing the selected V&V methods (cf. Section 2) against the ISO 25010 quality characteristics (cf. Section 3). The researchers should provide their answers filling each cell with a number corresponding to a *Likert* scale (1- Disagree, 2-Partially Disagree, 3- Partially Agree, 4- Agree, and N- Not Sure).

Q2: Complete the list by rating any other V&V methods that you believe could be applied to address one or more of the listed quality attributes.

This question was optional and provided to allow the expert to suggest and evaluate other V&V methods, not included in our initial list, that he considers relevant.

4.4 Metrics

Likert scales (1- Disagree, 2- Partially Disagree, 3- Partially Agree, 4- Agree, and N- Not Sure) were used for assessing the V&V methods against the quality characteristics in both questions. The aggregated metric on the agreement for each method/quality-characteristic set was obtained using the median value, which can be safely applied to *Likert* scales [11].

4.5 Execution strategy

The execution strategy consisted of identifying the population sample according to the strategy and distributing the survey instrument via email. Due to the format of the questions, the survey was provided by e-mail as an MS Word attachment. Participants should answer the survey within 15 days.

4.6 Statistical Techniques

The aggregation of the responses for the set of answers was conducted using the median value. Additionally, the Median Absolute Deviation (MAD), representing the degree of concordance between the experts, should be analyzed to further understand the representativeness of the median for the sample. Statistical visualization features to provide an overview of the results include tables and a bubble plot crossing information of V&V methods and quality characteristics.

4.7 Instrumentation

The questionnaire instrument had a title, a short description of the research goal, a note of consent stating that individual data will be handled anonymously, followed by the two questions. Additionally, as supporting documentation, short descriptions of the V&V methods and the ISO 25010 quality characteristics were also provided as an appendix.

4.8 Validity Assessment

Throughout the process of planning the survey, we identified some threats. Table 2 lists these potential threats and how we treated them in our survey. One of the main mitigation strategies was validating the instrument by asking other individuals to answer the survey, as part of a pilot study, before handing it over to the experts. This also allowed us to understand that, while we tried to keep it is simple as possible, answering the questionnaire still requires at least 20 minutes.

Table 2. Threats and Treatment

Threats	***Treatment***
Bad instrumentation	Revision and evaluation of the questionnaire about the format and formulation of the questions. Running a pilot study.
Inadequate explanation of the constructions.	Revision and evaluation of the questionnaire about the format and formulation of the questions. Running a pilot study.
Doubts of the experts on the purpose or on specific definitions	Including the research goal explanation and adding support information on the V&V methods and ISO 25010 quality characteristics.
Measurement and results reliability.	Using medians to aggregate individual Likert scale entries. Using the median absolute deviation to check on the agreement among the experts.
Statistical conclusion validity	This threat strongly depends on the sample size. A mitigation that could be used is running future survey replications and aggregating the results.

5 Operation

Our population sampling strategy, described in the sub-section 4.2, allowed us to identify 145 candidate subjects (PhDs in software engineering, active in one of the selected program committees, and with relevant publications on V&V methods). The survey was sent to them by e-mail as an MS Word attachment, which could be easily answered by using any MS Word compatible editor. Experts had 15 days to answer the survey. After this deadline the data collection was considered concluded.

At all, 19 experts (response rate of ~13%) from 7 different countries answered the survey. Taking into account the main factors that directly affect the response rate [6]: length of the form (number of pages), sending of the form through e-mail, duration to fill out the form (indeed, some experts mentioned that answering took them much longer than expected from our pilot study) and comparing with similar studies (e.g., [3]), where the response rate is commonly around 10%, we can consider our response rate satisfactory and according to our expectation.

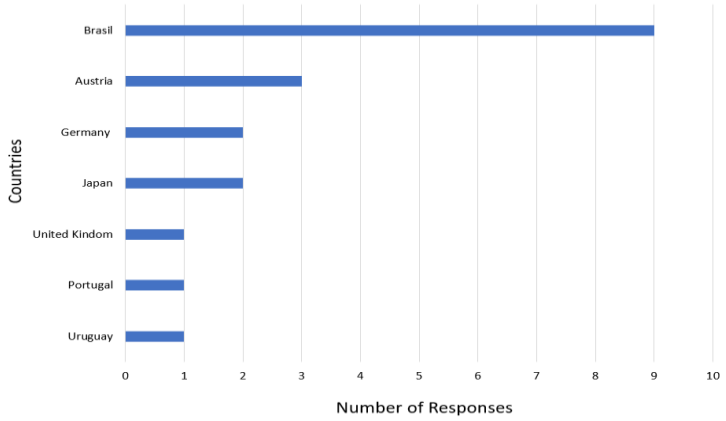

Fig.2 Number of responses (horizontal axis) per country.

Figure 2 represents the number of responses per country. It can be observed that most of the answers came from Brazil and Austria. This was probably related to the direct relationship of the authors with researchers from these countries.

6 Survey Results

The overall results relating the V&V and ISO 25010 quality characteristics are shown in Table 3. The rows contain each of the selected V&V methods and the columns show the ISO 25010 software quality characteristics: FS – Functional Suitability, PE – Performance Efficiency, C – Compatibility, U – Usability, R – Reliability, S – Security, M – Maintainability and P – Portability. The numbers in the cells correspond to the median values of the experts' answers on how suitable the method is for a given characteristic. For the purpose of calculating the median value "N - Not

Sure" responses were not considered. The Median Absolute Deviations (MAD) are shown within parentheses.

For the analysis of the survey data, we consider that the method is reported to address a quality characteristic if the median value of the answers of the respondents is greater than or equal to 3. The cells corresponding to these values are highlighted in grey in Table 3. It is possible to observe that, according to the experts, there is at least one of the selected V&V methods addressing each quality characteristic. Most of the methods address functional suitability.

Table 3. Suitability of V&V methods to address ISO 25010 quality characteristics.

Methods	***FS***	***PE***	***C***	***U***	***R***	***S***	***M***	***P***
1. Ad Hoc Testing	**3 (0)**	1 (0)	2 (0.5)	**3 (1)**	1 (0)	1.5 (0.5)	1 (0)	1 (0)
2. Exploratory Testing	**3 (1)**	2 (1)	2 (0.5)	**3 (1)**	2 (0)	2 (1)	2 (0)	2 (1)
3. Equivalence Partitioning	**4 (0)**	2 (1)	1.5 (0.5)	1 (0)	2 (1)	1 (0)	1 (0)	1 (0)
4. Pair wise Testing	**3.5 (0.5)**	1 (0)	2 (1)	1 (0)	2 (1)	2 (1)	1 (0)	1 (0)
5. Boundary-Value Analysis	**4 (0)**	2 (1)	2 (1)	1 (0)	2 (1)	2 (1)	1 (0)	1 (0)
6. Random Testing	**3 (1)**	2 (1)	2 (1)	1 (0)	2 (1)	2 (1)	1 (0)	1 (0)
7. Cause-Effect Graphing	**4 (0)**	2 (1)	2 (0.5)	1.5 (0.5)	2 (1)	2 (1)	1.5 (0.5)	1 (0)
8. Control Flow-Based Criteria	**3 (1)**	1.5 (0.5)	1 (0)	1 (0)	**3 (1)**	2 (1)	2 (1)	1.5 (0.5)
9. Data Flow-Based Criteria	**3 (1)**	1 (0)	1 (0)	1 (0)	**3 (1)**	2 (1)	2 (1)	1 (0)
10. Error Guessing	**3 (1)**	2 (1)	2.5 (0.5)	2 (1)	**3 (1)**	2 (1)	1 (0)	2 (1)
11. Mutation Testing	**3 (1)**	1 (0)	2 (1)	1 (0)	**3 (1)**	2 (1)	1 (0)	1 (0)
12. Operational Profile	**3 (1)**	**3 (1)**	**3 (1)**	**3 (1)**	**3 (1)**	2 (1)	1 (0)	2 (1)
13. Usability Inspection Methods	2 (1)	1.5 (0.5)	2 (1)	**4 (0)**	2 (1)	2 (1)	1 (0)	1 (0)
14. Finite-State Machines	**4 (0)**	2 (1)	2 (1)	2 (1)	**3 (1)**	**3 (1)**	2 (1)	1 (0)
15. Workflow Models	**4 (0)**	2 (0)	2 (1)	**3 (1)**	2 (1)	**3 (1)**	2 (1)	1 (0)
16. Walkthrough	**3 (1)**	2 (1)	2 (0)	2 (1)	2 (1)	2 (1)	**3 (0.5)**	2 (1)
17. Peer Review or desk checking	**3 (1)**	2 (1)	2 (1)	2 (1)	2 (1)	**3 (1)**	**3 (1)**	2 (1)
18. Technical Review	**3 (0)**	2 (1)	2 (1)	2 (1)	2.5 (0.5)	**3 (1)**	**3 (1)**	**3 (1)**
19. Inspection	**4 (0)**	2 (1)	2 (1)	**3 (1)**	2.5 (1.5)	**3 (1)**	**3 (1)**	**3 (1)**

The MAD represents the agreement between the experts. Figure 3 shows the overall distribution of the MAD values: 32%, 9%, 58%, and 1%, for the MAD values 0 (blue), 0.5 (orange), 1 (grey), and 1.5 (yellow), respectively. It can be seen that these values mainly oscillate in a range between 0 and 1 (except for one element that equals 1.5, concerning using inspections to address reliability). These overall low deviations indicate small differences between the opinions of the experts.

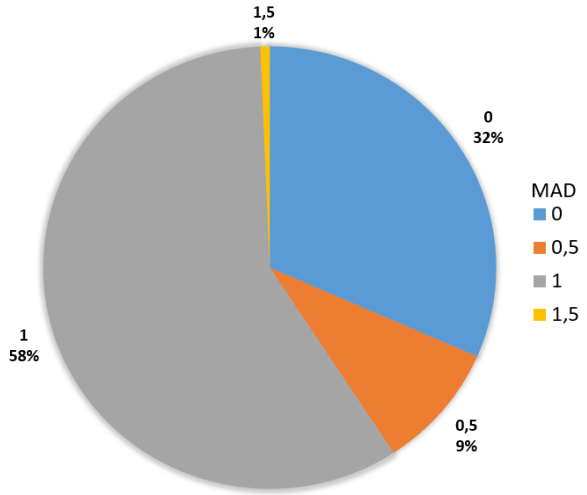

Fig 3. Median absolute deviations of the answers provided by the experts.

Figure 4 provides a summary of the relation between the V&V methods and the quality characteristics in a bubble plot. In this Figure, the size of the bubble: small, medium and large, represents the median value: 3, 3.5 and 4, respectively. The colors refer to the MAD value: Blue, Orange and Grey represent the values of 0, 0.5 and 1, respectively. Thus, the large blue plots represent combinations where the experts agree (median 4) with strong consensus (MAD 0) that the V&V method is suitable for addressing the quality characteristic. It is noteworthy that the grey dots, still represent a positive evaluation for the combination (median 3 and MAD 1).

The participants mentioned 20 other (more specific) V&V methods. Among these methods, the ones that were cited by more than one expert were: Model Checking, Penetration Testing, Stress Testing, and Fuzz Testing. It is noteworthy that all methods suggested by more than one expert are automated or semi-automated ones. This indicates that in a future survey trial such methods should probably be included.

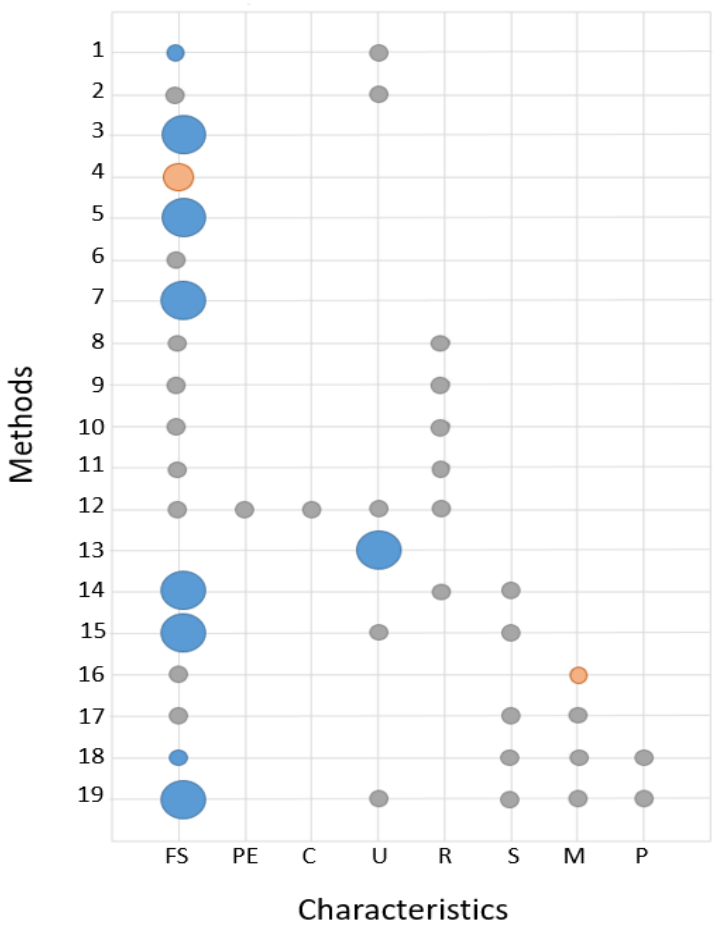

Fig 4. Map of V&V methods that were considered by experts to appropriately help addressing each of the ISO 25010 quality characteristics. Size of the bubble: small, medium and large, represent the median value: 3, 3.5 and 4, respectively. Colors refer to the agreement (MAD value): Blue, Orange and Grey represent the values of 0 (strong agreement), 0.5 and 1 (still a reasonable agreement).

7 Concluding Remarks

In this short paper we proposed to establish a relation between a set of V&V methods and the ISO 25010 quality characteristics, based on expert opinions. It is noteworthy that, to the best of our knowledge, such relation, while being extremely relevant for research and practice, has not yet been established.

Therefore, we compiled an initial list of V&V methods and carefully selected experts to answer our survey. At all, we received answers from 19 experts, all holding

PhDs in software engineering, being part of relevant program committees and having recent publications concerning V&V methods.

The resulting relations (suitability of the V&V methods to address the ISO 25010 quality attributes) are summarized in Figure 5. All the bubbles represented in this Figure concern an agreement on the relation between the method and the quality characteristic. Nevertheless, considering our sample size, the relations depicted by the small (median 3) grey (MAD 1) bubbles, while still representing an aggregated expert agreement on the relationship, should be taken with a grain of salt, given that in these cases the experts disagreed slightly more. Considering the overall sample, experts mostly provided consistent answers with small deviations from the median.

While our sample size is small and we do not claim for statistical conclusion validity and are aware of the importance of replications to reinforce our results, we still believe that the aggregated opinion of 19 experts can serve as a starting point for other researchers and practitioners, who currently completely lack information on the suitability of V&V methods to address ISO 25010 quality attributes.

Acknowledgments. The authors would like to thank all the survey respondents.

References

1. Bourque P., Fairley R.E: SWEBOK Guide V3.0, Guide to the Software Engineering Body of Knowledge, IEEE Computer Society (2004).
2. Endres A. and Rombach D.: A Handbook of Software and Systems Engineering, Addison Wesley (2003).
3. Felderer, M.; Auer, F. Software quality assurance during implementation: Results of a survey in software houses from germany, austria and switzerland. 9th International Conference, SWQD 2017, Vienna, Austria, (2017).
4. Feldt R., Marculescu B., Schulte J., Torkar R., Preissing P., Hult E.: Optimizing Verification and Validation Activities for Software in the Space Industry. Data Systems in Aerospace (DASIA), Budapest, (2010).
5. ISO25000 Software Product Quality, ISO/IEC 25010, http://iso25000.com/index.php/en/iso-25000-standards/iso-25010, Official site (2011).
6. Linåker, J., Sulaman, S. M., Maiani de Mello, R., Höst, M.: Guidelines for Conducting Surveys in Software Engineering. Technical Report Lund University, Sweden (2015).
7. Meyers G.J., Badgett T., Thomas T., Csandler C.: The Art of Software Testing. Wiley, 3rd edition (2011).
8. Torchiano, M., Fernández, D.M., Travassos, G.H. de Mello, R.M.: Lessons learnt in conducting survey research. In: Proceedings of the 5th International Workshop on Conducting Empirical Studies in Industry, pp. 33-39, (2017)
9. Wagner S.: Software Product Quality Control. Springer (2013).
10. Wiegers K.E.: Peer Reviews in Software: A Practical Guide Addison-Wesley (2002).
11. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering. Springer (2012).