TECHNICAL REPORT

RT – ES 753 / 17

Evidence of Usage-Based Reading Effects by Using the Structured Synthesis Method (SSM)

Paulo Sérgio Medeiros dos Santos
(pasemes@cos.ufrj.br)

Guilherme Horta Travassos
(ght@cos.ufrj.br)

PESC
Programa de Engenharia
de Sistemas e Computação

Systems Engineering and Computer Science Department

COPPE / UFRJ

Rio de Janeiro, Julho 2017

ABSTRACT

In this technical report, we present an example of the Structured Synthesis Method (SSM). For this example, we chose the classical domain of Software Inspection, in this case, the UBR inspection technique. This domain was deliberately chosen because it is a well-known domain in SE, particularly within the Empirical Software Engineering community where it has been extensively investigated and was one of the first topics to be the target of experimental studies. Thus, we used this in an attempt to draw attention to the application of the SSM method itself rather than to the synthesis results.

For details regarding the SSM, refer to Santos P.S.M., Travassos G.H. (2013) On the Representation and Aggregation of Evidence in Software Engineering: A Theory and Belief-based Perspective. Electronic Notes Theoretical Computer Science 292:95–118. doi: 10.1016/j.entcs.2013.02.008

RESUMO

Neste relatório técnico, apresentamos um exemplo de uso do Método de Síntese Estruturado (SSM). Para este exemplo escolhemos um domínio clássico das Inspeções de Software, neste caso particular, a técnica de inspeção UBR. Este domínio foi escolhido devido a ser bem conhecido em Engenharia de Software, particularmente pela comunidade de Engenharia de Software Experimental onde tem sido largamente investigado e foi um dos primeiros objetos de estudo da área. Com isso, chamamos atenção para o uso do SSM mais do que o entendimento do problema ou a síntese em si.

Para detalhes sobre o SSM, consulte Santos P.S.M., Travassos G.H. (2013) On the Representation and Aggregation of Evidence in Software Engineering: A Theory and Belief-based Perspective. Electronic Notes Theoretical Computer Science 292:95–118. doi: 10.1016/j.entcs.2013.02.008

1. Introduction

Inspection of software artifacts is a meaningful way of avoiding rework and improving software quality (Fagan 2002). The primary factors for this success are the relatively low cost of utilization and its capability in finding defects throughout the process. Moreover, software inspections can integrate the defect prevention and detection process.

Several aspects can influence the inspection cost-efficiency (number of defects per unit of time) and the types of defects identified. The characteristics related to the inspector level of experience or its technical specialty (e.g., programmer or tester) are usually cited among these aspects. As a consequence, an ad hoc inspection, in which there is no control over the inspector procedures, has an individualized cost-efficiency and no guarantee that of an adequate coverage of the artifact content or the types of defect (Porter and Votta 1998).

Usage-Based Reading (UBR) is an inspection technique whose primary goal is to drive reviewers to focus on crucial parts of a software artifact from the user's point-of-view. In UBR, faults are not assumed to be of equal importance, and the technique aims at finding the faults that have the most negative impact on the users' perception of system quality. For this, reviewers are given use cases in a prioritized order and inspect the software artifacts following the usage scenarios defined in the ordered use cases. Therefore, a central aspect on focusing inspection effort in UBR is the prioritization of use cases. UBR assumes that the set of use cases can be prioritized in a way reflecting the desired focusing criterion. If the inspection aims at finding the faults that are most critical to a certain system quality attribute, the use cases should be prioritized accordingly.

In this paper, we present a worked example of the Structured Synthesis Method (SSM). For this example, we chose the classical domain of Software Inspection, in this case, the UBR inspection technique. This domain was deliberately chosen because it is a well-known domain in SE, particularly within the Empirical Software Engineering community where it has been extensively investigated and was one of the first topics to be the target of experimental studies. Thus, we used this in an attempt to draw attention to the application of the SSM method itself rather than to the synthesis results. For details regarding SSM, we refer to Santos and Travassos (2013).

The next section describes the synthesis regarding the five-stage process of SSM. All details of this research synthesis, particularly the theoretical structures and the description of the constructs, can be found in the Evidence Factory tool at http://evidencefactory.lens-ese.cos.ufrj.br/synthesis/editor/291. A presentation of the tool features can be found in Santos et al. (2015). We end this report with final remarks regarding the aggregation.

2. The Usage-Based Reading Synthesis

2.1 Planning and definition

Using the structure suggested in SSM, the research question was defined as follows:

> *What are the expected effects from Usage-Based Reading inspection technique when it is applied for inspecting high-level design artifacts produced in the analysis phase of the software development process?*

The research question incorporates aspects related to technology, activity, and system leaving out any consideration of the actors' characteristics. Thus, no characteristics about organization, team or persons, such as software development experience, are determinant for the studies selection.

We defined 'Usage-Based Reading' as the only term of the search string. It was possible because UBR is a very specific software technology. Therefore, making the search string more detailed would only add the risk of leaving out papers, which did not include terms about the defined activity and system characteristics. As a result, we decided to consider the aspects of activity and system characteristics in the paper inclusion criteria. For exclusion criteria, on the other hand, we eliminated theoretical or analytical papers and articles not written in English. The last definition for paper selection is the digital libraries to be used, which in this case was Scopus (http://www.scopus.com).

2.2 Selection

We were able to find 15 technical papers in Scopus with the given search string, from which four were selected following the inclusion and exclusion criteria. The selection was performed in November'15. Among the excluded papers, one was a duplicate, one classified as theoretical (analyzing the contributions of three included papers), and the others did not fulfill the inclusion criteria.

The four included studies form a family of experiments aiming at investigating UBR performance in identifying faults on software artifacts. Two researchers participated in three of them. The first experiment (Thelin et al. 2001 – Study S1 – 27 participants) compared UBR with the ad-hoc inspection. Moreover, the other three studies (Thelin et al. 2003 – Study S2 – 34 participants), (Thelin et al. 2004 – Study S3 – 23 participants) and (Winkler et al. 2004 – Study S4 – 62 participants) compared UBR against a checklist based reading (CBR).

2.3 Quality assessment

Following SSM definitions, quality assessment was performed with quality checklists. Based on the study type, as all studies are *quasi*-experiments, the belief values for them have an inferior limit of 0.50. Then, we add to that base value the result from the scoring scheme for systematic studies. Table 1 presents the computed belief values for the four studies.

Table 1 – Belief values for moderation and causal relationships of theoretical structures

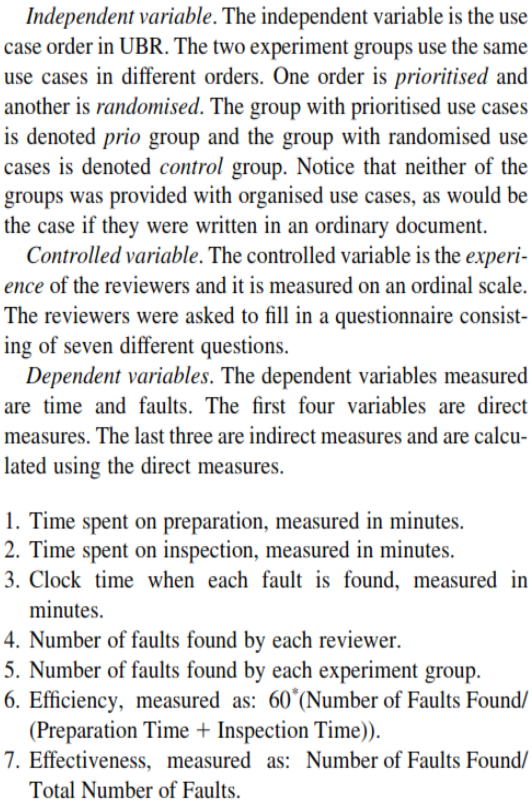
Study	Base belief value	Increase factor based on the study quality	Final belief value
S1	0.50	0.1858 (of 0.25)	0.6858
S2	0.50	0.2042 (of 0.25)	0.7042
S3	0.50	0.2042 (of 0.25)	0.7042
S4	0.50	0.1858 (of 0.25)	0.6858

It is possible to see that belief values are similar. It is a direct result of the fact that the first three papers have common authors. Thus, they tend to share the same textual structure when describing the procedures, analysis, and results. In the case of the fourth study, it is an external replication, which explains why the authors focused on reporting the same aspects to facilitate further comparison between the studies.

2.4 Extraction and Translation

All the experiments used the same set of instruments. Subjects inspected a real-world high-level design document, which consisted of an overview of the software modules and communication signals that are sent to/received from the modules. The system application domain is related to taxi management, and the design document specifies the three modules that compose the system: one taxi module used in the vehicles, one central module for the operators, and one integration module acting as a communication link between them. All faults were classified into three classes depending on the fault importance from the user's point-of-view. Class A or crucial faults represent faults in system functions that are crucial for a user (i.e., functions that are important for users and that are often used). Class B or important faults represent those which affect important functions for users (i.e., functions that are either important and rarely used or not as important but often used). Class C or minor faults are those that do not prevent the system from continuing to operate. Besides the number of faults, the experiments also report the efficiency (faults/hour) and effectiveness (faults/total faults).

Information extraction was largely facilitated by the quantitative nature of the studies. Each paper enumerated dependent and independent variables (Figure 1) so that it was straightforward to identify theoretical structures concepts.



Independent variable. The independent variable is the use case order in UBR. The two experiment groups use the same use cases in different orders. One order is *prioritised* and another is *randomised*. The group with prioritised use cases is denoted *prio* group and the group with randomised use cases is denoted *control* group. Notice that neither of the groups was provided with organised use cases, as would be the case if they were written in an ordinary document.

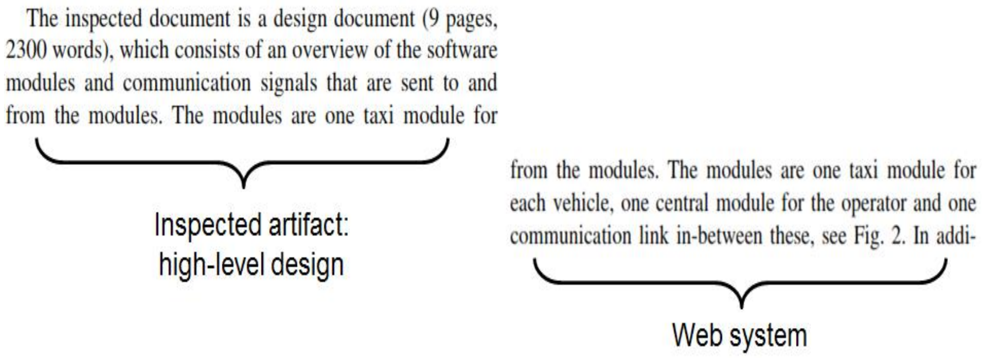
Controlled variable. The controlled variable is the *experience* of the reviewers and it is measured on an ordinal scale. The reviewers were asked to fill in a questionnaire consisting of seven different questions.

Dependent variables. The dependent variables measured are time and faults. The first four variables are direct measures. The last three are indirect measures and are calculated using the direct measures.

1. Time spent on preparation, measured in minutes.
2. Time spent on inspection, measured in minutes.
3. Clock time when each fault is found, measured in minutes.
4. Number of faults found by each reviewer.
5. Number of faults found by each experiment group.
6. Efficiency, measured as: 60*(Number of Faults Found/(Preparation Time + Inspection Time)).
7. Effectiveness, measured as: Number of Faults Found/Total Number of Faults.

Figure 1 – Study S1 variables listing

The context of experiments was detailed enough, which in controlled studies tend to be simpler than observational studies (Figure 2). Moreover, translation procedures were mostly unnecessary since studies' design was similar and used the same set of variables as surrogates. Causal relationships were extracted from the statistical tests used to answering the research questions. It is important to say that extraction and translation are solely based on what is reported. Thus, even though we knew important variables regarding the object of study at hand, theoretical structures should only have what is in the papers' text. For instance, we are aware that several, if not most, studies on software inspection consider the inspector's experience as a variable. Still, we could not include this variable into the theoretical structures, as the four studies did not observe this aspect.



The inspected document is a design document (9 pages, 2300 words), which consists of an overview of the software modules and communication signals that are sent to and from the modules. The modules are one taxi module for

from the modules. The modules are one taxi module for each vehicle, one central module for the operator and one communication link in-between these, see Fig. 2. In addi-

Figure 2 – Examples of concept identification for theoretical structure modeling

Given the similarity between studies, the theoretical structures for the four studies share most of the same concepts and relationships. Figure 3 depicts the theoretical structure modeled for the study S1 based on the information extracted. The only difference between theoretical structures from the four studies is related to the dependent variables. Two papers do not consider minor defects (class C) in their analysis. The authors do not provide any explicit justification for that, but we conjecture that it can be associated with publication space restrictions. Table 2 enumerates all effects along with its intensity and belief value (already adjusted with the discount from the *p-value*).

Table 2 – Effects reported in UBR primary studies

Study Effect	Effects showed as intensity (belief value)			
	S1	**S2**	**S3**	**S4**
Efficiency (total faults)	{SP} (0.66)	{SP} (0.67)	{WP, PO} (0.68)	{PO} (0.65)
Efficiency (crucial faults)	{PO, SP} (0.69)	{PO, SP} (0.70)	{WP, PO} (0.70)	{WP, PO} (0.68)
Efficiency (important faults)	{PO} (0.68)	{WP} (0.60)	{WP} (0.70)	{IF, WP} (0.69)
Efficiency (minor faults)		{WP} (0.52)	{WP} (0.70)	
Effectiveness (total faults)	{WP, PO} (0.64)	{PO} (0.63)	{PO} (0.70)	{SP} (0.67)
Effectiveness (crucial faults)	{PO, SP} (0.68)	{PO, SP} (0.68)	{PO, SP} (0.70)	{SP} (0.69)
Effectiveness (important faults)	{PO} (0.68)	{WP, PO} (0.58)	{PO} (0.70)	{IF, WP} (0.69)
Effectiveness (minor faults)		{IF, WP} (0.60)	{WP} (0.70)	
# Total faults	{SP} (0.69)	{PO} (0.63)	{PO} (0.70)	{SP} (0.67)
# Crucial faults	{PO, SP} (0.69)	{PO, SP} (0.68)	{PO, SP} (0.70)	{SP} (0.69)
# Important faults	{WP, PO} (0.69)	{WP, PO} (0.58)	{PO} (0.70)	{IF, WP} (0.69)
# Minor faults	{WP} (0.69)	{IF, WP} (0.60)	{WP} (0.70)	

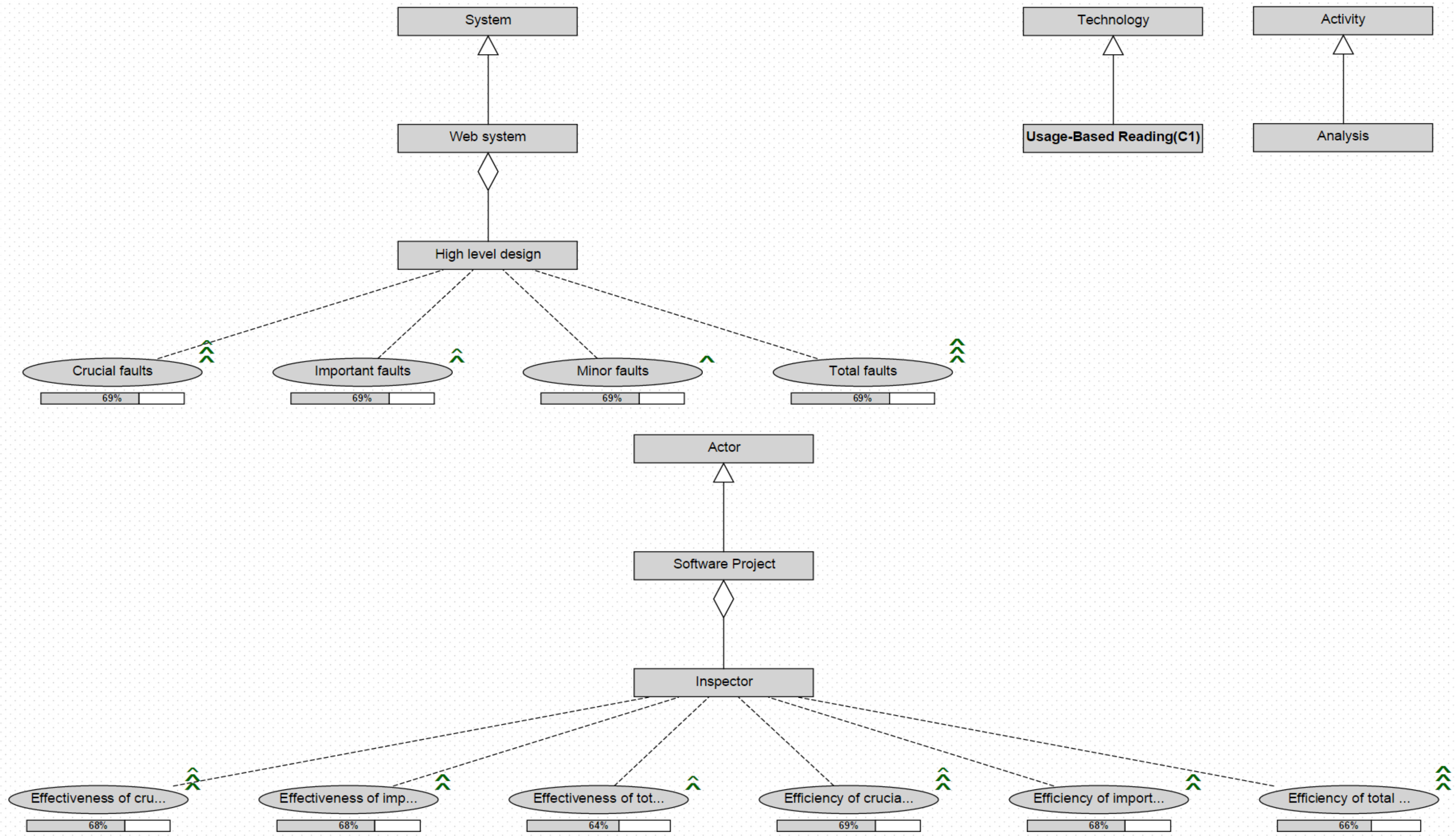


Figure 3 – Evidence model representing study S1 results (Thelin et al. 2001)

It is important to notice at this point that, although we are focusing on the descriptive, theoretical structures for UBR, they were modeled using the dismembering operation (). It means that first, we modeled comparative theoretical structures (comparing UBR with ad-hoc or CBR) and, then, based on the differences of the comparative cause-effect relationships, we determined the intensity of effects for UBR. We choose this strategy, instead of extracting two descriptive, theoretical structures from comparative studies as recommended in SSM, because papers contained percentage difference in most cases. Still, when individual data about each technology was present, we used it to calibrate the dismembering operation – that is, making it more precise than defined in Table 8 (Appendix A). Even indirect data, such as graphical data and boxplots, were used to that end. In Table 3, we list the effects for study S1 detailing how they were dismembered.

Table 3 – Dismembering operation values for study S1

Effect	Comparative	Descriptive for ad-hoc	Descriptive for UBR
Efficiency (total faults)	{WS}	{PO}	{SP}
Efficiency (crucial faults)	{SU}	{WP}	{PO, SP}
Efficiency (important faults)	{WS}	{WP}	{PO}
Effectiveness (total faults)	{WS}	{WP}	{WP, PO}
Effectiveness (crucial faults)	{SU}	{WP}	{PO, SP}
Effectiveness (important faults)	{WS}	{WP}	{PO}
# Total faults	{WS}	{PO}	{SP}
# Crucial faults	{SU}	{WP}	{PO, SP}
# Important faults	{WS}	{WP}	{WP, PO}
# Minor faults	{WS}	{WP, PO}	{WP}

The conversion rules used for comparative and descriptive values (when available) are enumerated in Table 4. We defined both comparative and descriptive rules because in some cases descriptive values were available. However, this led us to some inconveniences as the rules could conflict. For instance, in the case of ‘efficiency (crucial faults),’ the percentage difference between the inspection techniques is 95% and the mean values of identified faults per hour are 1.293 and 2.533 for ad-hoc and UBR, respectively. Therefore, if only the percentage difference was considered, then the descriptive values obtained from dismembering operation should have two units of distance (*e.g.*, WP and SP) since the 95% percentage difference is converted to {SU}. On the other hand, the approximate values of 1.293 and 2.533 are converted to {WP} and {PO} according to the defined rules, which has only one unit of difference between them. In these conflicting cases, to make the comparative and descriptive conversion rules compatible, we reduced the precision of the converted values. As a result, in this same example, the comparative value {SU} was dismembered to {WP} and {PO, SP} instead of {WP} and {PO}.

Table 4 – Conversion rules for effects quantitative values

Effect		Comparative qualitative intensity/difference	Quantative rule range
Comparative	Efficiency Effectiveness # defects	Indifferent (IF)	[0%, 0%]
		Weak difference (WI or WS)	(0%, 50%]
		Moderate difference (IN or SU)	(50%, 100%]
		Strong difference (FS or FI)	–[1]
Descriptive	Efficiency	Indifferent (IF)	0
		Weak impact (WN or WP)	(0, 2.5]
		Moderate impact (NE or PO)	(2.5, 5]
		Strong impact (SN or SP)	(5, ∞]
	Effectiveness	Indifferent (IF)	0
		Weak impact (WN or WP)	(0, 0.33]
		Moderate impact (NE or PO)	(0.33, 0.66]
		Strong impact (SN or SP)	(0.66, 1]
	# defects	Indifferent (IF)	0
		Weak impact (WN or WP)	(0, 4]
		Moderate impact (NE or PO)	(4, 8]
		Strong impact (SN or SP)	(8, 12]

2.5 Aggregation and analysis

To answer the research question defined for this worked example, only the dismembered theoretical structures relative to UBR were analyzed. Given their similarity, we were not able to identify any incompatibility between them. Thus, all four studies were analyzed together in a single aggregation. Some studies did not analyze (or report) some variables related to minor faults, but this is not impeditive for the aggregation since in SSM each effect is individually aggregated considering the papers in which they are present.

After this compatibility analysis and given the confidence level of each effect, Dempster's rule of combination could be computed. The combined theoretical structure is shown in Figure 4, and the detailed aggregation results are listed in Table 5. The first column shows the reported effect (i.e., benefit or drawback). The second column indicates the number of papers that have reported this effect. The third column shows the aggregated UBR effects intensity. The fourth column represents the aggregated belief on the respective effect. The fifth column lists conflict levels computed in each combination for the respective effect. For instance, the aggregation of four pieces of evidence leads to three combinations. Conflicts are always shown in the same order ((S1 $\oplus$ S3) $\oplus$ S4) $\oplus$ S2. This order was applied by the Evidence Factory tool, based on the order of the random IDs assigned to the evidence models. The sixth column registers the difference between maximum belief value of individual evidence for the respective effect and the aggregated value. The effects that were most strengthened where effectiveness and number of crucial faults.

[1] As we observed that the compared technologies are always able to identify defects (positive effects), we decided not to use strong difference.

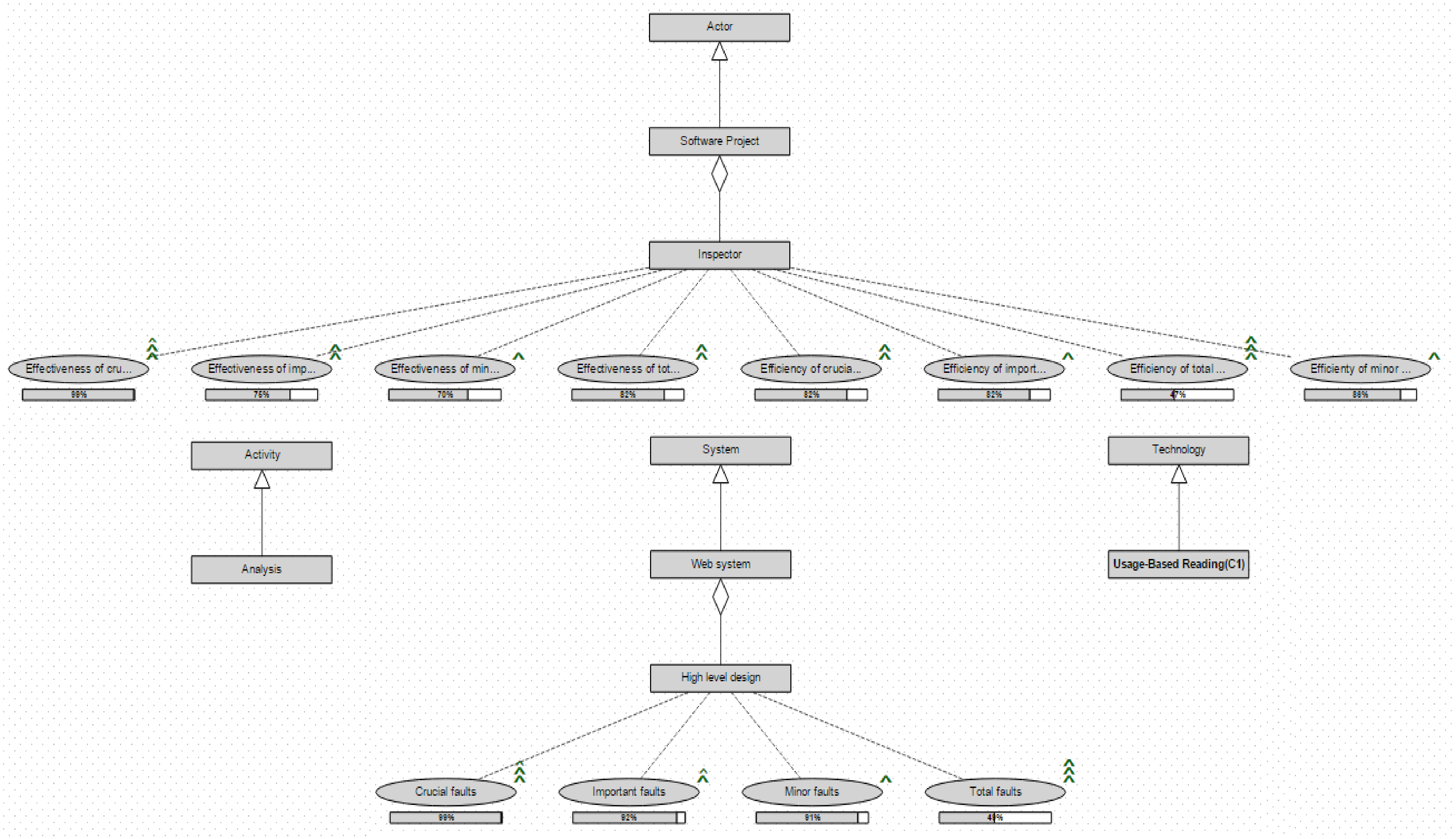


Figure 4 – Aggregated theoretical structure for UBR synthesis

Table 5 – Aggregated effects of UBR

Effect	Aggregation Results				
	#Papers	***Intensity***	***Belief***	***Conflicts***	***Difference***
Efficiency (total faults)	4	{SP}	0.47	0.45, 0.25, 0.49	-0.21
Efficiency (crucial faults)	4	{PO}	0.82	0.00, 0.00, 0.00	0.12
Efficiency (important faults)	4	{WP}	0.82	0.48, 0.27, 0.10	0.12
Efficiency (minor faults)	2	{WP}	0.86	0.00	0.16
Effectiveness (total faults)	4	{PO}	0.82	0.00, 0.60, 0.12	0.12
Effectiveness (crucial faults)	4	{PO, SP}	0.99	0.00, 0.00, 0.00	0.29
Effectiveness (important faults)	4	{PO}	0.75	0.00, 0.64, 0.00	0.05
Effectiveness (minor faults)	2	{WP}	0.70	0.00	0.00
# Total faults	4	{SP}	0,49	0.48, 0.28, 0.46	-0.21
# Crucial faults	4	{PO, SP}	0.99	0.00, 0.00, 0.00	0.29
# Important faults	4	{WP, PO}	0,93	0.00, 0.48, 0.00	0.23
# Minor faults	3	{WP}	0,91	0.00, 0.00	0.21

Before analyzing the aggregated results, we should first define how conflicts should be resolved. Although we have not had any incompatibility between theoretical structures, we can notice major conflicts between study results. There are three main factors associated with these conflicts. The first comes from the fact that we dismembered results from comparisons between UBR and ad hoc, and UBR and CBR. Therefore, it is expected some differences among results. The second aspect is related to the dismembering operation itself. As defined in SSM, dismembering is imprecise and suggested to be used only in some specific situations. Thus, it is a potential source of differences between results as well. The last aspect considered for explaining results is that the second combination (between S4 and resulting aggregation from S1 and S3) has the highest frequency of conflict occurrence – half effects had conflicts in the second combination. Interestingly enough, it is the combination involving the study S4, which is the only study that is an external experiment of UBR.

The combined belief values presented in Table 5 were computed using the basic conflict resolution strategy of SSM, which ignores the conflict by redistributing it among hypotheses. However, to use this strategy SSM establishes that all conflicts must be lower than 0.50 or the mean conflict is below 0.33 (as in this particular case of 3 combinations we have 1/3 = 0.3333). Hence, we understood that the best strategy to handle conflicts in this aggregation was incorporation. In other words, by using the dismembering function and aggregating results for a comparison of different techniques, we are much more interested in the trend than the specific result within the *Likert* scale. It is directly related to the incorporation conflict strategy, which tends to produce relatively more imprecise results. Next, in Table 6 the new belief values, after conflicts resolution, are presented.

Table 6 – Aggregated effects of UBR after conflicts resolution by incorporation

Effect	**Aggregation Results**				
	#Papers	***Intensity***	***Belief***	***Conflicts***	***Difference***
Efficiency (total faults)	4	{PO, SP}	0.85	*(INCORPORATED)*	0.17
Efficiency (crucial faults)	4	{PO}	0.82	0.00, 0.00, 0.00	0.12
Efficiency (important faults)	4	{WP}	0.82	0.48, 0.27, 0.10	0.12
Efficiency (minor faults)	2	{WP}	0.86	0.00	0.16
Effectiveness (total faults)	4	{PO, SP}	0.87	*(INCORPORATED)*	0.17
Effectiveness (crucial faults)	4	{PO, SP}	0.99	0.00, 0.00, 0.00	0.29
Effectiveness (important faults)	4	{WP, PO}	0.77	*(INCORPORATED)*	0.07
Effectiveness (minor faults)	2	{WP}	0.70	0.00	0.00
# Total faults	4	{PO, SP}	0.99	*(INCORPORATED)*	0.29
# Crucial faults	4	{PO, SP}	0.99	0.00, 0.00, 0.00	0.29
# Important faults	4	{WP, PO}	0.93	0.00, 0.48, 0.00	0.23
# Minor faults	3	{WP}	0.91	0.00, 0.00	0.21

We also present details of one conflicting aggregation to illustrate the conflict incorporation procedure (Table 7). As previously defined in SSM, instead of redistributing the conflict among all hypotheses, the idea of incorporation is to stretch the range of effect intensity by putting the conflict value into a contiguous range that includes the conflicting pair of hypotheses sets. For instance, in the first combination of Table 7 (between studies S1 and S3), there is a conflict value of 0.455 between the hypotheses {SP} from study S1 and {WP, PO} from study S3 which is assigned to the hypothesis {WP, PO, SP}. Thus, in this case, we have the positive trend for the effect that includes all positive values of the *Likert* scale ({WP, PO, SP}) and not a precise intensity ({WP}, {PO} or {SP}) for it. The same operation is performed in the other conflicts. After the three combinations, the results of aggregation are presented at the bottom of Table 7. The hypothesis {PO, SP} was chosen based on the criterion defined in SSM. Although $\text{Bel}_{1,2,3,4}$({WP, PO, SP}) has the largest value of 0.987, the $Bel_{1,2,3,4}$({PO, SP}) contributes with more than 50% of its value, since 0.854/0.987 = 0.865. As a result, it was not selected. Furthermore, in the case of $Bel_{1,2,3,4}$({PO, SP}) the value of $Bel_{1,2,3,4}$({PO}) contributes with 19% (0.166/0.854 = 0.194) and $Bel_{1,2,3,4}$({SP}) = 0.297 with 35% (0.297/0.854 = 0.348). Since both $Bel_{1,2,3,4}$({PO}) and $Bel_{1,2,3,4}$({SP}) values contribute with less than 75% to $Bel_{1,2,3,4}$({PO, SP}), then the hypothesis {PO, SP} was instead.

Table 7 – Details of calculations for combining results of ‘efficiency (total faults)’ effect (conflicts are resolved by incorporation)

Combination of studies S1 and S3

m_3 m_1	**{WP,PO} (0.678)**	**Θ (0.322)**
{SP} (0.656)	Ø (0.445)	{SP} (0.211)
Θ (0.344)	{WP,PO} (0.233)	Θ (0.111)

Combination of study S4 with the resulting combination of studies S1 and S3

m_4 $m_{1,3}$	**{PO} (0.649)**	**Θ (0.351)**
{SP} (0.211)	Ø (0.137)	{SP} (0.074)
{WP,PO} (0.233)	{PO} (0.151)	{WP,PO} (0.082)
{WP,PO,SP} (0.445)	{PO} (0.289)	{WP,PO,SP} (0.156)
Θ (0.111)	{PO} (0.072)	Θ (0.039)

Combination of study S2 with the resulting combination of studies S1, S3, and S4

m_2 $m_{1,3,4}$	**{SP} (0.675)**	**Θ (0.325)**
{SP} (0.074)	{SP} (0.050)	{SP} (0.024)
{PO, SP} (0.137)	{SP} (0.092)	{PO,SP} (0.045)
{PO} (0.512)	Ø (0.346)	{PO} (0.166)
{WP,PO} (0.082)	Ø (0.055)	{WP,PO} (0.027)
{WP,PO,SP} (0.156)	{SP} (0.105)	{WP,PO,SP} (0.051)
Θ (0.039)	{SP} (0.026)	Θ (0.013)

Final combined probabilities and belief values

$\boldsymbol{m_{1,2,3,4}}$**({SP})** = 0.050 + 0.092 + 0.105 + 0.026 + 0.024 = 0.297	$\boldsymbol{Bel_{1,2,3,4}}$**({SP})** = 0.297
$m_{1,2,3,4}$({PO, SP}) = 0.346 + 0.045 = 0.391	$\boldsymbol{Bel_{1,2,3,4}}$**({PO, SP})** = 0.166 + 0.297 + 0.391 = 0.854
$\boldsymbol{m_{1,2,3,4}}$**({PO})** = 0.166	$\boldsymbol{Bel_{1,2,3,4}}$**({PO})** = 0.166
$\boldsymbol{m_{1,2,3,4}}$**({WP,PO})** = 0.027	$\boldsymbol{Bel_{1,2,3,4}}$**({WP,PO})** = 0.166 + 0.027 = 0.193
$\boldsymbol{m_{1,2,3,4}}$**({WP,PO,SP})** = 0.055 + 0.051 = 0.106	$\boldsymbol{Bel_{1,2,3,4}}$**({WP,PO,SP})** = 0.166 + 0.297 + 0.391 + 0.027 + 0.106 = 0.987
$\boldsymbol{m_{1,2,3,4}}$**(Θ)** = 0.013	$\boldsymbol{Bel_{1,2,3,4}}$**(Θ)** = 1

Result: {PO, SP} since $\boldsymbol{Bel_{1,2,3,4}}$**({PO, SP}) = 0.854**

At this point, with conflicts discussed and resolved, we focus on the results themselves. It is noticeable the large agreement between studies regarding results associated with crucial faults. It is manifested in the high belief value of 0.99 observed in efficiency, effectiveness, and number of crucial faults. The high belief values resulting from aggregation should be analyzed in perspective, as each aggregation has its specificities. In this case, the 0.99 belief value should not be necessarily interpreted as an ‘almost certainty’ (*i.e.*, belief value of 1), but rather as a virtually full agreement among four strong evidence (*i.e., quasi*-experiments). Thus, in other words, the current body of knowledge indicates that UBR seems to have a direct impact to crucial faults since it is possible to observe similar results in four different studies which even compare different technologies (ad hoc and CBR).

Another interesting finding that can be observed in the aggregated results is the relative difference between the intensity of effects associated with crucial and minor faults. The results suggest that UBR has a larger impact on crucial faults than minor faults. It is precisely the most important aspect of UBR as it focuses inspections on the most important type of faults. It was observed in all dimensions explored in the studies: efficiency, effectiveness, and the number of faults. UBR has a {PO} impact over efficiency relative to crucial faults while it has {WP} for efficiency relative to minor faults. For effectiveness, we found {PO, SP} for crucial faults and {WP} for minor faults. It was the same for the number of crucial faults. Thus, this consistency in the difference between crucial and minor faults among the studies is another important result strengthened in the aggregation.

Based on this analysis and the overall results detailed in Table 6, we have enough input to answer the research question defined for this synthesis. UBR inspection technique can safely be used for identifying most important (*i.e.*, crucial) faults in high-level design, with a high level of efficiency and effectiveness. It still can be used for less important faults, although with relatively less efficacy. These effects seem to result from the basic mechanism behind UBR, which is the assumption that the proper prioritization of use cases can help identifying relatively more important faults.

The scope in which the aggregation findings can be claimed to be valid are explicit in the aggregated theoretical structure (Figure 4). In all studies, the same Web system's high-level design models were inspected using UBR. Thus, it is difficult to argue with any generalization beyond this context. Still, the cause of the observed effects is theoretically reproducible in other contexts with different kinds of systems and software artifacts, since UBR working mechanism is based on use case prioritization, which is, at least theoretically, independent of the inspected software artifacts. Moreover, the studies did not explicitly consider the participation of graduate students as an important factor influencing the findings. Arguably, this is because most subjects have experience in SE industry. Following this line of reasoning, we understand that industry professionals can be included within the findings external validity. That is why we used the concept 'Inspector' to refer generically to the actor.

Besides external validity, we should extend our considerations to other types of validity threads. We believe that the most important internal validity threat is the potential bias associated with the fact that the same researcher that authored SSM conducted the synthesis. Thus, from the studies selection to the definition of concepts and their relationships, practically all steps were subjected to this issue. It was the main motivation for choosing an inspection technique as the theme for research synthesis so that the domain aspects would not represent a confounding factor during the synthesis process. Regarding construct validity, we should point the use of the dismembering operation, which represents a validity threat in itself as it increases the imprecision of effects intensity. To minimize this lack of accuracy, when apart from the percentage difference the absolute quantitative values were available they were used to improve the effects precision.

3. Conclusion

The goal of this paper is to provide a worked example of the SSM. As discussed in the introduction, the software inspection theme was deliberately chosen, as it is an acknowledged research topic within Software Engineering. We tried to present all details necessary to undertake a research synthesis using SSM. We hope this can serve as supplemental material for understanding and applying the method.

Furthermore, as all the four aggregated studies are quantitative, it was possible to see that SSM produces outcomes consistent with the input data. The synthesis strengthened the evidence regarding the effectiveness and efficiency of UBR regarding the crucial faults, which is exactly intended with the inspection technique. Still, researchers must be aware that the set of studies to synthesize greatly influences the consistency and reliability of the resulting synthesis. The synthesis of bad studies will inevitably lead to bad results. In this regard, SSM has relatively less and more transparent phases. The extraction and translation step is relatively less objective than the other ones since it depends on conceptual development. On the other hand, the aggregation and analysis step since they are carried out based on the theoretical structures formal representation.

Nevertheless, the results related to the UBR synthesis have their value on their own. Researchers interested in this theme can use this synthesis to guide future studies in this topic.

References

Fagan M (2002) A History of Software Inspections. In: Broy PDM, Denert PDE (eds) Software Pioneers. Springer Berlin Heidelberg, pp 562–573

Porter A, Votta L (1998) Comparing Detection Methods For Software Requirements Inspections: A Replication Using Professional Subjects. Empir Softw Eng 3:355–379. doi: 10.1023/A:1009776104355

Santos PSM, Nascimento IE, Travassos GH (2015) A Computational Infrastructure for Research Synthesis in Software Engineering. In: XVIII Ibero-American Conference on Software Engineering, Track: XVII Workshop on Experimental Software Engineering. Lima, Peru, pp 309–322

Santos PSM, Travassos GH (2013) On the Representation and Aggregation of Evidence in Software Engineering: A Theory and Belief-based Perspective. Electron Notes Theor Comput Sci 292:95–118. doi: 10.1016/j.entcs.2013.02.008

Thelin T, Andersson C, Runeson P, Dzamashvili-Fogelstrom N (2004) A replicated experiment of usage-based and checklist-based reading. In: 10th International Symposium on Software Metrics, 2004. Proceedings. IEEE, pp 246–256

Thelin T, Runeson P, Regnell B (2001) Usage-based reading—an experiment to guide reviewers with use cases. Inf Softw Technol 43:925–938. doi: 10.1016/S0950-5849(01)00201-4

Thelin T, Runeson P, Wohlin C (2003) An experimental comparison of usage-based and checklist-based reading. IEEE Trans Softw Eng 29:687–704. doi: 10.1109/TSE.2003.1223644

Winkler D, Halling M, Biffl S (2004) Investigating the effect of expert ranking of use cases for design inspection. In: Euromicro Conference, 2004. Proceedings. 30th. pp 362–371

Appendix A. Aggregation of comparative theoretical structures

This appendix details the strategy used for aggregating comparative evidence denominated here dismembering operation. The only important difference between descriptive and comparative theoretical structures is the way that *causal* relationships are described. In comparative theoretical structures, *effects* are defined relative to the two causes observed in evidence. Given this difference, it is necessary to set an analog scale describing the comparison. A seven-point *Likert* scale with the following values is defined: strongly inferior (SI), inferior (IN), weakly inferior (WI), indifferent (IF), weakly superior (WS), superior (SU), and strongly superior (SS). Also, we redefine the *frame of discernment*: Θ = {SI, IN, WI, IF, WS, SU, SS}.

In fact, besides the effects' scale distinction, all the described procedures for aggregation are still applicable to both descriptive and comparative evidence. The descriptive evidence is characterized by its focus on describing possible benefits and drawbacks of a single cause whereas comparative evidence tries to do that relatively to another cause under the same category (*e.g.*, two inspection techniques). Despite this difference, we understand that aggregation is still possible since we can find the notion of causality in both kinds of evidence.

We define two additional strategies to aggregate descriptive and comparative evidence together: (i) determining a comparative theoretical structure based on the comparison of two descriptive theoretical structures and, the reverse operation, (ii) dismembering a comparative theoretical structure into two descriptive ones. In both cases, the notion of compatible theoretical structure is maintained: it is only possible to compare theoretical structures having the same *value* and *variable concepts*. Hence, dismembering a comparative theoretical structure produce two compatible descriptive theoretical structures with the same *value* and *variable concepts*.

The comparison of two descriptive theoretical structures is performed in the following manner. Using the defined *Likert* scale as an approximation for an interval scale, the maximum distance between the seven-point scale extremes (i.e., {SN} and {SP}) is 6, and the minimum is 0 when the compared values are the same. Based on this, we define a conversion rule between the descriptive and comparative scales:

- **{SI} or {SS}** when the difference is equal or larger than 3 units (e.g., from {IF} and {SP} = 3 up to {SN} and {SP} = 6);
- **{IN} or {SU}** when the difference is equal to 2 units;
- **{WI} or {WS}** when the difference is equal to 1 unit;
- **{IF}** when there is no difference.

As a convention, we say 'superior' when the first compared cause is better than the second and 'inferior' otherwise. For instance, if there is one descriptive evidence for two inspection techniques with $m_{\text{1-t1-\#defects}}(\{\text{PO}\}) = 0.3$ and $m_{\text{1-t2-\#defects}}(\{\text{IF}\}) = 0.9$ then, as the difference between {PO} and {IF} is equal to 2 units, the conversion would result in $m_{\text{1-t1/t2-\#defects}}(\{\text{SU}\}) = 0.3$[2]. Another relevant aspect of the conversion rule is the definition of the belief value. As the comparative scale is defined from two evidence,

[2] Clarification about the notation used: 1-t2-#defects should be read as 'evidence 1 related to the technique t2 for the effect #defects and 1-t1/t2-#defects as 'evidence 1 comparing t1 and t2 in relation to #defects'.

we take the minimum belief value between the pair. Thus, in the above example we have min(0.3, 0.9) = 0.3.

The dismembering procedure, on the other hand, should only be considered when raw descriptive data is not available, but comparative data such as numeric difference or qualitative description about differences. Otherwise, even if the study reports a comparative evidence, whenever the raw descriptive data is available it should be used to model descriptive theoretical structures. Therefore, the dismembering procedure only provides a rough way to estimate the individual effects for each cause considered in the comparison. To that end, we have defined an 'inverse' conversion rule based on the comparison procedure. Dismembering precision depends on the difference of effects' intensity and direction.

Table 8 – Conversion rules from comparative to descriptive theoretical structures

	Comparison between cause 1 and 2	Comparative causes	Dismembered value for cause 1	Dismembered value for cause 2
Non-negative effects	1 > 2	{SS}	{SP}	{IF}
		{SU}	{PO,SP}	{IF,WP}
		{WS}	{WP,PO,SP}	{IF,WP,PO}
	1 = 2	{IF}	{IF,WP,PO,SP}	{IF,WP,PO,SP}
	1 < 2	{WI}	{IF,WP,PO}	{WP,PO,SP}
		{IN}	{IF,WP}	{PO,SP}
		{SI}	{IF}	{SP}
Non-positive effects	1 > 2	{SS}	{IF}	{SN}
		{SU}	{WN,IF}	{SN,NE}
		{WS}	{NE,WN,IF}	{SN,NE,WN}
	1 = 2	{IF}	{SN,NE,WN,IF}	{SN,NE,WN,IF}
	1 < 2	{WI}	{SN,NE,WN}	{NE,WN,IF}
		{IN}	{SN,NE}	{WN,IF}
		{SI}	{SN}	{IF}
One effect non-negative and the other non-positive	1 > 2	{SS}	{IF,WP,PO,SP}	{SN,NE,WN,IF}
		{SU}	{IF,WP,PO}	{NE,WN,IF}
		{WS}	{IF,WP}	{WN,IF}
	1 = 2	{IF}	{IF}	{IF}
	1 < 2	{WI}	{WN,IF}	{IF,WP}
		{IN}	{NE,WN,IF}	{IF,WP,PO}
		{SI}	{SN,NE,WN,IF}	{IF,WP,PO,SP}
Note: when the comparative value is an interval, we assume the worst case (more imprecise)		{WS, SU} (non-negative effects)	{LP,PO,FP}	{IF,WP,PO}

For instance, in the best case, when compared causes are both non-negative, and the first is strongly superior ({SS}) to the second, then the only possibility to meet these considerations when dismembering is that the first cause is strongly positive ({SP}) and the second indifferent ({IF}). It is because, by definition, strongly superior must have three units of difference and given that causes do not have a distinct direction we conclude that one is {SP} and the other is {IF}. It is described in the first line of Table 8. Following this reasoning, the worst case for dismembering is a situation where both causes are non-negative (or non-positive) but do not have a difference ({IF}). Given these considerations, there are four possible equally acceptable answers for dismembering in this case, since both dismembered descriptive causes could assume values {SN}, {NE}, {WN}, and {IF} representing a non-negative comparative indifference (described in the fourth line of Table 8) – or {IF}, {WP}, {PO}, and {SP} for a non-positive comparative indifference (outlined in the eleventh line of Table 8). In Table 8, we enumerate all possible combinations[3] for dismembering comparative theoretical structures.

We suggest using dismembering only when there is some indication about the effects direction. It occurs, for instance, when raw data about the comparison is not available in the published report, but there are charts such as boxplots or dispersion showing the direction. Alternatively, in qualitative cases, where authors report that both causes had positive (or negative) effects, but one was superior to another.

[3] Non-negative and non-positive cases can be converted to negative and positive cases by just removing {IF} from dismembered effects values. In this case, as the maximum difference between two descriptive values is 2 units (e.g., between {WP} and {SP}), the comparative values can not assume {SI} or {SS}.