

Do words make a difference?

An Empirical Study on the Impact of Taxonomies on the Classification of Requirements

Kim Lauenroth
adesso AG
Stockholmer Allee 20
44269 Dortmund, Germany
kim.lauenroth@adesso.de

Erik Kamsties
Dortmund University of
Applied Sciences and Arts
44227 Dortmund, Germany
erik.kamsties@fh-dortmund.de

Oliver Hehlert
adesso AG
Stockholmer Allee 20
44269 Dortmund, Germany
oliver.hehlert@adesso.de

Abstract—Requirements taxonomies help to classify and channel the requirements in a project. A very simple taxonomy is the distinction between functional and non-functional requirements. Furthermore, a taxonomy helps to decide if a statement is a requirement at all or just something else (e.g., 'information'). The quality of a taxonomy is important as we do not want to put a statement in the wrong category. In this paper, we argue that we need to take cognitive psychology into account in this task of requirements classification. Cognitive psychology focuses on the abilities and limitations of the human mind. We present a controlled experiment and a replication in which we compare three requirements taxonomies. The participants had to evaluate a set of requirements based on the given taxonomies. The results of these experiments show that there are differences between the taxonomies: Interestingly, the question whether a statement is identified as a requirement or not depends on the taxonomy. These experiments present initial results, we assume that these results are related to phenomena of cognitive psychology. We conclude that the wording should be carefully taken into account in the definition of the categories of a high quality requirements taxonomy.

I. INTRODUCTION

A great amount of information is collected during the requirements elicitation phase and also at later phases of software development. In RE practice, not every piece of information should be interpreted as a requirement, since introductory statements, motivations, rationales, background information, and irrelevant information also flood in. Thus, from a practical perspective, it is important to *clearly mark requirements as such*, because, for instance, a requirement demands an acceptance test case, but not background information or rationales.

This collected information needs to be organized, and *requirements taxonomies* support this activity. A requirements taxonomy defines different categories of requirements and allows the requirements engineer to say that a given requirement belongs to the category *X*, *Y*, or *Z*. These taxonomies are an important tool in requirements engineering practice and are used, e.g., as checklist in requirements elicitation, or to structure requirements templates or documents.

In an ideal world, the customer clearly indicates the requirements, for instance by a particular phrase, e.g. 'The system shall...'. In practice however, the requirements engineer is often faced with the task of identifying requirements. Thus, beside this obvious task of categorizing requirements, a requirements taxonomy fulfills an implicit task: it supports the requirements engineer in separating real requirements from other information. If a statement does not fit into one of the categories, it must be something else, and not a requirement.

In this paper, we focus on this second, implicit *binary classification* task of separating real requirements from other information, which is of practical relevance for the following reasons:

- 1) Overlooking important requirements for a new software systems is unanimously accepted in research and industry as one of the greatest threats to software project success. *A high-quality requirements taxonomy should support the requirements engineer in identifying real requirements.*
- 2) Other information that becomes part of the requirements specification increases the size of the specification and affects the overall process. *A high quality requirements taxonomy should support the requirements engineer in identifying other information that is not a requirement.*

From the cognitive viewpoint, requirements classification is a simple but demanding task. The requirements engineer has to examine a particular requirement against a given taxonomy to decide to which category a requirement belongs.

Cognitive psychology focuses on the abilities and limitations of the human mind. One interesting effect is called the *Invisible Gorilla*. It shows that our attention with respect to unexpected events and information is very limited, especially when we are working on a particular cognitive task. Requirements engineering work is full of cognitive tasks, e.g., listening to stakeholders during interviews, or reading a specification during review activities. We are therefore interested in the question, how requirements classification could be affected by phenomena from cognitive psychology.

The remainder of this paper is structured as follows. In Section II we discuss related work. Section III describes the investigated taxonomies in more detail. In Section IV, we will present findings from cognitive psychology and discuss their potential impact on the classification of requirements. We further present the experiment design in Section V and the results of this experiment as well as a replication in Section VI. Section VII discusses threats to the validity of our experiment. Section VIII closes the paper with a conclusion and a perspective towards future research.

II. RELATED WORK

The related work in the intersection between cognitive psychology and requirements engineering is to the best of our knowledge very small. The information systems engineering community is active in this direction with a dedicated workshop series: COGNISE - International Workshop on Cognitive Aspects of Information Systems Engineering. The survey of Lenberg et al. from 2015 [1] shows that there is substantial work at the intersection of software engineering and psychology, but very little work at the intersection of requirements engineering and psychology.

Mohanani [2] examines the impact of cognitive bias during the development conceptual designs from requirements. Object-oriented modelling has been examined from a psychological perspective, e.g., by Dawson [3], Hadar [4], and Tegarden/Sheetz [5]. Aranda applies findings from cognitive psychology to distributed requirements elicitation [6]. Moody applies results from cognitive psychology to the evaluation of visual notations (cf. [7]). In our own work, we have argued for the benefits of taking psychological effects into account in RE research [8].

RE research presents several requirements taxonomies (cf., e.g. [9] or [10] for an overview of several definitions for non-functional requirements), and there is a significant amount of research that focuses on the *automated* identification and classification of requirements (cf., e.g. [11], [12], or [13]).

However, to the best of our knowledge, there is up to now no work on the impact of psychological effects on the *manual* classification of requirements.

III. REQUIREMENTS TAXONOMIES

A wide range of taxonomies has been published, ranging from very simple ones to complex, domain-specific taxonomies with more than 50 categories [13].

We selected three simple taxonomies for our experiment. The first taxonomy is the classical definition from IEEE [14] (short T_{IEEE}) and is a real binary classification. The second and third are two common taxonomies from the literature: functional vs. non-function requirements (short T_{NFR}) [15] and functional vs. quality requirements vs. constraints (short T_{QR}) [16].

These taxonomies define *explicitly* different categories of requirements. As we can find other types of statements from a customer as well, usually another category "information" is added to provide a complete taxonomy which is capable of

classifying every statement. Yet, this category is *implicit* in the original description of the taxonomies.

A. Taxonomy T_{IEEE} :

IEEE610 [14] defines a requirement as ...

- 1) a condition or capability needed by a user to solve a problem or achieve an objective.
- 2) a condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed documents.
- 3) a documented representation of a condition or capability as in (1) or (2).

This taxonomy defines only one explicit category (requirement) and one implicit category (not a requirement). We take this taxonomy into account because it represents in our point of view one of the most fundamental tasks in requirements engineering: Recognizing if a given statement is a requirement or not.

B. Taxonomy T_{NFR} :

Ralph R. Young defines in his book [15]:

- **Functional Requirement:** Functional requirements describe what the system or software must do.
- **Non-Functional Requirement:** Non-functional requirements specify system properties, such as reliability and safety.

This taxonomy distinguished two categories of requirements and a third implicit category (neither functional nor non-functional requirement). The important keyword for both categories is the term *functional*.

The distinction between the two categories is created by the negation of the term *functional*. Such a definition is called *dualism* which in the essence means that a thing can be in one out of two given categories [17].

C. Taxonomy T_{QR} :

The IREB glossary defines [16]:

- **Functional Requirement:** A requirement concerning a result or behavior that shall be provided by a function of a system (or of a component or service).
- **Quality Requirement:** A requirement that pertains to a quality concern that is not covered by functional requirements.
- **Constraint:** A requirement that limits the solution space beyond what is necessary for meeting the given functional and quality requirements.

This taxonomy distinguished three categories of requirements and a fourth implicit category (neither functional requirement nor quality requirement nor constraint). The three categories defined by this taxonomy use three keywords: *functional*, *quality*, and *constraint*. This taxonomy does not incorporate a dualism as the T_{NFR} taxonomy.

These three taxonomies provide several categories but we reduce all of them to a *binary classification* as shown in Table I as we are interested only in the *identification* of requirements.

TABLE I
CLASSIFICATION RESULTS FOR EXEMPLARY TAXONOMIES

Taxonomy	Classification Result	Defined Categories
T_{IEEE}	Requirement	Requirement
	Information	Information (but not a requirement)
T_{NFR}	Requirement	Functional, non-functional requirement
	Information	Information (but not a requirement)
T_{QR}	Requirement	Functional, quality requirement, constraint
	Information	Information (but not a requirement)

We use the term *classification quality* to denote the degree to which the classification of statements based on a taxonomy meets an expert's classification (a more formal definition of the term is provided in Chapter V). The focus of this paper is on the *differences in classification quality of taxonomies* rather than the classification quality of the taxonomies themselves. Intuitively, it could be expected that the classification quality of the taxonomies is equal when reduced to a binary decision between *requirement* and *information*. Yet, findings from cognitive psychology discussed in the following section lead us to the assumption that there are differences.

IV. FINDINGS FROM COGNITIVE PSYCHOLOGY

This section presents findings from cognitive psychology and discusses our expectation about their impact on the classification of requirements.

Thinking in Working Memory. Cognitive tasks (e.g. solving a puzzle) are performed by the human in the so called working memory [18]. Similar to the working memory of a computer, the human working memory has a limited capacity. If the working memory is stressed during a particular task (e.g. because a person has to keep several information units in the working memory), the performance of the current task is decreased and the probability of mistakes is increased (cf., e.g., [19]).

Applying a taxonomy (e.g. classifying a statement as requirement or other information) is a task that is performed in the working memory. The statement under examination has to be evaluated against the elements of the taxonomy. To perform this task, the statement under examination together with the elements of the used requirements taxonomy have to be kept in the working memory. A taxonomy with several categories may stress the working memory more than a taxonomy with only a few categories. Such a stress may therefore lead to more classification mistakes and consequently may show a lower score in all above presented measures.

Priming - Impact of Words on Memory and Perception. Priming is a so called implicit memory effect in which a stimulus (e.g., reading or hearing a word) influence the response to another stimulus [20]. Daniel Kahnemann illustrates the effect of priming with a very simple example [21]:

If you have recently seen or heard the word EAT, you are temporarily more likely to complete the word fragment SO_P as SOUP than as SOAP. The opposite would happen, of course, if you had just seen WASH.

The priming effect is not limited to completing word fragments, it especially extends to the human perception. For

example, a person that is primed with the word EAT is more likely to recognize food in his surrounding (cf. [21]). The elements of a requirements taxonomy are defined by certain keywords (e.g. 'functional' in functional requirement). These keywords may trigger a priming effect that influences the memory or the perception of the person that is using the taxonomy.

This priming effect may be positive if the keywords of the taxonomy elements are well defined since the perception of requirements engineer will easily identify the relevant words that allow for proper classification. Such a proper definition of keywords should lead to higher scores in all above presented measures. On the contrary, the priming effect may be negative, if the keywords of the taxonomy elements are poorly defined. In this case, the requirements engineer may overlook important keywords that would allow for a proper classification. This effect would lead to lower scores in all above presented measures. A possible example of such a negative effective could be the dualism that created by the NFR taxonomy. Every statement that is not considered a functional requirement could be considered automatically as a non-functional requirement, even if it is not a requirement at all.

The Illusion of Attention or the Invisible Gorilla. Christopher Chabris and Daniel Simons [22] present an experiment in which the participants have to watch a video. The video shows two groups of people (one group wearing black, the other white shirts). Both groups pass basket balls. The task of the participants was to count the number of passed balls between the players wearing white shirts. After watching the video, the participants had to report the number they had counted.

The participants were further asked if they had noticed anything unusual during the counting task. Nearly half of the subjects said that they noticed nothing unusual. However, the video contained something unusual: a person wearing a gorilla costume walked into the scene, stopped in the middle of players, thumped its chest and walked away. This intermezzo lasted for about 9 seconds. Nearly half of the participants missed the gorilla¹.

This experiment shows that our attention with respect to unexpected events and information is very limited, especially when we are working on cognitive tasks. Therefore, Chabris and Simons call this effect the *Illusion of Attention* or the *Invisible Gorilla*. It is important to note that this effect is not

¹This experiment has become known as the monkey business illusion and is available on YouTube: https://www.youtube.com/watch?v=IGQmdoK_ZfY

limited to real-time stimuli such as balls and players, but apply to people's cognitive information processing in general [23].

As stated above, applying a taxonomy is a cognitive task and presumably more demanding than the pass counting exercise from the gorilla experiment. The requirements engineer has to carefully examine a given statement. Therefore, it is possible that people miss unexpected information in a stressful situation (e.g., during a workshop or an interview situation). This effect could cause requirements engineers oversee important words that would lead to the proper classification of a statement. Thereby he could miss an important requirement mentioned by a stakeholder or include a statement that is not a requirement at all.

The priming effect and the illusion of attention could even create a combined effect in requirements work with taxonomies. The words that define the taxonomy could cause a priming effect that leads to a certain expectation during requirements work. For example, people subconsciously expect certain keywords and oversee other unexpected keywords that would have resulted in a different categorization of a requirement.

V. EXPERIMENT DESIGN

A. Definition of the Hypothesis

The objective of our empirical study is to evaluate if there is a difference in the classification quality of different requirements taxonomies. The requirements classification task is performed by a person (typically a requirements engineer or a stakeholder) in a given project situation (e.g., in a project meeting or in his office) on a set of given statements based on a given taxonomy.

We want to examine the impact of the different taxonomies (our independent variable) on the classification quality (the dependent variable).

Following the discussion on findings from cognitive psychology (see Section IV), we state the following hypothesis: *There is a difference in the classification quality between the three investigated taxonomies (T_{IEEE} , T_{NFR} , and T_{QR}). The metrics to measure the classification quality are described in the following section.*

B. Definition of Metrics

We evaluate the results per participant based on the classification theory for binary classification (cf., e.g., [24]) as follows. A statement is classified by *participants* of an empirical study and *RE experts*. Thereby, we get an expert opinion which we consider as reference point and several participant results. This leads to four cases:

- **True Positive (TP):** A statement is a requirement and is recognized by the participant as requirement
- **False Positive (FP):** A statement is *not* a requirement but is recognized by the participant as requirement
- **False Negative (FN):** A statement is a requirement but is *not* recognized by the participant as requirement
- **True Negative (TN):** A statement is *not* a requirement and is *not* recognized by the participant as requirement

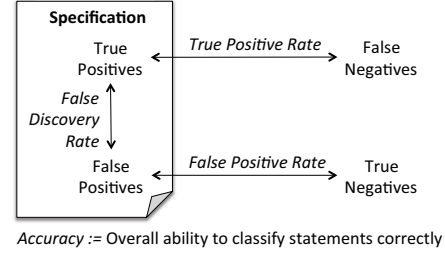


Fig. 1. Visualization of Metrics related to a Hypothetical Specification

Classification theory provides several measures for the quality of a given classification. To derive measures for the requirements classification quality of requirements taxonomies, we assume that the statements that a person has selected as requirement (the true and the false positives) become part of a hypothetical specification and select the following four measures for the operationalization of requirements classification quality of a person (see Fig. 1):

- **TPR - True Positive Rate (Recall):** $\frac{t_p}{t_p + f_n}$. The TPR can be interpreted as the coverage of the requirements specification compared to all theoretically possible requirements.
- **FPR - False Positive Rate (Fall-out):** $\frac{f_p}{t_n + f_p}$. The FPR can be interpreted as the percentage of other information that becomes part of the requirements specification compared to all theoretically possible other information. (Remark: $FPR = 1 - TNR$ (True Negative Ratio or specificity)).
- **FDR - False Discovery Rate (Fall-out):** $\frac{f_p}{t_p + f_p}$. The FDR can be interpreted as the percentage of other information that becomes part of the requirements specification compared to overall number of statements in the specification. (Remark: $FDR = 1 - PPV$ (Positive Predictive Value or precision)).
- **ACC - Accuracy:** $\frac{t_p + t_n}{t_p + t_n + f_p + f_n}$. With respect to the requirements specification, the ACC can be interpreted as the overall ability to recognize requirements and other information correctly.

C. Design

Every participant uses all three taxonomies to reduce the effect of intellectual abilities of single participants on the result. In order to control learning effects, we permute the order of the taxonomies, which results in $3! = 6$ groups (see Table II).

An experiment participant gets 3 sets with 20 statements each (10 statements are requirements and 10 are not). The sets are disjoint, i.e. every participant evaluates every statement exactly once and evaluates every set with a different requirements taxonomy.

D. Experiment Setting and Material

To test the above mentioned hypothesis, we have defined an experimental setting where a participant has to evaluate statements based on the three different requirements taxonomies in a quiet and stressless laboratory situation.

TABLE II
GROUPS AND ORDER OF TAXONOMIES

Group	1st Tax.	2nd Tax.	3rd Tax
1	T_{IEEE}	T_{NFR}	T_{QR}
2	T_{IEEE}	T_{QR}	T_{NFR}
3	T_{NFR}	T_{IEEE}	T_{QR}
4	T_{NFR}	T_{QR}	T_{IEEE}
5	T_{QR}	T_{IEEE}	T_{NFR}
6	T_{QR}	T_{NFR}	T_{IEEE}

We created a set of 60 statements (in *German*, 30 statements represent a requirement and 30 statements are other information). The statements and the classifications have been developed in an iterative way by three independent industrial experts and were derived from their project experience. They reflect typical statements in a requirements document (see examples below), but they are neither derived from the same specification nor from a single application domain. All statements have been classified independently by all three experts and in case of disagreement modified until all three experts agreed with the classification of each statement.

The following statements are examples for a requirement *The system shall offer a comfortable search function* and for other information *The system should not be too expensive*.

Every statement is presented on a single page of paper with the assigned taxonomy. The participant has to tick a box for his classification decision. This individual presentation of statements and highly simplified answering scheme minimizes the possibility of distraction for the participant and eases the classification task for the participants.

Each set starts with the presentation of a definition of the taxonomy that has to be applied to the set. Between each taxonomy, the participant has to solve a puzzle from classical intelligence tests as a relaxing exercise. The purpose of this exercise is to prevent boredom with the classification task. The sets are created in a randomized fashion to prevent from cheating.

The overall experiment session took 90 minutes including a brief introduction into the experiment, distribution of material, evaluation of statements. The participants were able to complete their task in the given time. We made no attempts to measure individual effort as we do not have a theory and a hypothesis about the impact of taxonomies on the time for classification.

VI. RESULTS FROM 2 EXPERIMENTS

We performed the described experiment two times²:

- 28 master's students from a master's lecture in requirements engineering from the University of Applied Sciences and Arts in Dortmund in 2014 (short: group 2014).
- 35 master's students from a master's lecture in requirements engineering from the University of Applied Sciences and Arts in Dortmund in 2015 (short: group 2015)

²Please contact the authors to obtain the raw data of the experiment (metrics per participant, R scripts for statistical analysis, and used statements).

All master's students came from the Computer Science department and they were familiar with all three taxonomies since the taxonomies were discussed in detail in their requirements engineering lecture.

In the remainder of this section, we will first give an overview of the results of both experiments. We further present our approach for the statistical analysis, present the results for each metric, and close with a short interpretation of the results.

A. Overview of Results

We achieved a total number of 1664 out of 1680 possible classifications for group 2014 and 2097 out of 2100 possible classifications for group 2015. The difference is explained by the fact that some statements have not been classified by the participants or were not interpretable because of multiple answers.

From the classification results, we calculated all metrics (TPR, FPR, FDR, and ACC) in total and per participant. The total results are presented in Table III and IV. The results per participants for both experiments are shown as boxplots in Figure 2. Min, max, and median values for both experiments can be found in Appendix IX in Table IX and X.

	TPR	FPR	FDR	ACC
IEEE	.8333	.2960	.2628	.7685
NFR	.8436	.3993	.3275	.7204
QR	.9278	.5326	.5326	.6980

TABLE III
TOTAL METRICS GROUP 2014

	TPR	FPR	FDR	ACC
IEEE	.9122	.3150	.2529	.7997
NFR	.9253	.4290	.3192	.7471
QR	.9424	.5954	.3899	.6719

TABLE IV
TOTAL METRICS GROUP 2015

Based on the results, we observed the following:

- The overall results indicate that the students were able to recognize requirements properly with a median TPR between .8397 (for T_{IEEE}) and 1.00 (T_{QR}) and a median accuracy between .65 (T_{QR}) and .85 (T_{IEEE}).
- The results of group 2014 appear to be weaker than the results of group 2015.
- Very few participants of the experiment show bad classification results. The lower quartile of the lowest accuracy value (for T_{QR}) is .59. The weakest accuracy values are .4211 (for T_{QR}) and .5263 (for T_{NFR}).
- The median values for all 4 metrics are close to the overall results for each metric (see Tables III and IV).
- The boxplots indicate a difference in the classification quality between all three taxonomies. The differences for the metric FPR appear to be very large between the 3 taxonomies. The differences for the other metrics are smaller, but visible.
- The direction of the differences is not the same for all 4 metrics. T_{QR} appears to be the best taxonomy for the

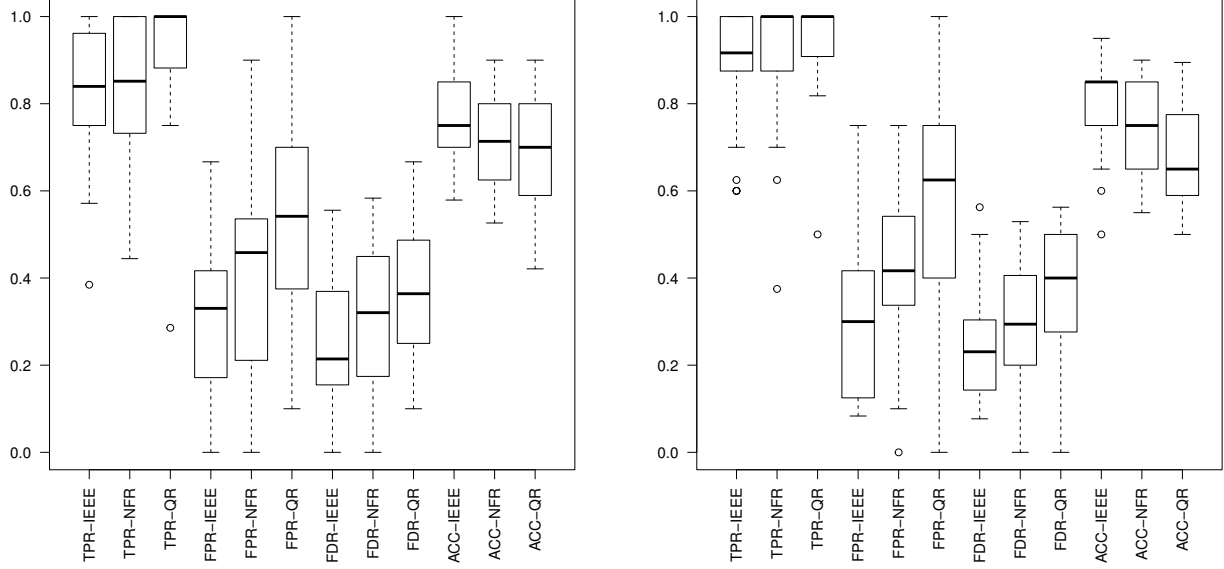


Fig. 2. Distribution of the 4 metrics - group 2014 left / group 2015 right

metric TPR and the worst taxonomy for the metrics FPR, FDR, and ACC.

B. Procedure for Statistical Analysis

For the statistical analysis of our results, we performed a four step approach.

1) Step 1 - Determine Statistically Significant Differences:

In the first step, we evaluate the statistical significance of the difference between the 3 taxonomies for each metric. We selected the one-way analysis of variance (ANOVA) test [25]. The ANOVA is a suitable statistical tool for our setting, since the ANOVA tests if the means of several groups are equal or not. In our setting, the groups are defined by the metric results of the different taxonomies.

In the following paragraphs, we report on the fulfilment of the prerequisites for the application of an ANOVA test (cf. [25] for the list of prerequisites):

- *Independence of observations:* All participants of our experiment classified the requirements on their own. We therefore consider the observations as independent from each other.
- *Normality:* We applied a Shapiro-Wilk normality test [26] to all groups per metrics and taxonomies. The result (W- and p-values) is shown in Appendix IX in Table XI and XII. The test results show that several metrics do not follow a normal distribution, since the p-value is below the threshold .05.
- *Homogeneity of variances:* We applied a Levene test [27] to all groups (metrics) to determine the homogeneity of variance. The results (F- and p-values) are shown in Appendix IX in Table XIII and XIV. The p-value for all metrics is above the threshold .05. We therefore conclude that the variances of all metrics are equal.

Although several metrics do not fulfil the prerequisite of normality distribution for an ANOVA, we have decided to apply the ANOVA instead of a Kruskal-Wallis test for two reasons. Firstly, the literature on the ANOVA argues that the ANOVA test is robust against a violation of the normality assumption if the sample sizes are equal. Samples size are equal because we compare the effects in the same group of people. Secondly, we consider the requirements classification task as a kind of intelligence test and the results of intelligence tests typically follow a normal distribution [28].

2) *Step 2 - Determine Effect Size:* In the second step, we calculated the effect size for each metric and group. In addition to the statistical significance of a difference, the effect size is an important measure to determine whether the observed effect is small or large. For an ANOVA, we followed the suggestions from [29] and selected eta-squared (η^2) as measure for the effect size. According to [30], .02 is a small, .13 is a medium, and .26 is a large effect.

3) *Step 3 - Determine Size of Difference:* In the third step, we evaluated the actual size of the difference between the taxonomies for each metric. We followed the suggestions from [25] and selected the Tukey's HSD (honest significant difference) test (short TukeyHSD) to determine the pairwise differences between the individual metrics (lower and upper bound of the difference and mean difference), and to calculate confidence intervals for each difference. The prerequisites for the TukeyHSD test are identical to the prerequisites of the ANOVA (see above). Therefore, the TukeyHSD test is applicable to our results.

4) *Step 4 - Combine Results of both Experiments:* In the final step, we combined the results of the TukeyHSD test for both experiments in a forest plot to compare the differences of both groups per taxonomy and to get a combined view on the results of both experiments. The combined view is calculated

from the mean of all results (lower and upper bound of the difference and mean difference).

In the following, we will present the combined results for both experiments per metric.

C. Metric TPR - True Positive Rate

Table V shows the results for the ANOVA tests for the metric TPR. The results for both groups show not acceptable p-values (above .05). However, the p-value for group 2014 is only a little above .05. The effect size for both groups are rather small (close to .02).

	F-value	p-value	eta-squared
Stud (2014)	3.0205	0.0543	0.0694
Stud (2015)	0.9016	0.4091	0.0174

TABLE V
ANOVA RESULT FOR METRIC TPR

The results of the TukeyHSD test (see Fig. 3) support the results of the ANOVA. The p-value for all pairs of metrics are not acceptable (above .05). There is no real indication for a difference between the taxonomies for the metric TPR.

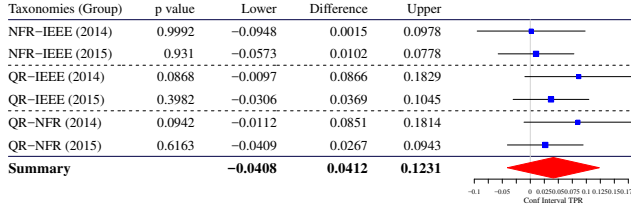


Fig. 3. Forest plot for TPR

D. Metric FPR - False Positive Rate

Table VI shows the results for the ANOVA tests for the metric FPR. The results for both group show very small p-values (close or below .001). The effect size for both groups is medium (above .13) to large (close to .26).

	F-value	p-value	eta-squared
Stud (2014)	9.2237	0.0002	0.1855
Stud (2015)	15.7208	0.0000	0.2356

TABLE VI
ANOVA RESULT FOR METRIC FPR

The results of the TukeyHSD test (see Fig. 4) show that there is a significant difference between the taxonomies T_{QR} and T_{IEEE} in both groups (p-value close to 0). For both groups, there also is a significant difference between T_{NFR} and T_{QR} . For group 2015 the difference between T_{IEEE} and T_{QR} and T_{NFR} is also significant (p-value below .05). For each pair of metrics, the forest plot also shows that the means (squares on the lines) for both groups are very close to each other. The combined results (diamond) also indicates that there is a real difference for the metric FPR.

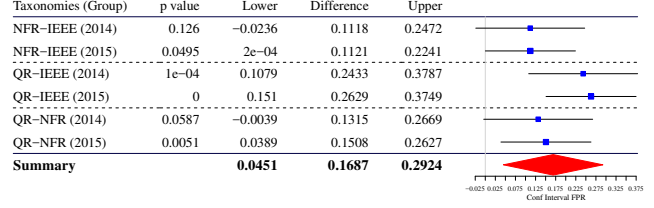


Fig. 4. Forest plot for FPR

E. Metric FDR - False Discovery Rate

Table VII shows the results for the ANOVA tests for the metric FDR. The results for group 2014 show an acceptable p-value (below .05) for group 2014 and a strong p-value for group 2015 (below .001). The effect size is between medium and small for group 2014 and is medium (above .13) for group 2015.

	F-value	p-value	eta-squared
Stud (2014)	3.4107	0.0378	0.0777
Stud (2015)	8.9882	0.0003	0.1498

TABLE VII
ANOVA RESULT FOR METRIC FDR

The results of the TukeyHSD test (see Fig. 5) shows that there is a real difference between the taxonomies T_{QR} and T_{IEEE} in both groups with strong p-values (close to 0). The p-values for the other pairs of metrics are not acceptable (above .05).

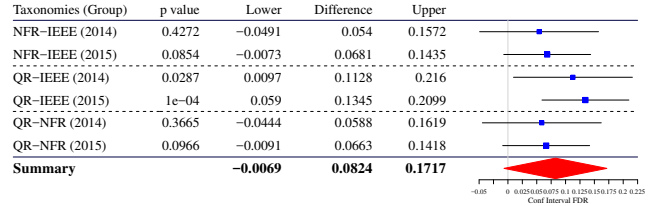


Fig. 5. Forest plot for FDR

F. Metric ACC - Accuracy

Table VIII shows the results for the ANOVA tests for the metric ACC. The results for group 2014 show a weak p-value (above .05) for group 2014 and very strong p-value for 2015 (close to 0). The effect size is between medium and small for group 2014 and is between medium and large for group 2015.

	F-value	p-value	eta-squared
Stud (2014)	2.5798	0.0820	0.0599
Stud (2015)	12.2559	0.0000	0.1938

TABLE VIII
ANOVA RESULT FOR METRIC ACC

The results of the TukeyHSD test (see Fig. 6) support these indifferent results. For group 2015, there is a significant difference between T_{QR} and T_{IEEE} (p-value close to 0) and between T_{QR} and T_{NFR} (p-value close to .01). The p-values for the other pairs of metrics are not acceptable (above .05).

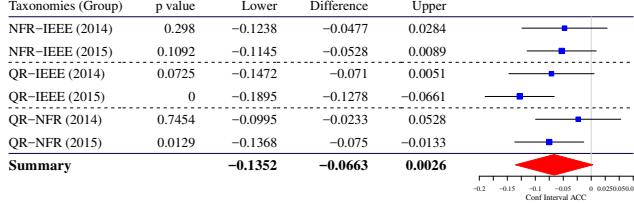


Fig. 6. Forest plot for ACC

G. Interpretation of Results

The results of our experience partially confirm our hypothesis. With respect to TPR, all three taxonomies appear to be close to each other with respect to classification quality. Our results did not show any statistically significant result for this metric. If we look at the absolute numbers (see Figure 2 and Tables IX and X), the TPR is between 0.84 and 1.0 (median). That is, there is not much room for significant differences. At least 84% of the requirements are recognized as such. The T_{QR} taxonomy performed best (median), also the variance in the individual results of the participants is considerably smaller.

For the metric FPR, the results provide statistically significant evidence that the selection of the taxonomy has an impact on the number of false positives. The taxonomy T_{IEEE} leads the lowest number of false positives, whereas the taxonomy T_{QR} leads the highest number of false positives. We could also show that there is a significant difference in the false positive rates of T_{NFR} compared to T_{IEEE} for the 2015 group. The dualism of T_{NFR} could be an explanation: if a statement was considered as "not a functional requirement" it was too often classified as a non-functional requirement. However, it is actually an information.

If we look at the absolute numbers (see Figure 2 and Tables IX and X), FPR ranges between 0.3 and 0.62 (median). That is, depending on the taxonomy between 30% to 62% of other information is classified as requirement. The variance in the individual results is large.

A higher TPR comes at the price of a higher FPR. The most balanced taxonomy, that is the one with the highest value for ACC, is T_{IEEE} in both groups. With respect to ACC (and also FDR), the results of group 2015 provide statistically significant evidence that the taxonomy T_{IEEE} leads to better results than T_{QR} .

As all three taxonomies are used in the same way - to separate requirements from other information, the question arises what the cause of the differences between the taxonomies really is. One difference between the taxonomies are the denominators (terms that describe the categories of a taxonomy such as *quality* in case T_{QR}) may have an impact on the abilities of a person to classify requirements and on the proper classification of a given statement.

All in all, the results of the two experiments have to be interpreted with care. The results originate from two experiments with a total number of 63 participants and material in German language. Further experiments with a much larger number of participants and other languages are necessary to confirm the presented results. Nevertheless, the results are promising and

motivate us to perform further experiments (see Section VIII for our plans on future research).

VII. THREATS TO VALIDITY

Following the guidelines presented in [31], we discuss threats to the conclusion, internal, construct, and external validity.

Conclusion Validity: The reference point for the measures in our experiment is a list of statements that has been evaluated and classified by three industrial RE expert. We therefore consider our measurement approach as reliable. We exclude most random effects in our experimental setting since our experiment consists of a simple classification task.

Internal Validity: We explicitly mitigated the learning effect in the experimental design by creating groups with a different order of the taxonomies applied. We excluded effects related to the order of taxonomies by randomizing the order of taxonomies for each participant. We mitigated the effect of getting bored during the experiment by including dedicated relaxation exercises. The participants of the experiment were computer science master's students. The generalization of our results to RE professionals therefore could be limited. We argue that the requirements classification task is an easy to learn task. The results show that our participants were able to classify requirements properly. Nevertheless, we intend to replicate our experiment with RE professionals to examine if requirements professionals show the same results.

Construct Validity: The experimental task is simple (classify a statement) and easy to evaluate (compare the experimental results against the expert evaluation). The measurement approach and the measures are based on proven approaches from classification theory. We consider the threats to the construct validity of our experiment as very small.

External Validity: A threat to the external validity of our results is the language. The experimental material was developed using the German language. Therefore, it is possible that our results cannot be reproduced in other languages. We argue that the considered cognitive effects are language independent and therefore, it is very likely that they occur in other languages. We are especially convinced that the effects related to the term *non-functional requirement* will occur in every language that translates this dualism into their own language. For example, non-functional requirement is translated into "requisito no funcional" (Spanish), "wymaganie niefunkcjonalne" (Polish), or "requisiti non funzionali" (Italian) (cf. [16]).

VIII. CONCLUSION

From a high level point of view, our paper argues that selecting the wrong taxonomy could have a very practical impact on daily work: requirements engineers recognize too much other "information" as requirements. Our early experimental results provided initial evidence that this impact could be real. The three taxonomies really have a different classification quality for otherwise identical statements. We have especially observed statistical significant effects for the false positive rate:

between 30% and 62% of other information is classified as requirements depending on the taxonomy.

On the other hand, requirements engineers potentially miss important requirements because they do not recognize them as requirements. We found that between 84% to 100% of the requirements were identified as such. Differences between the taxonomies were not significant.

The design of the experiment is simple and we controlled ordering effects, and we were able to address most other influencing factors such as cheating. Thus, we are convinced that our results are reliable and originate from the above presented cognitive effects, especially from the *priming effect* and from the *illusion of attention*.

Nevertheless, our experiment did not focus on a particular cognitive effect to isolate one effect as the origin of the observed results. Also the theory of how these effects affect the classification task in detail needs more investigation. We plan to address these issues in future research.

Before we start working in this direction, we plan to obtain more evidence for the results presented in this paper since our results originate from two experiments with 63 master's student participants. Furthermore, our results originate from requirement in the German language. To verify our results in a broader context, we plan to replicate the experiment with professionals from industry instead of master students. We further plan to replicate the experiment with requirements written in English to verify if the results can be transferred to another language. To achieve a much larger number of participants, we plan to create a web-based version of our experiment that allows us to distribute and perform the experiment in a more convenient way.

If we collect more evidence to confirm our hypothesis, a practical consequence of the empirical results could be to classify statements in several rounds. In an initial round, statements ought to be binary classified just as *requirement* or *information*. And in subsequent rounds, requirements are classified more thoroughly.

Another avenue for future research is to move from simple, tree-like taxonomies to ontologies for requirements classification. Ontologies could provide a richer basis for classification and possibly lead to a higher classification quality.

Although this paper presents early results, our results are promising and motivate us to perform further research in this direction. For us, the most important conclusion from our paper is, that research on requirements taxonomies and requirements engineering research in general can benefit from ongoing research in cognitive psychology. And by the way, yes, words make a difference. At least when you are looking for non-functional requirements or quality requirements.

REFERENCES

- [1] P. Lenberg, R. Feldt, and L. G. Wallgren, "Behavioral software engineering: A definition and systematic literature review," *Journal of Systems and Software*, vol. 107, pp. 15–37, 2015.
- [2] R. Mohanani, "Implications of requirements engineering on software design: A cognitive insight," in *Proceedings of the 38th International Conference on Software Engineering Companion*, ser. ICSE '16. New York, NY, USA: ACM, 2016, pp. 835–838. [Online]. Available: <http://doi.acm.org/10.1145/2889160.2889254>
- [3] L. Dawson, "Cognitive processes in object-oriented requirements engineering practice: analogical reasoning and mental modelling," in *Information Systems Development*. Springer, 2013, pp. 115–128.
- [4] I. Hadar, "When intuition and logic clash: The case of the object-oriented paradigm," *Science of Computer Programming*, vol. 78, no. 9, pp. 1407–1426, 2013.
- [5] D. P. TEGARDEN and S. D. SHEETZ, "Cognitive activities in oo development," *Int. J. Hum.-Comput. Stud.*, vol. 54, no. 6, pp. 779–798, Jun. 2001. [Online]. Available: <http://dx.doi.org/10.1006/ijhc.1999.0462>
- [6] G. N. Aranda, A. Vizcaino, A. Cechich, and M. Piattini, "A cognitive-based approach to improve distributed requirements elicitation processes," in *Proceedings of the Fourth IEEE International Conference on Cognitive Informatics*, ser. ICCI '05. IEEE Computer Society, 2005, pp. 322–330. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1259589.1259925>
- [7] D. L. Moody, "The physics of notations: Toward a scientific basis for constructing visual notations in software engineering," *Software Engineering, IEEE Transactions on*, vol. 35, no. 6, pp. 756–779, Nov 2009.
- [8] K. Lauenroth and E. Kamsties, "People's capabilities are a blind spot in RE research and practice," in *Requirements Engineering: Foundation for Software Quality - 22nd International Working Conference, REFSQ 2016, Gothenburg, Sweden, March 14-17, 2016, Proceedings*, 2016, pp. 243–248. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-30282-9_17
- [9] M. Glinz, "On non-functional requirements," in *15th IEEE International Requirements Engineering Conference, RE 2007, October 15-19th, 2007, New Delhi, India*, 2007, pp. 21–26.
- [10] P. Loucopoulos, J. Sun, L. Zhao, and F. Heidari, "A systematic classification and analysis of nfrs," in *19th Americas Conference on Information Systems, AMCIS 2013, Chicago, Illinois, USA, August 15-17, 2013*, 2013. [Online]. Available: <http://aisel.aisnet.org/amcis2013/SystemsAnalysis/RoundTablePresentations/1>
- [11] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, "Automated classification of non-functional requirements," *Requir. Eng.*, vol. 12, no. 2, pp. 103–120, 2007. [Online]. Available: <http://dx.doi.org/10.1007/s00766-007-0045-1>
- [12] A. Casamayor, D. Godoy, and M. R. Campo, "Identification of non-functional requirements in textual specifications: A semi-supervised learning approach," *Information & Software Technology*, vol. 52, no. 4, pp. 436–445, 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.infsof.2009.10.010>
- [13] E. Knauss and D. Ott, "(semi-) automatic categorization of natural language requirements," in *Requirements Engineering: Foundation for Software Quality - 20th International Working Conference, REFSQ 2014, Essen, Germany, April 7-10, 2014, Proceedings*, 2014, pp. 39–54. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-05843-6_4
- [14] IEEE, "Standard glossary of software engineering terminology. ieee standard 610.12-1990," 1990.
- [15] R. Young, *The Requirements Engineering Handbook*. Artech House, 2004.
- [16] M. Glinz, "A glossary of requirements engineering terminology - version 1.6 may 2014," 2014. [Online]. Available: http://www.ireb.org/fileadmin/IREB/Download/Homepage%20Downloads/IREB_CPRE_Glossary_16.pdf
- [17] (2014) Dualism (stanford encyclopedia of philosophy), <http://plato.stanford.edu/entries/dualism/>. [Online]. Available: <http://plato.stanford.edu/entries/dualism/>
- [18] R. G. Morrison, "Thinking in working memory," in *The Cambridge Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrison, Eds. Cambridge University Press, 2005, pp. 457–474.
- [19] A. Baddeley and G. Hitch, "Working memory," *The psychology of learning and motivation: Advances in research and theory*, vol. 8, pp. 47–89, 1974.
- [20] E. Tulving, D. L. Schacter, and H. A. Stark, "Priming effects in word-fragment completion are independent of recognition memory," *Journal of experimental psychology: learning, memory, and cognition*, vol. 8, no. 4, pp. 336–342, 1982.
- [21] D. Kahnemann, *Thinking - Fast and Slow*. Penguin Books, 2011.

- [22] C. Chabris and D. Simons, *The Invisible Gorilla and other Ways our Intuition Deceives Us*. HarperCollins Publishers, 2010.
- [23] S. B. Most, "What's 'inattentional' about inattentional blindness?" *Consciousness and Cognition*, vol. 19, no. 4, pp. 1102–1104, 2010.
- [24] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.
- [25] D. C. Montgomery, *Design and analysis of experiments*. John Wiley & Sons, 2008.
- [26] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.
- [27] H. Levene *et al.*, "Robust tests for equality of variances," *Contributions to probability and statistics*, vol. 1, pp. 278–292, 1960.
- [28] C. Burt, "Distribution of intelligence," *British Journal of Psychology*, vol. 48, no. 3, pp. 161–175, 1957. [Online]. Available: <http://dx.doi.org/10.1111/j.2044-8295.1957.tb00614.x>
- [29] L. S. Meyers, G. Gamst, and A. J. Guarino, *Applied multivariate research: Design and interpretation*. Sage, 2006.
- [30] J. Cohen, "Statistical power analysis for the behavioral sciences lawrence earlbaum associates," *Hillsdale, NJ*, pp. 20–26, 1988.
- [31] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, and B. Regnell, *Experimentation in Software Engineering*. Springer, 2012. [Online]. Available: <http://dx.doi.org/10.1007/978-3-642-29044-2>

IX. APPENDIX

	min	median	max
TPR (IEEE)	0.3846	0.8397	1.0000
TPR (NFR)	0.4444	0.8516	1.0000
TPR (QR)	0.2857	1.0000	1.0000
FPR (IEEE)	0.0000	0.3304	0.6667
FPR (NFR)	0.0000	0.4583	0.9000
FPR (QR)	0.1000	0.5417	1.0000
FDR (IEEE)	0.0000	0.2143	0.5556
FDR (NFR)	0.0000	0.3205	0.5833
FDR (QR)	0.1000	0.3640	0.6667
ACC (IEEE)	0.5789	0.7500	1.0000
ACC (NFR)	0.5263	0.7136	0.9000
ACC (QR)	0.4211	0.7000	0.9000

TABLE IX

MIN, MEDIAN, AND MAX VALUE PER PARTICIPANT, METRIC AND TAXONOMY (2014)

	min	median	max
TPR (IEEE)	0.6000	0.9167	1.0000
TPR (NFR)	0.3750	1.0000	1.0000
TPR (QR)	0.5000	1.0000	1.0000
FPR (IEEE)	0.0833	0.3000	0.7500
FPR (NFR)	0.0000	0.4167	0.7500
FPR (QR)	0.0000	0.6250	1.0000
FDR (IEEE)	0.0769	0.2308	0.5625
FDR (NFR)	0.0000	0.2941	0.5294
FDR (QR)	0.0000	0.4000	0.5625
ACC (IEEE)	0.5000	0.8500	0.9500
ACC (NFR)	0.5500	0.7500	0.9000
ACC (QR)	0.5000	0.6500	0.8947

TABLE X

MIN, MEDIAN, AND MAX VALUE PER PARTICIPANT, METRIC AND TAXONOMY (2015)

	W (IEEE)	p (IEEE)	W (NFR)	p (NFR)	W (QR)	p (QR)
TPR	0.9028	0.0133	0.8951	0.0088	0.5913	0.0000
FPR	0.9398	0.1090	0.9654	0.4639	0.9657	0.4701
FDR	0.9507	0.2063	0.9653	0.4619	0.9739	0.6869
ACC	0.9690	0.5543	0.9464	0.1605	0.9598	0.3458

TABLE XI

RESULT OF SHAPIRO-WILK NORMALITY TEST PER METRIC AND TAXONOMY (2014)

	W (IEEE)	p (IEEE)	W (NFR)	p (NFR)	W (QR)	p (QR)
TPR	0.7559	0.0000	0.6609	0.0000	0.5954	0.0000
FPR	0.9147	0.0100	0.9706	0.4610	0.9739	0.5596
FDR	0.9511	0.1227	0.9701	0.4460	0.9259	0.0211
ACC	0.9306	0.0291	0.9212	0.0154	0.9082	0.0066

TABLE XII

RESULT OF SHAPIRO-WILK NORMALITY TEST PER METRIC AND TAXONOMY (2015)

	F-value	p-value
TPR	1.3109	0.2752
FPR	1.2079	0.3041
FDR	0.1097	0.8962
ACC	0.6624	0.5184

TABLE XIII

RESULT OF LEVENE TEST PER METRIC AND TAXONOMY (2014)

	F-value	p-value
TPR	1.0056	0.3694
FPR	0.4822	0.6188
FDR	0.9396	0.3941
ACC	1.4788	0.2328

TABLE XIV

RESULT OF LEVENE TEST PER METRIC AND TAXONOMY (2015)