Understanding and Supporting Software Architectural Decisions

Tofan, Dan

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 17-11-2019
Chapter 7

Improve Group Architectural Decisions


As found in Chapter 3, many software architectural decisions are group decisions rather than decisions made by individuals. Consensus in a group of decision makers increases the acceptance of a decision among decision makers and their confidence in that decision. Furthermore, going through the process of reaching consensus means that decision makers understand better the decision (including the decision topic, decision options, rationales, and potential outcomes). However, as found in Chapter 4, little guidance exists on group architectural decision making, and on how to increase consensus.

In this chapter, we propose and evaluate how a process (named GADGET) helps architects increase consensus when making group architectural decisions. Specifically, we investigate how well GADGET increases consensus in group architectural decision making, by understanding its practical applicability, and by comparing GADGET against group architectural decision making without using any prescribed approach.

We conducted two empirical studies. First, we conducted an exploratory case study to understand the practical applicability of GADGET in industry. We investigated whether there is a need to increase consensus, the effort and benefits of GADGET, and potential improvements for GADGET. Second, we conducted an experiment with 113 students from three universities to compare GADGET against group architectural decision making without using any prescribed approach.
Study results indicate that GADGET helps decision makers increase their consensus, captures knowledge on architectural decisions, clarifies the different points of view of different decision makers on the decision, and increases the focus of the group discussions about a decision. In addition, we used the feedback from industrial practitioners to refine GADGET. From the experiment, we obtained causal evidence that GADGET increases consensus among participants better than group architectural decision making without using any prescribed approach.
7.1 Introduction

Current approaches for architectural decision-making are aimed at individual architects. This includes the REGAIN approach presented in Chapter 6. However, the survey in Chapter 3 indicates that software architects make most of their decisions in groups, rather than individually. Other researchers also report that most architectural decisions are made in groups (Miesbauer and Weinreich, 2013).

Unfortunately, little is known about group architectural decisions, and how to improve group architectural decision making. The mapping study on architectural decisions in Chapter 4 indicates a small number of papers on how to make group architectural decisions.

Group architectural decisions bring additional challenges compared to individual architectural decisions, such as the need for communication among decision makers and increased consensus between decision makers and other stakeholders (Svahnberg, 2004).

Increasing consensus among decision makers is a critical factor of group decision making. On the one hand, low consensus in early architectural decisions may lead to misunderstandings within the group of decision makers (Svahnberg, 2004). Such misunderstandings may cause problems. For example, if a stakeholder feels that her point of view about a decision was not taken seriously, that stakeholder might not accept the final software system. On the other hand, benefits of consensus include higher acceptance and better understanding of the architectural decision by all involved stakeholders. Furthermore, consensus increases confidence in the correctness of the architectural decision (Svahnberg, 2004). Therefore, consensus needs to be addressed explicitly as part of group architectural decision making. However, as mentioned before, no approach from software architecture literature targets explicitly the increase of consensus in group architectural decision making.

Regarding the scope of this chapter, we focus on consensus (i.e. ‘we have some general agreement and we understand each other’s perspectives’) instead of unanimity (i.e. ‘all of us have the same perspectives’). Consensus has two main components: general agreement and mutual understanding among stakeholders involved in making a decision (Tastle and Wierman, 2007).
7. Improve Group Architectural Decisions

Therefore, in this chapter, we investigate how to increase general agreement and mutual understanding amongst decision makers. Finally, we focus on inexperienced architects, rather than senior architects, because senior architects usually have enough experience to handle group decision making.

In this chapter, we evaluate GADGET (Group Architectural Decisions with repertory Grid Technique), which is a group decision making process for helping architectural decision makers (e.g. architects and other stakeholders who have a decision-making role) increase consensus about their decisions. The key target group of GADGET is inexperienced architects, as aforementioned, because they need support to reach consensus in group decision making. In addition, GADGET aims at groups that are recently formed and which do not have common procedures and processes in place, and therefore may benefit from a standardized way of interaction. The process offers guidance for increasing consensus incrementally, making explicit the knowledge of the decision makers, and helping them structure their group interactions.

This chapter contributes with empirical evidence of how GADGET increases consensus in group architectural decision making. The validation has two parts:

- a case study with seven students and thirteen practitioners
- an experiment with 113 students to answer research questions that emerged from the case study

This chapter is organized as follows. Section 7.2 presents GADGET. Next, GADGET is evaluated in Section 7.3 in a case study, and in Section 7.4 in an experiment. We discuss validity threats for the evaluation in Section 7.5, and related work in Section 7.6. Finally, Section 7.7 presents conclusions.

7.2 The GADGET Process

GADGET extends our previous work on making and capturing architectural decisions with the Repertory Grid technique (that we also used previously for architectural knowledge acquisition in Chapters 5 and 6) with the idea of group evaluations and feedback from the Delphi technique (Linstone and Turoff, 2002).
The **Delphi technique** is a ‘method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem’ (Linstone and Turoff, 2002). In Delphi, participants answer questions on a complex problem in several iterations, receive a summary of answers from all other participants, and are given the opportunity to revise their answers for the next iteration. After several iterations, the answers converge and determine the solution to the complex problem.

In addition to Delphi, we also considered other techniques to be included in GADGET, namely **brainstorming** (Osborn, 1963) and **nominal group** (Delbecq and Van de Ven, 1971). However, we preferred Delphi for the following reasons. **Brainstorming** is strong at generating new, creative ideas, while performing evaluations. Since our goal was to increase consensus, these characteristics were not high priority for GADGET. The **nominal group** technique has similar steps as Delphi, but the evaluation step is anonymous. We preferred that GADGET has an open evaluation step, so that participants can communicate and understand faster each other’s perspectives.

Here are the five steps used by GADGET, using the topic of an architectural decision (e.g. choice of database, architectural patterns, JavaScript framework, or platform technologies) as input. Since the steps are not domain-specific, GADGET can be used also for making group decisions in other domains. In this chapter, we focus on evaluating GADGET for making group decisions in the software architecture domain.

![Diagram of GADGET process steps and outcomes](image-url)

**Figure 7.1.** GADGET process steps and outcomes.

Each step consists of the following.

- **Decision topic**
- **Legend:**
- **Process Step**
- **Outcome**
- **1. Indicate alternatives and concerns**
- **2. Discuss alternatives and concerns**
- **3. Prioritize and rate concerns**
- **4. Discuss differences**
- **5. Iterate**
- **Common set of alternatives and concerns**
- **Individual priorities and ratings**
- **Consensus and persisting divergences**
1. **Indicate alternatives and concerns:** Decision-makers individually indicate their alternatives and concerns for the decision topic. Also, decision-makers can indicate what alternatives or concerns to remove from previous iterations (see Step 5). The rationale for this step is to ensure that any potentially relevant alternative and concern is considered in the decision making process. The output of this step is a set of alternatives and concerns from each decision-maker. For example, for making a decision about the JavaScript framework, one of the decision-makers indicates three alternatives (e.g. Angular, Ember, and Backbone), and four concerns (e.g. testability, performance, learning curve, and existing skillsets).

2. **Discuss alternatives and concerns:** Decision-makers have a group discussion on the alternatives and concerns, with the purpose of consolidating them in a common set of agreed alternatives and concerns. The rationale for this step is to clarify and potentially add or remove alternatives and concerns that are included in the decision making process. For example, more alternatives can be added and some concerns can be clarified (e.g. what is minimum acceptable performance of a JavaScript framework).

3. **Prioritize concerns and rate concerns against alternatives:** Decision-makers individually prioritize concerns using the hundred-dollar approach (i.e. assign a priority to each concern from 0 to 100, so that the sum of priorities is 100), which was evaluated in Chapter 6. In addition, decision-makers individually rate each alternative against every concern, using a five-level Likert scale, with values ranging from ‘1-strongly disagree’ to ‘5-strongly agree’. Decision-makers may use supplementary values such as ‘not applicable’ and ‘don’t know’. The rationale for this step is to ensure that alternatives and the importance of concerns are considered when making the decision (some stakeholders may consider alternatives and concerns more or less important than others). The output of this step is the set of ratings and priorities from each decision-maker.

4. **Discuss differences:** Based on the ratings and priorities of concerns from Step 3, metrics are calculated for priorities and ratings. For ease of interpretation and usability of GADGET, only four metrics are used for the ratings and priorities indicated by participants in Step 3: a) average of ratings of alternatives based on concerns, b) average
priorities of concerns, c) the range of ratings of alternatives based on concerns, and d) the range of priorities (i.e. difference between highest and smallest values of priorities and ratings). These metrics help decision-makers understand how their own perspectives compare to the perspectives of the other decision-makers. This generates a ‘soft’ pressure towards convergence. If differences in ranges are small enough, then there is an acceptable degree of consensus among decision makers. Otherwise, the decision-makers with highest differences present their rationales to stimulate focused discussions about the differences in perceptions. During these discussions, participants are either willing to modify their priorities and ratings, or they ‘agree to disagree’. The expected output of this step is increased consensus, and/or explicit list of persisting divergences, which, if too big, might require an additional iteration.

5. Iterate from Step 1: This step is needed if decision-makers choose to update the ratings and priorities provided in Steps 1 and 3. The discussions in Step 4 may modify the perspectives of the decision-makers, which could lead to new alternatives and concerns, or different priorities and ratings of concerns. The rationale of this step is that it enables decision-makers to capture their updated perspectives.

### 7.3 GADGET Case Study

We conducted an exploratory case study to explore the practical applicability of GADGET for the purpose of evaluating GADGET with respect to its impact on consensus among decision makers from the viewpoint of a group of decision makers, in the context of architectural decisions. Case studies are very well suited for exploratory research questions (Runeson et al., 2012), since case studies offer flexibility to study a phenomenon (e.g. group decision making) in its real-world context. Case studies rely on observations to form tentative hypotheses and confirmatory research questions, which can be further investigated in subsequent studies. Next, we report the case study using the guidelines from (Runeson and Höst, 2009).

#### 7.3.1 Case Study Design

We defined the following three case study research questions:
RQ1. Is there a practical need for increasing consensus in group architectural decision making?

As discussed in Section 7.1, there is very little work on consensus in group architectural decision making. Therefore, before investing efforts into developing approaches for increasing consensus, we investigated whether such approaches are needed. If there is a practical need to increase consensus in group architectural decision making, then an approach such as GADGET may satisfy this need.

RQ2. What are the effort and benefits offered by GADGET?

The rationale for RQ2 is that practitioners are usually interested in the actual benefits of a new approach (or GADGET in our case) and effort (i.e. time) involved in using it. If an approach has low benefits and requires high effort, then practitioners are unlikely to use such approach. Researchers need to pay attention to effort and benefits of a new approach, to avoid proposing approaches that practitioners are unlikely to use.

RQ3. What are potential improvements to GADGET?

The rationale for RQ3 is that we were open to improve GADGET to ensure it satisfies the needs of its potential users. To improve GADGET, we needed feedback from participants in the case study.

To recruit participants, we invited practitioners from the local community of architects in Groningen. In addition, to obtain more data, we invited graduate students with practical experience, who took the software architecture course offered at the University of Groningen.

The case study used groups of three to four participants. Each case study session for each group consisted of three steps:

1. Participants received an overview of the case study session in which they participated, the GADGET process, and an example to illustrate the GADGET process.
2. Participants used GADGET on an architectural decision topic they had been involved with in their recent activity. Participants entered alternatives, concerns, and ratings into a shared online spreadsheet that we had prepared in advance.
3. Participants provided feedback on GADGET in a group discussion. To focus the group discussions, we prepared the set of discussion items in Table 7.1. We used the discussion items for RQ1 only during the sessions with practitioners, and skipped these questions in the sessions with students, since we were interested in identifying the real-world need for GADGET, as indicated by practitioners.

<table>
<thead>
<tr>
<th>ID</th>
<th>Discussion Item</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Do conflicting perspectives occur in group architectural decision making?</td>
<td>RQ1</td>
</tr>
<tr>
<td>2</td>
<td>What is the impact of conflicting perspectives in group architectural decision making?</td>
<td>RQ1</td>
</tr>
<tr>
<td>3</td>
<td>What approaches have you used so far in consensus building? (If any)</td>
<td>RQ1</td>
</tr>
<tr>
<td>4</td>
<td>What did you like/dislike about the proposed process?</td>
<td>RQ2, RQ3</td>
</tr>
<tr>
<td>5</td>
<td>Would you use this process in your practice?</td>
<td>RQ2, RQ3</td>
</tr>
<tr>
<td>6</td>
<td>Did you change your opinion about alternatives? Why (not)?</td>
<td>RQ2</td>
</tr>
<tr>
<td>7</td>
<td>How did the process help?</td>
<td>RQ2</td>
</tr>
<tr>
<td>8</td>
<td>How can the process be improved?</td>
<td>RQ3</td>
</tr>
<tr>
<td>9</td>
<td>In which situations would you apply the process?</td>
<td>RQ2, RQ3</td>
</tr>
</tbody>
</table>

We made audio recordings of the sessions, with the prior permission of the participants. For analyzing the feedback from participants, two researchers independently performed content analysis on the transcriptions of the recordings and observer’s notes, to identify codes corresponding to sentences, phrases or paragraphs, as recommended by (Krippendorff, 2004). Then, in case of differences in interpretation, researchers discussed and resolved the differences. We grouped the codes from the content analysis to answer the three research questions: on need for consensus in group architectural decision
making (RQ1), effort/benefits of GADGET (RQ2), and possible improvement for GADGET (RQ3).

7.3.2 Results

7.3.2.1. Case Study Participants and Execution

Table 7.2 summarizes the groups of students and practitioners that participated in the case study, and the decision topics that were addressed during the sessions. Years of experience refer to practical experience in software engineering. Groups S1, S2 and P2 opted to use topics that we prepared in advance, and all other groups used decision topics from their activities. Regarding tool support for the sessions, groups P3 and P4 chose to use an early version of the dedicated tool support for GADGET (detailed in Chapter 8). The other four groups chose to use the shared online spreadsheet.

<table>
<thead>
<tr>
<th>Group id</th>
<th>Group size</th>
<th>Group type</th>
<th>Average years of experience</th>
<th>Decision topic</th>
<th>Number of GADGET iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4</td>
<td>Students</td>
<td>4.62</td>
<td>Enterprise Resource Planning system</td>
<td>1</td>
</tr>
<tr>
<td>S2</td>
<td>4</td>
<td></td>
<td>4.50</td>
<td>JavaScript framework</td>
<td>1</td>
</tr>
<tr>
<td>P1</td>
<td>3</td>
<td>Practitioners</td>
<td>9</td>
<td>Buy or build critical component</td>
<td>2</td>
</tr>
<tr>
<td>P2</td>
<td>3</td>
<td></td>
<td>9</td>
<td>Communication system</td>
<td>1</td>
</tr>
<tr>
<td>P3</td>
<td>4</td>
<td></td>
<td>3.66</td>
<td>Operating system</td>
<td>2</td>
</tr>
<tr>
<td>P4</td>
<td>3</td>
<td></td>
<td>6</td>
<td>Programming language</td>
<td>2</td>
</tr>
</tbody>
</table>
As an example on the execution of the sessions, participants in S1 indicated concerns such as ‘low price’, ‘high security’, ‘high level of customer service’, and ‘low learning curve’. For S1, step two of GADGET resulted in seven alternatives (e.g. SAP Business One, Microsoft Dynamics, NetSuite) and eleven concerns for the first session. In Step three of GADGET, members of S1 prioritized concerns using the hundred-dollar approach. In addition, participants rated each alternative against each concern on a one-to-five scale, indicating how well an alternative satisfies a concern. Participants were familiar with some of the consolidated alternatives, but needed more time to learn about the others. During the session, they searched for information on the alternatives on the internet, and used the results for the ratings. In Step four of GADGET, members of S1 discussed the differences between the values they assigned, starting with the ratings that had the highest ranges. Participants discussed 14 ratings during the only iteration of the process. Participants reached consensus for eleven ratings.

Finally, we spent 20 minutes to obtain feedback on GADGET through a group discussion. We encouraged participants to provide feedback on their experiences, using the questions in Table 7.1.

7.3.2.2. Analysis Results

Next, we present the results of the content analysis, for the three categories corresponding to RQ1, RQ2, and RQ3.

**RQ1 - Need for consensus in group architectural decision making**

Regarding occurrences of conflicting perspectives (item 1 in Table 7.1), two architects indicated that conflicting perspectives related to a decision do not occur very often, and four architects indicated that they occur very often. Increasing the number of decision makers increases the number of conflicting perspectives, since decision makers have different priorities for concerns, and tradeoffs need to be found.

From the content analysis, we identified a positive and a negative impact of conflicting perspectives (item 2 in Table 7.1). On the one hand, participants indicated that conflicting perspectives is often time consuming (as one architect phrased it: ‘long and often almost endless discussions’). On the other
hand, participants indicated that the outcome of the decision is better if there are conflicting perspectives, because it encourages decision makers to address concerns of more stakeholders.

Regarding approaches for increasing consensus (item 3 in Table 7.1), from the content analysis we learnt architects lack structured approaches. Instead, architects use unstructured group discussions to increase consensus.

Overall, there is a need for increasing consensus in group architectural decision making in a systematic way, since 1) conflicting perspectives occur in practice, 2) conflicting perspectives help make better decisions, and 3) architects lack structured approaches for increasing consensus.

**RQ2 - Effort and benefits**

Regarding effort, we observed that GADGET requires one to three hours per group. Regarding benefits, the main benefit that emerged from the content analysis was increasing consensus among decision makers on the architectural decision. This benefit was indicated by five participants. A participant in the first session expressed this: ‘that’s what I really liked about the process: not focusing on the decision making in the first place, but on agreeing on a viewpoint.’ Additionally, a participant stated: ‘we learnt from it, you see other points of view, you also see your own gaps and misconceptions.’ The overall message from participants was that GADGET helped them increase consensus, by developing an increased shared understanding of each other’s perspectives, as a result of discussing the differences between them in a structured manner.

Several other additional benefits emerged from the content analysis:

1. **Increased focus** of the group discussions (appearing three times in the content analysis). According to a participant, decision makers are ‘less likely to run off-topic’. Moreover, participants considered that the process offered a structured way of increasing consensus, with prioritization of items for discussion, allowing them to ‘focus on stuff that is important.’

2. **Rationale** – participants appreciated that GADGET helps them capture the rationale for the decision, in addition to making the decision. Specifically, GADGET provides the rationale through its metrics, and maps concerns to participants. Therefore, architects can
see not only the outcome of the group decision, but also the intermediary steps that lead to the outcome.

3. **Reusability** – participants indicated that GADGET output (i.e. alternatives, concerns, and ratings) has high potential for reusability. For example, after making a group decision with GADGET, if a decision on the same topic needs to be made in the future, then alternatives, concerns, and ratings may be reused. In addition, some concerns may be reused across different decisions, especially across decisions that have strong dependencies (e.g. security-related concerns are reusable across most decisions for architecting a security-intensive system).

4. **Clarity of problem** – architects indicated that GADGET helped them clarify their point of view on the decision, by forcing architects to make explicit what matters to them in the decision.

**RQ3 – Improvements**

During the case study with the first group of participants, they indicated the need for increasing consensus on the priorities of concerns. Therefore, we updated GADGET to include prioritization of concerns (i.e. step three of GADGET), and we used the updated version of GADGET with the rest of the groups.

Here are the additional improvements suggested by participants throughout the sessions, and what we did about them:

- Participants suggested to optimize the time needed to use GADGET, by avoiding idle time in a face-to-face meeting, which happens when participants need different amounts of time to finish a step. For example, step three of GADGET (i.e. prioritize and rate concerns, see Section 7.2) can take place outside of a face-to-face meeting. Based on this suggestion, we removed time constraints on using GADGET in face-to-face meetings.

- Allow decision makers to eliminate less promising alternatives in later iterations. Based on this suggestion, we made explicit in the GADGET description (see step one in Section 7.2) that decision makers can also indicate what alternatives and concerns to remove when iterating.

- Participants considered that spreadsheets lacked dedicated features, such as the ability to trace divergent perspectives among decision makers. One of the architects indicated that he "wants to spend most of
the time on discussions, instead of working with the tool.’ We used this feedback for developing dedicated, user-friendly tool support for GADGET in Chapter 8.

7.3.3 Discussion

The exploratory case study offered us insights on GADGET. The increase in consensus from using GADGET was visible not only in the input from participants (e.g. ratings), but also in the feedback from participants. For example, a participant mentioned: ‘I trust the knowledge my teammates have from their respective fields. After noting they are more informed than I am, I would gladly accept their vision of the alternative, and I would concede to their rating.’ Additionally, other participants mentioned that strong arguments from peers in their groups convinced them to adjust their ratings.

Overall, the benefits of GADGET include: increased focus of the discussions, captured rationales of the decisions, potential for reusability of captured knowledge on decisions, and time savings. Still, there is further room for improving GADGET: offering additional prioritization approaches for concerns, and adding confidence levels to ratings. Also, tool support for GADGET needs to be user-friendly (i.e. low learning curve, and reducing the time required to learn and use GADGET).

7.3.3.1. Recommendations for Practitioners

From our experience with using GADGET, we recommend the following:

- Regarding threshold values for step four of GADGET (i.e. discuss differences), the recommended thresholds guideline values for differences are one for ratings and ten for priorities
- Regarding the number of iterations, two iterations for GADGET provide sufficient opportunities for decision-makers to reach consensus (i.e. general agreement on the decision, and mutual understanding of each other’s perspectives)
- GADGET is particularly useful when the following conditions are met:
  - The topic of the architectural decision is important enough for a group decision.
7.4 GADGET Experiment

o The architectural decision has several promising alternatives, so that spending time to evaluate them systematically is worthy.
o The decision makers have the maturity and openness to adopt and apply a systematic approach for their decision.

7.3.4 Implications for Research

Although there is a need for consensus in group architectural decision making, when making group architectural decisions, decision makers typically do not use any structured approach for increasing consensus. This means that decision makers use an 'as-is' or 'natural' approach which occurs when decision makers increase consensus without using any predefined approach. We call this approach ADHOC - the approach of increasing consensus in group architectural decisions without using any structured approach. Overall, the ADHOC approach seems to be popular in practice.

Exploratory case studies, such as the one we reported in this section, are useful for obtaining insights and generating hypotheses for further research (Yin, 2003). This case study brought initial evidence that GADGET increases consensus. Moreover, this case study helped us generate research questions and hypotheses for comparing GADGET with ADHOC, which we report in Section 7.4. Validity threats are reported in Section 7.5.

7.4 GADGET Experiment

The exploratory case study offered insights and initial evidence into the need for increasing consensus in group architectural decisions, as well as the effort and benefits offered by GADGET. One of the insights was that, in practice, consensus is often increased without using any structured approach (i.e. ADHOC). Therefore, we conducted an experiment to compare GADGET (i.e. a new approach) with ADHOC (i.e. the existing frequently used approach). This comparison allows drawing conclusions whether GADGET improves the current state of practice. Next, we report the experiment using the guidelines from (Jedlitschka et al., 2008).

In this experiment, we used ADHOC (as motivated in the previous section) for the control groups, and GADGET for the treatment groups. By comparing GADGET with ADHOC, we could better understand if GADGET increases
consensus, compared to ADHOC. This was a further research step compared to the exploratory case study in Section 7.3, in which we brought initial evidence that GADGET increases consensus, but we did not compare GADGET with another approach.

We chose to compare GADGET with ADHOC, instead of another process, for two reasons:

1. **Practical relevance.** Since ADHOC is popular in practice (as found in the case study in Section 7.3), the comparison with ADHOC helps practitioners understand what they can expect from adopting GADGET.

2. **Lack of a reference process.** As we found out in Chapter 4, there is no reference process in the literature for group architectural decision making to use as a baseline for comparison.

### 7.4.1 Research Goal and Questions

The goal of the experiment was to compare GADGET with ADHOC for the purpose of understanding them with respect to their impact on consensus among decision makers from the viewpoint of decision makers, in the context of group decision making for software architecture.

From our research goal, we derive the following two research questions.

**RQ1. Compared to ADHOC, what is the impact of GADGET on increasing consensus among group architectural decision makers?**

**Rationale:** This research question aims at offering evidence on how GADGET compares against ADHOC at increasing consensus among decision makers. In the case study in Section 7.3, we found that GADGET has the potential to increase consensus. However, an ad-hoc and unsystematic approach (i.e. ADHOC) can also help achieve consensus. If ADHOC has the same effect as GADGET, then it makes little sense for decision makers to use GADGET, since ADHOC has less overhead than GADGET.

**RQ2. How do perceptions on GADGET vs. ADHOC differ among decision makers?**

**Rationale:** The perception of an approach influences strongly the actual intention to use that approach (Venkatesh and Davis, 2000). A positive
perception of an approach likely leads to a higher intention to use the approach, which, in turn, results in actual usage of the approach. For example, if some architects perceive that GADGET brings benefits such as capturing rationale and correctness, without significant extra effort, then these architects are likely to use GADGET in their future activity. Therefore, understanding the perceptions on GADGET helps us understand the actual potential future usage of GADGET.

We present the metrics for answering RQ1 and RQ2 in Sections 7.4.4 and 7.4.5.

7.4.2 Participants

There are certain constraints when selecting participants for experiments. If the experiment has insufficient participants, then it is difficult to obtain relevant results. Also, if the sample is not representative enough, then the results of the experiment can be debated. However, a trade-off needs to be made between the number of participants and their representativeness. Kubickova and Ro (Kubickova and Ro, 2011) indicate that students are used as research subjects in an increasingly large number of scientific studies in various disciplines (e.g. in 80% of consumer research studies), despite continuous debates which have been going on for several decades on the scientific value of using students as research subjects (Kubickova and Ro, 2011).

Such debates also exist in software engineering research. A study on freshmen, graduate students, and industry people found no conclusive results on differences between these types of participants (Runeson, 2003). Another study suggests that students ‘may work well’ as subjects for software engineering studies (Svahnberg et al., 2008).

We chose to use a high number of participants with a good-enough representativeness for inexperienced software architects, who can benefit much from a structured approach for increasing consensus in their group architectural decisions. Furthermore, since we aim at establishing causal relationships, using students is preferable than using practitioners: students help reducing variations and thus confounding factors, so they help increase the internal validity of the study. Participants in our experiment were graduate and undergraduate software engineering students, who took a Software
Architecture course, in which they were presented the concept of architectural decisions. We conducted the experiment with students from three universities: University of Groningen in Netherlands, University of Vienna in Austria, and University of Pretoria in South Africa. We also describe the practical experience of students in software engineering (see Section 7.4.6) in order to interpret the results of the study according to the background of the students.

For validity and ethical purposes, we ensured that students had commitment for the study, and that the study contributed to participants’ education, as recommended by (Berander, 2004). To this end, we followed a checklist for integrating student empirical studies with our research and teaching goals (Carver et al., 2009; Galster et al., 2012). Below we present several items from Carver’s checklist for our study.

1. **Ensure adequate integration of the study into the course topics.**
   The course lectures stressed the importance of making architectural decisions. In the introduction of the experiment session, we explained to students how the session helps them make architectural decisions.

2. **Write up a protocol and have it reviewed.** We prepared the set of steps to follow and discussed them with two other researchers not involved in the study. The ethics committee from the University of Pretoria reviewed the protocol and approved it, with minor modifications (e.g. use a random id, instead of initials on the forms). Reviews of ethics committees from the other universities were not required.

3. **Obtain subjects’ permission for their participation in the study.**
   We told students about the session at least one week in advance. We also told students that the session addresses advanced topics in architecture, and that participation is voluntary, with no influence on their grades. By showing up for the session, students consented to participate. In addition, students from the University of Pretoria signed a consent form to indicate explicitly their consent, as required by the ethics committee.

4. **Build or update a lab package.** We built a lab package for the experiment, so that it could be replicated by other researchers. We developed the lab package at the University of Groningen. Later on,
researchers from University of Vienna and University of Pretoria used the same lab package to replicate the experiment.

7.4.3 Experimental Materials and Tasks

The lab package included the following experimental materials:

1 **Experimental case.** We used a predefined case with the purpose of putting students in roles in which they had to make a group architectural decision during the session. The case was based on a real-world architectural decision, which we learnt about from interviewing one of the original decision makers (Tselios et al., 2012). The case had a five-page description with all the details that students needed to make the group decision: description of the organization, roles, problem, concerns, and alternatives. The case included three decision maker roles: Department Manager, IT Architect, and Business Analyst. Each student took one of the roles during the experiment.

2 **Tasks descriptions.** Each student received descriptions of the tasks that included detailed instructions of the steps.

3 **Shared spreadsheet.** Students who used GADGET received access to a shared Google spreadsheet. Each group received a separate spreadsheet. Each spreadsheet included GADGET-specific fields (e.g. ratings, priorities) for each decision maker, and instructions on how to use the spreadsheet.

4 **Post-questionnaire on perceptions.** At the end of session, students filled out a post-questionnaire about their educational background and experience, as well as their perceptions on various aspects of the group decision making process (detailed in Section 7.4.5).

5 **Post-questionnaire on consensus.** This questionnaire included questions about prioritizing concerns and rating how well the alternatives satisfy concerns. Students filled them out from their role’s point of view, but also from the perspective of the other two group members and how they would fill them out. For example, a student could indicate a set of concerns’ priorities for her role, a different set of concerns’ priorities for one of her colleagues, and a totally different
set of concerns’ priorities for the other colleague. We explain further the rationales for these measurements in Section 7.4.4.

Table 7.3 shows an example of an item from the post-questionnaire on consensus for capturing an IT Architect’s point of view. The topic of the architectural decision described in Table 7.3 is choosing the newsletter system that an organization is using for communicating with its customers. Alternative A is to replace the current legacy system with a third-party software-as-a-service solution. Alternative B is to pay a partner to develop a new, modern system. Alternative C is to use an open source platform and various plugins. Alternative D is to enhance the current legacy system. Students who had the role of IT Architects filled out this item with their own values for priorities of concerns (whose sum had to be 100). In addition, students filled out ratings from one to five, indicating strong disagreement, disagreement, neutral, agreement, or strong agreement on how well each of the alternatives described in the case (i.e. A, B, C, and D) satisfied each of the concerns.

To help students maintain their focus throughout the experiment, we simplified the post-questionnaire on consensus. We asked students to rate alternatives from the other roles’ points of view for the ratings of two concerns, instead of six concerns. Thus, post-questionnaire items for IT Architects’ point of view only had the last two rows (i.e. cost-efficient, training time), while the items for the Business Analyst role included only the first two rows, and the items for the Department Manager included only the middle two rows. This simplification helped us reduce the risk of obtaining random data as a potential reaction to being asked to perform a tedious task, by helping students to maintain their focus.

Table 7.3. Example of post-questionnaire item for capturing an IT Architect’s priorities of concerns, and ratings of the four alternatives (i.e. A, B, C, and D) against two concerns.

<table>
<thead>
<tr>
<th>Concerns</th>
<th>Priorities</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher security</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Better delivery time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easily scalable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More cost-efficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Better training time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 7.2 shows the steps of the experimental process. First, we made a short presentation with the plan for the session, and an overview of tasks. Second, we selected students randomly to form groups of three students, since architectural decisions involve typically three persons, as presented in Chapter 4. When the number of students was not divisible by three, we included each extra student in existing groups. Third, we distributed the groups into two groups: half of the participants remained in the same room (control group), and the other half went to a different room (treatment group). Fourth, students read the experimental case and tasks descriptions. Fifth, students made the group decisions. Finally, students filled out the post-questionnaires on perceptions and consensus. During the session, we were available to answer questions from students, if necessary.

![Experimental Process Diagram]

Figure 7.2. Students followed the above steps for the experimental process.

In general, for an experiment, a null hypothesis \((H_0)\) states that the treatment causes no difference (e.g. using GADGET does not make any difference when compared to an ad-hoc decision making approach). The alternative hypothesis \((H_1)\) states that the treatment makes a difference (e.g. GADGET may help or hinder reach consensus, compared to an ad-hoc approach) (Wohlin et al., 2012). Based on the analysis of the data from the experiment, the null hypothesis can be rejected and the alternative hypothesis can be accepted. The analysis uses statistical tests to determine statistically significant differences between the data from the control group (e.g. ADHOC) and data from the treatment group (e.g. GADGET). Next, we present the hypotheses, including
their null and alternative hypotheses, on the differences caused by the treatment in our experiment (i.e. GADGET).

7.4.4 Hypotheses for RQ1 – Consensus

To answer RQ1, we define metrics for operationalizing consensus among decision makers. As mentioned in Section 7.1, we consider two components of consensus: general agreement and mutual understanding. We define hypotheses and metrics on both components of consensus.

7.4.4.1. Hypothesis on General Agreement

Regarding general agreement, we defined a metric that counts how many groups reached agreement on their group architectural decision. For example, if no group reached agreement on their group architectural decisions, then this metric is zero. Using this metric, we propose the following hypothesis.

\( H_{a0} \): ADHOC and GADGET result in the same general agreement among group decision makers.

\( H_{a1} \): GADGET results in higher general agreement than ADHOC.

7.4.4.2. Hypothesis on Mutual Understanding on the Priorities of Concerns

Regarding mutual understanding among group decision makers, a group has high mutual understanding on a decision, if group members are also able to indicate accurately the perspectives of the other group members on that decision. For example, let us consider three architects (Anne, Bob, and Charlie) who need to make a group architectural decision on which framework (e.g. A, B, C, or D) to use for a new software system. High mutual understanding among the three architects means that, after discussions, each of the three architects is able to indicate accurately what the other two architects think about the performances of each framework. In contrast, low mutual understanding means the input from the other group members was not taken seriously, which resulted in misunderstandings among architects on each other’s perspectives (e.g. at the end of the discussion, Charlie has no idea what Anne thinks about the performance of the C framework, although Anne mentioned this during the discussion).
Priorities of concerns are a ratio type of data, which means that calculating differences between priorities is allowed. For the metric related to the mutual understanding on the priorities of concerns, we calculate the sum of absolute differences between the priorities assigned by a student, and the priorities that the student’s group colleagues estimated. Based on these assumptions, equation (1) summarizes the metric for calculating mutual understanding on priorities (MUP) of concerns, for a decision with six concerns (see Table 7.3) in a group of three decision makers. $p_{ai}$ stands for the priority indicated by architect A for the $i$ concern, from A’s point of view. $p_{ij}$ stands for the priority estimated by colleague $j$ for the $i$ concern, as colleague $j$ estimates that A indicated. MUP ranges from 0 to 100. Lower values for the metric mean higher mutual understanding among group decision makers, due to smaller differences between estimated and actual priorities.

$$MUP_a = \sum_{j=1}^{2} \sum_{i=1}^{6} |p_{ai} - p_{ij}|$$

Using the above metric, we propose the following hypothesis.

$H_{b0}$: ADHOC and GADGET result in the same level of mutual understanding on priorities of concerns among group decision makers.

$H_{b1}$: GADGET results in higher mutual understanding on priorities of concerns than ADHOC.

### 7.4.4.3. Hypothesis on Mutual Understanding on Ratings

Ratings of alternatives are provided on a 5-point Likert scale, which may be considered an ordinal type of data. This means that summing differences among ratings (similar to eq. (1) in Section 7.4.4) is problematic. Instead of summing differences among ratings, we use the standard deviation to measure the variation among ratings. Similar to the metric for priorities, we calculate the standard deviation for one’s own ratings, and the ratings that the other decision makers in the group estimated for one’s ratings. Lower values for the standard deviation indicate higher mutual understanding on ratings among group decision makers, due to smaller variation between estimated and actual priorities.

Using the standard deviation metric, we propose the following hypothesis.
7. Improve Group Architectural Decisions

$H_{c0}$: ADHOC and GADGET result in the same level of **mutual understanding on ratings of alternatives against concerns** among group decision makers.

$H_{c1}$: GADGET results in higher **mutual understanding on ratings of alternatives against concerns** than ADHOC.

### 7.4.5 Hypotheses for RQ2 - Perceptions

To answer RQ2, we defined metrics to measure the perceptions of the group decision makers about the process they use (i.e. GADGET or ADHOC). Based on existing literature, we propose three categories of perceptions: on benefits of using GADGET, challenges related to the use of GADGET, and satisfaction from using a group decision making process. For each category, we propose several perception items. Each perception item is operationalized by indicating the level of agreement with items in the post-questionnaire, using a five-point Likert scale (i.e. from strong disagreement to strong agreement). The items in the post-questionnaire originate from the initial GADGET evaluation in Section 7.3, and literature on decisions. Table 7.4 shows the perception categories, perception and post-questionnaire items, as well as the literature source for the items.
Table 7.4: Mapping of perception categories, metrics, and post-questionnaire items.

<table>
<thead>
<tr>
<th>ID</th>
<th>Perception category</th>
<th>Perception metric</th>
<th>Post-questionnaire item</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Benefits</td>
<td>Reevaluation of initial perspective</td>
<td>After discussing the case with my team I changed my mind regarding the importance of one or more concerns (Hartwig, 2010; Schweiger et al., 1986)</td>
</tr>
<tr>
<td>M2</td>
<td>Revalues extra points</td>
<td>The discussion with my team revealed valid points that I would not be able to consider on my own (Hartwig, 2010; Schweiger et al., 1986)</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>Reusability</td>
<td>The artefacts (documents, notes, tables, spreadsheets, etc.) that my team created during the decision-making session could be reused to examine similar situations in the future (Jansen et al., 2007; Tang and van Vliet, 2009)</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>Rationale</td>
<td>The artefacts that my team created during the decision-making session could be used to justify to other people the reasons we made this decision (Jansen et al., 2007; Tang and van Vliet, 2009)</td>
<td></td>
</tr>
<tr>
<td>M5</td>
<td>Clarifies problem</td>
<td>After the decision-making session, my team had a clearer view on ASO’s problem (Lai et al., 2002)</td>
<td></td>
</tr>
<tr>
<td>M6</td>
<td>Improves decision making skills</td>
<td>The decision-making session improved my decision-making skills (Hardgrave et al., 2003)</td>
<td></td>
</tr>
<tr>
<td>M7</td>
<td>Challenges</td>
<td>Understandability of process</td>
<td>It was too difficult for me to understand what I was required to do (Hardgrave et al., 2003)</td>
</tr>
<tr>
<td>M8</td>
<td>Clarity of instructions</td>
<td>The instructions were clear enough (Hardgrave et al., 2003)</td>
<td></td>
</tr>
<tr>
<td>M9</td>
<td>Time for decision</td>
<td>I believe that the decision-making session required too much time (Chapter 5)</td>
<td></td>
</tr>
<tr>
<td>M10</td>
<td>Effort</td>
<td>I believe that the decision-making session required too much effort (Chapter 5)</td>
<td></td>
</tr>
<tr>
<td>M11</td>
<td>Preparation time</td>
<td>It took me too long to understand what I was required to do in the decision-making session (Chapter 5)</td>
<td></td>
</tr>
<tr>
<td>M12</td>
<td>Satisfaction for future collaboration</td>
<td>I would be willing to work with the same team on other projects in the future (Schweiger et al., 1986)</td>
<td></td>
</tr>
<tr>
<td>M13</td>
<td>Satisfaction on cooperation</td>
<td>Working together with my teammates was an enjoyable experience (Schweiger et al., 1986)</td>
<td></td>
</tr>
<tr>
<td>M14</td>
<td>Enjoyment</td>
<td>I enjoyed the decision-making session (Schweiger et al., 1986)</td>
<td></td>
</tr>
<tr>
<td>M15</td>
<td>Commitment</td>
<td>I strongly support my group’s final decision (Schweiger et al., 1986)</td>
<td></td>
</tr>
<tr>
<td>M16</td>
<td>Overall satisfaction</td>
<td>I am satisfied with my group’s decision (Schweiger et al., 1986)</td>
<td></td>
</tr>
</tbody>
</table>

Based on the 16 metrics in Table 7.4, we define 16 hypotheses, as follows. Since the hypotheses are similar and only the metrics vary, we formulate a generic hypothesis, which is adaptable to each of the 16 hypotheses.
7. Improve Group Architectural Decisions

$H_{M0}$: ADHOC and GADGET result in similar perceptions on the $M_i$ metric (where $M_i$ varies from M1 to M16), among group decision makers.

$H_{M1}$: ADHOC and GADGET result in different perceptions on the $M_i$ metric among group decision makers.

In summary, the independent variable for this experiment is the group decision making process (i.e. GADGET or ADHOC). The dependent variables for RQ1 and RQ2 are summarized in Table 7.5.

Table 7.5. Summary of dependent variables for each research question.

<table>
<thead>
<tr>
<th>RQ</th>
<th>Hypothesis</th>
<th>Metric description</th>
<th>Scale type</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>$H_{a0}$</td>
<td>General agreement</td>
<td>Nominal</td>
<td>Yes/no</td>
</tr>
<tr>
<td></td>
<td>$H_{b0}$</td>
<td>Sum of differences between priorities of concerns</td>
<td>Ratio</td>
<td>Zero or more</td>
</tr>
<tr>
<td></td>
<td>$H_{c0}$</td>
<td>Standard deviation of ratings</td>
<td>Ratio</td>
<td>Zero or more</td>
</tr>
<tr>
<td>RQ2</td>
<td>$H_{M0}$</td>
<td>Perception metrics</td>
<td>Interval</td>
<td>1 to 5</td>
</tr>
</tbody>
</table>

7.4.6 Results

The experiment took place in three sessions. The first session took place with 18 students at the University of Groningen. The second session took place with 72 students at the University of Vienna. The third session took place with 23 students at the University of Pretoria. All sessions took place in a similar manner, and no deviations from the protocol occurred. After performing the experimental sessions, we discarded data from 11 students, due to missing or incomplete values. The valid data from the remaining 102 students was analyzed as follows.

7.4.6.1. Analysis Procedure

To analyze the collected data, we defined analysis procedures for investigating the hypotheses in Sections 7.4.4 and 7.4.5. Table 7.6 summarizes the analysis procedures for all hypotheses. We used the Mann-Whitney U test because it is well suited for comparing two independents samples (i.e. the treatment/GADGET and control/ADHOC groups). Furthermore, this statistical test is non-parametric (i.e. it makes no assumption regarding the normal distribution of the data), which is suitable to this experiment, since we cannot
assume that the data is normally distributed. Still, we checked the normality of
the data using the Shapiro-Wilk test, to confirm the validity of using a non-
parametric test. We used IBM SPSS for applying statistical tests.

<table>
<thead>
<tr>
<th>Research question</th>
<th>Hypothesis keywords</th>
<th>Hypothesis number</th>
<th>Analysis procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1 Consensus</td>
<td>Agreement</td>
<td>$H_{a0} - H_{a1}$</td>
<td>Binomial test</td>
</tr>
<tr>
<td></td>
<td>Mutual understanding</td>
<td>$H_{b0} - H_{b1}$</td>
<td>Mann-Whitney U tests</td>
</tr>
<tr>
<td></td>
<td>(priorities of concerns)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mutual understanding</td>
<td>$H_{c0} - H_{c1}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(ratings of alternatives against concerns)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ2 Perceptions</td>
<td>Benefits, challenges and satisfaction</td>
<td>$H_{M0} - H_{M1}$</td>
<td>$M_i$ covers $M_1$ to $M_{16}$</td>
</tr>
</tbody>
</table>

7.4.6.2. Participants’ Background

Regarding background, we asked participants to indicate their number of years of practical experience in software engineering. Figure 7.3 summarizes the results. Five students declined to respond. Overall, a third of the students had more than one year of practical experience. Participants’ levels of experience are balanced across the treatment (i.e. GADGET) and control (i.e. ADHOC) groups.

Figure 7.3. Summary of the years of practical experience in software engineering of the students.
7.4.6.3. Answer to RQ1 - Consensus

To answer RQ1, we tested the three hypotheses on the two components of consensus (i.e. agreement and mutual understanding) summarized in Table 7.6. Regarding the hypothesis on agreement, we found that all groups from both treatments reached consensus. Therefore, we cannot reject the null hypothesis $H_{a0}$ (detailed in Section 7.4.4), and conclude that both GADGET and ADHOC result in agreement among group decision makers.

Table 7.7 summarizes the values for the hypotheses on mutual understanding on priorities of concerns and ratings. For example, the average values for the metrics on priorities (as defined in Section 7.4.4) were 133.91 for students in the control group (ADHOC), and 102.69 for students in the treatment group (GADGET). We checked the normality of the data using the Shapiro-Wilk test, and we found out that the data was not normally distributed ($p$-value = 0.011). The non-parametric Mann-Whitney U test on $H_5$ returned a statistically significant difference ($p$-value = 0.0003). Therefore, we reject the null hypothesis (i.e. $H_{a0}$ in Section 7.4.4), and conclude that GADGET results in higher consensus for priorities of concerns among group decision makers.

Regarding the hypothesis on mutual understanding on ratings of alternatives against concerns, we found lower standard deviations of ratings in the GADGET group. The average values for metrics on ratings (as defined in Section 7.4.4) was 1.29 for students in the control group (ADHOC), and 1.12 for students in the treatment group (GADGET). The Shapiro-Wilk test for normality indicated the data was not normally distributed ($p$-value = 0.015). The Mann-Whitney U test returned a statistically significant difference ($p$-value = 0.00001). Therefore, we reject $H_{o0}$, and we conclude that GADGET results in higher consensus for ratings among group decision makers.
7.4. GADGET Experiment

Table 7.7. Medians and means for ADHOC and GADGET for the metrics on mutual understanding.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Hypothesis keywords</th>
<th>Metric description</th>
<th>Median (mean) ADHOC</th>
<th>Median (mean) GADGET</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_b$</td>
<td>Mutual understanding on priorities of concerns</td>
<td>Sum of differences between priorities of concerns</td>
<td>130 (133.91)</td>
<td>95 (102.69)</td>
<td>0.0003</td>
</tr>
<tr>
<td>$H_c$</td>
<td>Mutual understanding on ratings</td>
<td>Standard deviation of ratings</td>
<td>1.31 (1.29)</td>
<td>1.17 (1.12)</td>
<td>0.00001</td>
</tr>
</tbody>
</table>

7.4.6.4. Answer to RQ2 - Perceptions

We tested the 16 hypotheses (in Section 7.4.5) on students’ perceptions on the GADGET and ADHOC approaches. Table 7.8 summarizes the results for the 16 hypotheses corresponding to M1 to M16. The Shapiro-Wilk test for normality indicated that data for all metrics was not normally distributed ($p$-value = 0.000). After applying Mann-Whitney U tests, we found statistically significant ($p < 0.05$) differences on eight metrics. We rejected $H_{M30}$, $H_{M40}$, $H_{M70}$, $H_{M90}$, $H_{M100}$, $H_{M120}$, $H_{M130}$, and $H_{M160}$. We accepted their corresponding alternative hypotheses: $H_{M31}$, $H_{M41}$, $H_{M71}$, $H_{M91}$, $H_{M101}$, $H_{M121}$, $H_{M131}$, and $H_{M161}$.

Table 7.8. Perceptions on ADHOC and GADGET (1 = strong disagreement, 5 = strong agreement). The shaded rows highlight the statistically significant results.

<table>
<thead>
<tr>
<th>ID</th>
<th>Perception category</th>
<th>Perception metric item</th>
<th>Median (mean) ADHOC</th>
<th>Median (mean) GADGET</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Benefits</td>
<td>Reevaluation of initial perspective</td>
<td>3 (2.90)</td>
<td>3 (3.17)</td>
<td>.192</td>
</tr>
<tr>
<td>M2</td>
<td></td>
<td>Reveal extra points</td>
<td>3 (3.33)</td>
<td>3 (3.17)</td>
<td>.375</td>
</tr>
<tr>
<td>M3</td>
<td></td>
<td>Reusability</td>
<td>3 (2.76)</td>
<td>4 (3.55)</td>
<td>.019</td>
</tr>
<tr>
<td>M4</td>
<td></td>
<td>Rationale</td>
<td>4 (2.86)</td>
<td>4 (3.77)</td>
<td>.005</td>
</tr>
<tr>
<td>M5</td>
<td></td>
<td>Clarify problem</td>
<td>4 (3.88)</td>
<td>4 (3.83)</td>
<td>.821</td>
</tr>
<tr>
<td>M6</td>
<td></td>
<td>Improve skills</td>
<td>3 (3.27)</td>
<td>3 (3.25)</td>
<td>.641</td>
</tr>
<tr>
<td>M7</td>
<td>Challenges</td>
<td>Understand process (negative statement)</td>
<td>1 (1.31)</td>
<td>2 (1.64)</td>
<td>.007</td>
</tr>
<tr>
<td>M8</td>
<td></td>
<td>Clear instructions</td>
<td>4 (4.27)</td>
<td>4 (4.15)</td>
<td>.352</td>
</tr>
</tbody>
</table>
7.4.7 Discussion

Controlled experiments are particularly useful for establishing causal relationships (Wohlin et al., 2012). In this experiment, we compared the impact of the group decision making approach (i.e. GADGET or ADHOC) on two components of consensus: mutual understanding and general agreement. We found out that GADGET performs better than ADHOC at increasing mutual understanding among decision makers, for both priorities of concerns and ratings of alternatives against concerns. We found no difference between GADGET and ADHOC at the general agreement.

Additionally, we found statistically significant differences between perceptions (RQ2) on GADGET vs. ADHOC as follows:

- Regarding perceptions on the benefits of GADGET vs. ADHOC approaches, reusability of created artefacts (e.g. alternatives, rationale) while using the approaches was significantly higher for GADGET. In addition, the GADGET approach allowed better capturing of the rationales for the architectural decisions than ADHOC. However, we found no significant differences on reevaluating the initial perspectives, revealing extra points, problem clarification, and improving decision making skills.

- Regarding perceptions on the challenges of using GADGET vs. ADHOC, we found the following significant differences. GADGET users had more difficulties understanding the process than ADHOC
users, which reflects the learning curve of GADGET. In addition, GADGET users perceived a higher time and effort to make decisions compared to ADHOC, which reflects the effort of using a structured approach for group decision making. However, we found no differences on the clarity of the instructions and the preparation time.

- Regarding perceptions on the satisfaction of using GADGET vs. ADHOC, we found significantly higher willingness for future collaboration with the same team members for ADHOC. Also, ADHOC users reported higher satisfaction on cooperation and overall higher satisfaction with their decisions than GADGET users. However, we found no significant differences on enjoying the session, and on one’s commitment to one’s group final decision.

### 7.4.7.1. Interpretation of Results

These findings mean the following:

- Regarding consensus among decision makers, this experiment indicates GADGET’s positive effect on increasing consensus. The combined evidence from the case study in Section 7.3 and the experiment in this section indicates that practitioners can use GADGET to increase consensus in their architectural decisions.

- Regarding the results on the benefits of GADGET vs. ADHOC, the results on reusability and capturing rationales in the experiment confirmed the results from the case study. These benefits help practitioners avoid architectural knowledge vaporization, and reduce maintenance costs. For the remaining four items on benefits (i.e., reevaluation of initial perspective, revealing extra points, clarifying problem, and improving decision making skills), the results in Table 7.8 indicate no differences between GADGET and ADHOC, which means these four items are not key benefits of GADGET.

- Regarding meaning of results on the challenges of GADGET vs. ADHOC, the results indicate there is a higher cost for decision makers in terms of time and effort for using GADGET. These results were obtained in the context of first-time users of GADGET and not-first time users of ADHOC (since participants were very likely to have made other group decisions before the experiment, given their years of experience, as shown in Figure 7.3). We can expect that the effort of using GADGET would decrease for subsequent uses, after passing its
learning curve. Still, the lack of differences on instructions clarity and preparation time (in Table 7.8) suggests that participants could learn about GADGET from the written instructions they received. Overall, although GADGET has a learning curve, we expect practitioners to progress fast on the learning curve.

- Regarding meaning of results on satisfaction on using GADGET vs. ADHOC, we note that ADHOC scored more favorably than GADGET. However, the results on GADGET still show positive satisfaction from using GADGET. Overall, practitioners who use GADGET for the first time can expect positive satisfaction, although lower than ADHOC, which is more familiar to practitioners.

### 7.4.7.2. Cross-study Discussion

From the case study and the experiment, we learnt that GADGET increases consensus among participants. Furthermore, GADGET helps make better decisions, by encouraging decision makers to evaluate systematically alternatives. Finally, GADGET reduces architectural knowledge vaporization by capturing the rationale of the group decision.

In general, the behavior of the decision makers can be explained by Bryson’s (Bryson, 1996) rationale for reaching consensus in a group. Bryson indicates that, during group decision making, some participants are in a learning mode and other participants are in a strategic mode. Participants in a learning mode are less certain on their preferences, and are willing to change their preferences. However, participants in a strategic mode are more certain on their preferences and less likely to change them.

In the case study and the experiment, we noticed the learning and strategic modes of participants while using the GADGET approach. However, these modes varied per alternative. For example, a participant of the case study was in a learning mode for an alternative, but in a strategic mode for a different alternative. Overall, GADGET ensured that participants in the strategic mode could present their arguments to participants in the learning mode, and help achieve consensus among participants.
7.5. Validity Threats

7.4.7.3. Limitations of GADGET

There are a few limitations for applying GADGET in practice. GADGET assumes participants in the group decision making are on a similar hierarchy level, and no politics are involved in the decision making. Other factors include social relationships among participants. For example, if the group has high cohesion, then the group decision making process might be easier to adopt and follow. Still, more work is needed to understand these limitations and their influence on the adoption and results of group decision making processes, such as GADGET.

7.5 Validity Threats

Using guidelines from (Jedlitschka et al., 2008) and (Wohlin et al., 2012), we present construct, internal, external, and conclusion validity threats for the case study (detailed in Section 7.3) and experiment (detailed in Section 7.4).

7.5.1 Case Study Validity Threats

Construct validity - to avoid this threat, we conducted the case study not only with students (two groups), but also with practitioners (four groups). Furthermore, we prevented interviewer (i.e. to please researchers) and response biases (i.e. responses that make participants look good) by encouraging participants to criticize GADGET openly. In turn, this helped us collect areas for improvement, as reported in Section 7.3.2.2. Finally, participants were anonymized and had no incentive (e.g. grades, money) to please researchers.

Internal validity threats were not applicable for the case study, since we did not attempt to show any causality relationship.

External validity - to address this threat, we involved practitioners in the case study. Furthermore, the students who participated in the case study also had practical experience.

Conclusion validity - the study conclusions were drawn based on the results from the content analysis of interviews with participants, using guidelines from the literature (Krippendorff, 2004). To ensure accurate conclusions, two researchers were involved in the content analysis of the interviews with
participants. The researchers made sure that there was high agreement in their interpretation of the data.

7.5.2 Experiment Validity Threats

We addressed **construct validity** by operationalizing the constructs in our experiment; we defined metrics for each hypothesis (see Sections 7.4.4 and 7.4.5). Furthermore, to avoid impact on participants’ behavior, we made clear to the participants that the experiment would not have any impact on their grades. Additionally, to avoid hypotheses guessing and evaluation apprehension, we did not tell participants our hypotheses.

To address **internal validity** threats, such as the instrumentation validity threat, we made a pilot for the experiment (Tselios et al., 2012), to increase the clarity of the experimental package. For example, we increased the readability of the paper forms, so that subjects can easily understand their tasks. We addressed the **mortality** validity threat by integrating the study with the software architecture course (see Section 7.4.2), so that subjects joined it voluntarily for the educational value. We distributed participants randomly to the groups to avoid selection threats. Furthermore, by using students we increased internal validity, since using practitioners means larger variation in confounding variables such as domains, types of previous projects, or previous experiences.

Another instrumentation validity threat is that students took roles (i.e. department manager, IT architect, or business analyst) for which they had little or no experience. To address this threat and to avoid relying on the experience of participants, we gave each student printouts with the description of their corresponding role. This description contained all the information they needed to make the decision and to participate in the group decision process. This, it was not necessary that students required external sources of information during the experiment, or previous experience.

Regarding **external validity**, Kitchenham et al. regard students as relatively close to the population of interest, because they are the next generation of software professionals and close to novice software developers (Kitchenham et al., 2002). We consider our results as applicable to inexperienced architects, rather than senior architects. Since inexperienced architects need more support
than senior architects, it is reasonable to use students in the experiment, instead of senior architects. Moreover, the nature of tasks students had to perform did not require experience levels of senior architects, as students had sufficient knowledge to perform their tasks. To ensure the commitment of the participants, we made sure that the experiment contributes to participants’ education (see Section 7.4.2). To check whether or not GADGET is also applicable to more experienced or senior architects, we need to conduct a future similar experiment with practitioners.

Regarding **conclusion validity**, statistical tests have various assumptions, and violating them may lead to poor conclusions. We used non-parametric tests that make fewer assumptions, such as Mann-Whitney. By conducting the experiment with a large sample of students from multiple universities, we aimed at increasing tests’ statistical power. Another potential threat is that some metrics (e.g. perceptions) tend to be less reliable than others (e.g. ratings). To address this threat, we piloted our study (Tselios et al., 2012) to clarify wording, and avoid misunderstandings.

### 7.6 Related Work

There have been a few approaches and studies on group architectural decisions. Zannier et al. describe real-world architectural decisions, and ask for more work on understanding real-world group architectural decisions (Zannier et al., 2007). Kazman et al. propose an extension of CBAM (Kazman and Klein, 2001) that considers explicitly the preferences of group architectural decision makers (Kazman et al., 2005). Recently, Rekha and Muccini analyze real-world group architectural decision making (Rekha and Muccini, 2014). Nowak and Pautasso analyze situational awareness in group architectural decision making (Nowak and Pautasso, 2013b). Gaubatz et al. propose automatic enforcements of constraints in group architectural decisions (Gaubatz et al., 2015). In this chapter, we focus on a particular aspect of group architectural decision making (i.e. increasing consensus), which has not been addressed in previous work.

Related work on processes for group architectural decision making include the following. Babar et al. studied the feasibility of groupware support for architecture evaluation, with applicability on architectural decisions (Babar et
improve group architectural decisions (al-naeem et al., 2006). al-naeem et al. propose using the analytical hierarchy process in group architectural decision making (al-naeem et al., 2005). nakakawa et al. propose a theoretical model on group architectural decision making for enterprise software systems (nakakawa et al., 2010). sousa et al. present a process for group architectural decision making, in which a facilitator helps the group interactions (sousa et al., 2006). in this chapter, the proposed gadget process does not require a facilitator, while our focus is on presenting empirical evidence on the gadget process.

related work on approaches that capture architectural knowledge and help group architectural decisions include the following. falessi et al. reported an experiment with students on documenting the rationales of group architectural decisions (falessi et al., 2006). mohan and ramesh propose a traceability framework for group architectural decisions (mohan and ramesh, 2007). zimmermann et al. propose a framework for capturing architectural decisions which can help group architectural decisions (zimmermann et al., 2009). in this chapter, we provide evidence that the gadget process reduces architectural knowledge vaporization.

tang (tang, 2011) mentions communication issues that may appear in group architectural decision making, but no process improvement is offered. also, kazman et al. (kazman et al., 1999) describe the importance of consensus for the atam approach, but without describing how to increase consensus for architectural decisions. furthermore, the attribute-driven design method (wojcik et al., 2006) does not indicate how to increase consensus in group architectural decisions. in contrast, in this chapter we provide evidence on how gadget increases consensus in group architectural decisions.

7.7 conclusions

in this chapter, we propose and evaluate gadget, a process for increasing consensus in group architectural decisions. consensus is conceptualized in terms of its two main components: general agreement and mutual understanding. gadget is based on our previous work on using the repertory grid technique to capture architectural knowledge in chapters 5 and 6, as well as the delphi technique (linestone and turoff, 2002). to evaluate gadget thoroughly, we conducted a case study and an experiment, which
increases the validity of our findings. Thirteen practitioners and eight students participated in the case study, and 113 students participated in the experiment.

From the case study, we identified the need for increasing consensus in group architectural decisions. In addition, we found that GADGET helps practitioners increase consensus in group architectural decisions. From the experiment, we found that GADGET and ADHOC resulted in agreement among group decision makers, while GADGET resulted in higher mutual understanding than ADHOC. GADGET provides significantly higher reusability of architectural decisions and more captured rationales than ADHOC. However, GADGET requires more effort than ADHOC.

The results of the two studies in this chapter indicate that GADGET helps practitioners, and particularly inexperienced architects to increase consensus in group architectural decisions, and capture the rationales of architectural decisions. Still, group architectural decision making is a multifaceted topic, since in practice group decisions can be influenced by factors such as hierarchy levels, hidden agendas, or politics. Such factors were out of scope for this chapter. Overall, for architectural decisions in which such factors do not play a role, GADGET is particularly useful for increasing consensus in group architectural decisions and capturing the rationales of the decisions.

The studies in Chapters 6 and 7 encouraged us to offer tool support for REGAIN and GADGET, so that practitioners can easily use REGAIN and GADGET. In the next chapter, we report our efforts towards implementing such tool support.

7.8 Acknowledgments

We thank Konstantinos Tselios for his help in conducting the studies, and participants for their efforts.