# An Empirical Investigation of the Component-Based Performance Prediction Method Palladio

Ralf Reussner[1], Steffen Becker[2], Anne Koziolek[1], Heiko Koziolek[3]

[1]Karlsruher Institut für Technologie, Germany
{ralf.reussner,anne.koziolek}@kit.edu
[2]Fachgruppe Softwaretechnik, Universität Paderborn, Germany
steffen.becker@uni-paderborn.de
[3]ABB Corporate Research, Ladenburg, Germany
heiko.koziolek@de.abb.com

**Abstract.** Model-based performance prediction methods aim at evaluating the expected response time, throughput, and resource utilization of a software system at design time, before implementation, to achieve predictability of the system's performance characteristics. Existing performance prediction methods use monolithic, throw-away prediction models or component-based, reusable prediction models. While it is intuitively clear that the development of reusable models requires more effort, the actual higher amount of effort had not been quantified or analysed systematically yet. Furthermore, the achieved prediction accuracy of the methods when applied by developers had not yet been compared. To study the effort, we conducted a controlled experiment with 19 computer science students who predicted the performance of two example systems applying an established, monolithic method (Software Performance Engineering) as well as our own component-based method (Palladio) in 2007. This paper summarizes two earlier papers on this study. The results show that the effort of model creation with Palladio is approximately 1.25 times higher than with SPE in our experimental setting, with the resulting models having comparable prediction accuracy. Therefore, in some cases, the creation of reusable prediction models can already be justified, if they are reused at least once.

**Keywords:** Performance Prediction, Empirical Study, Controlled Experiment

## 1 Introduction

As current applications always ask for maximum performance, performance problems are continuously prevalent in many software systems [20]. Model-based prediction methods [1] try to tackle these problems during early design phases to avoid the problem of implementing architectures which are not able to fulfil certain performance goals. They counter the still popular "fix-it-later" attitude towards performance problems. Many of these methods use designer-friendly UML-based models for software developers, and transform them into formal models (e.g., queueing networks, stochastic Petri-nets, stochastic process algebras), from which performance measures (e.g., response times, throughput) can be derived analytically or via simulation.

During the last decade, researchers have proposed several monolithic prediction approaches (such as SPE [20], uml2LQN [17], umlPSI [2], survey in [1]) and several component-based (CB) prediction approaches (such as CB-SPE [7], ROBOCOP [8], and Palladio [6], survey in [5]). CB approaches try to leverage the benefits of componentry in the sense of Szyperski [21] by reusing well-documented component specifications. This is of particular interest for performance prediction methods, as CB software designs limit the degree of freedom for implementation by (at least partially) reusing existing components. This can also lead to higher performance prediction accuracy. In addition, reusable component prediction models can be composed isomorphically to the software architecture, thereby lowering the effort of performance modelling.

Palladio features highly parametrized component performance specifications, which are better suited for reuse than those of other approaches, because they include more context dependencies (i.e., dependencies to external service calls, usage profile, resource environment). The effort for creating such parametrized, CB models is naturally higher than for throw-away models. However, until now this higher effort has not been investigated systematically. Therefore, it is an open question when it is justified.

Based on this observation, we conducted a controlled experiment in 2007 comparing the effort of applying SPE (as an example for a method with throw-away models) and Palladio (as an example for a method with reusable models). In this paper, we summarize the results from two earlier papers [14, 13] which answer following questions: (Q1) “What is the duration of modelling and predicting with both methods?” and (Q2) “What is the quality of the models in terms of prediction accuracy?”. A more recent paper [15] furthermore investigated the effort reduction from reusing component-based models and found that reusing Palladio models can save time, because effort to reuse can be explained by a model that is independent of the inner complexity of a component.

In our 2007 experiment, we let 19 computer science students apply the methods in an experimental setting. They analysed two CB software systems and assessed the performance impact of additional five design alternatives (e.g., introducing caches, replication, etc.). By using tools accompanying the methods (SPE-ED and PCM-Bench), they predicted response times for two different usage profiles. Therefore we assessed the effort for the combination of applying the method and the corresponding tools.

Our results for question (Q1) show that modelling the whole task (that is the initial system and five additional design alternatives) took in average 1.25 times longer with Palladio than with SPE. For question (Q2), we found that the models created with both approaches allowed a reasonable prediction accuracy to correctly assess the performance of the design alternatives.

This paper is organized as follows. Section 2 presents the basics of model-driven performance prediction and briefly introduces SPE and Palladio. Afterwards, Section 3 explains the experimental design, before Section 4 describes the results. Section 5 discusses the validity of the empirical study. Related work is summarized by Section 6, while Section 7 concludes the paper and sketches future work.

# 2 Model-Driven Performance Prediction

## 2.1 Background

Several model-driven performance prediction approaches have been proposed [1], all of which follow a similar process model (Fig. 1). First, developers annotate plain software design models (e.g., UML models) with estimated or already measured performance properties, such as the execution time for an activity or the number of users concurrently issuing requests.

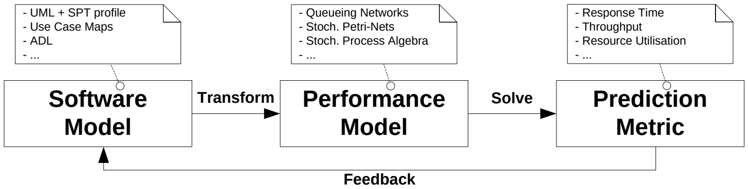


**Fig. 1.** Performance Prediction Process

Second, model transformations automatically convert the annotated software models into established performance formalisms such as queueing networks (QN), stochastic Petri nets (SPN), or stochastic process algebras (SPA). Existing analytical or simulation-based solution techniques then automatically derive and report performance measures, such as response times for specific use cases, maximum throughputs, or the utilization of resources, which is crucial for identifying performance bottlenecks. Developers compare the predicted results to their requirements and decide whether to change their design or to start implementation. Only a few approaches implement an automated feedback of the prediction results into the software design model.

For our experiment, we compared our component-based Palladio method [6] with the mature, monolithic Software Performance Engineering (SPE) method [20]. We chose SPE as it has been applied in practice and provides a reasonably usable tool support, unlike many other approaches [11] solely proposed by academics. The following two sections briefly describe the two approaches, which both follow the process model sketched above.

## 2.2 SPE

The SPE method was the first elaborated, practically applicable comprehensive approach for early, design-time performance prediction for software systems [19]. SPE primarily targets software architects and performance analysts during early development stages. They identify key scenarios (i.e., use cases critical to the overall system performance) and set performance goals for the scenarios (e.g., max. response time) based on the requirements.

Afterwards, developers use a software execution model (Execution Graph, EG) to describe steps within such a performance-critical scenario. EGs are similar to UML activity diagrams and allow annotating each step with resource requirements, for example the number of needed CPU instructions.

With a so-called overhead matrix, software resource requirements in EGs (e.g., a database access) can be mapped to system resources (e.g., 10 ms for a hard disk access per database access). Several scenarios and the corresponding user arrival rates on different machines can be combined to form a system execution model.

EGs do not necessarily reflect actual componentization of a system, but provide an abstraction of the most performance-relevant steps in a scenario. This is useful for conducting performance analyses as early as possible during the life-cycle of a system, when many details are still unknown. It also limits the developers' effort for initial modelling. However, dependencies on the specific project context are not made explicit, but are mixed with component specifics. Thus, it is usually not possible to readily reuse the resulting performance models when reusing the software components. Additionally, the models cannot be used for model-driven development, as their performance-related abstraction does not provide enough information for other purposes like code generation.

The SPE methodology has been applied in industrial settings. Several anonymized case studies are provided in [20].

### 2.3 Palladio Component Model

The Palladio Component Model (PCM) [6] is a meta-model for specifying and analysing component-based software architectures with focus on performance prediction.

This meta-model is divided among the separate developer roles of a component-based development process: The component developer produces independent, reusable component specifications. The other roles (software architects, system deployers, domain experts and quality-of-service analysts) provide information on the project-specific context, such as binding of the components, their allocation to hardware and their usage. The meta-model provides each role with a domain-specific language suited to capture their specific knowledge [6].

To support the creation of reusable component performance models, the component specifications are parametrized by influence factors whose later values are unknown to the component developer. In particular, these are the performance measures of external service calls, which depend on the actual binding of the component's required interfaces (provided by the software architect), the actual resource demands which depend on the allocation of the components to hardware resources (provided by the system deployer), and performance-relevant parameters of service calls (provided by the domain expert).

The parametric behavioural specification used in the PCM as part of the software model is the *Resource Demanding Service Effect Specification* (RD-SEFF) which is a control and data flow abstraction of single component services, also similar to UML activity diagrams. It specifies control flow constructs like loops, or branches only if they affect external service calls. Additionally, they abstract component internal computations in so called *internal actions* which only contain the resource demand (e.g. reading 100 Bytes from a hard drive) of the action but not its concrete behaviour. Calling services and parameter passing are specified using *external call actions*, which only refer to the component's required interfaces to stay independent of the component binding. Hence, unlike EGs, RD-SEFFs reflect the componentization of the system and allow to create component specifications that can be reused in other project contexts. In this experiment, we thus measure the additional effort required to reflect the componentization in the Palladio models (in contrast to the SPE models).

| **Goal** | **Empirically validate the applicability of the performance prediction approach Palladio from a user's point of view.** |
|---|---|
| Question 1 | What is the duration of predicting the performance? |
| Metric 1.1 | Average duration of a prediction |
| Hypothesis 1 | A Palladio prediction needs 1.5 as long as an SPE prediction |
| Question 2 | What is the quality of the created performance prediction models? |
| Metric 2.1 | Relative deviation of predicted mean response times of the participants and of the reference model. |
| Metric 2.2 | Percentage of correct design decisions. |
| Metric 2.3 | Normalized deviation in design decision rankings. |
| Hypothesis 2 | The created models are similar to the reference model. |

**Table 1.** GQM plan overview

## 3 Empirical Investigation

For the empirical investigation, we formulated a goal, two question and derived metrics using the Goal-Question-Metric approach [4]. The goal of this work is:

> **Goal**: Empirically validate the applicability of the performance prediction approach Palladio from a user's point of view.

Questions and metrics are presented in section 3.1. Then, Section 3.2 presents the experiment's design and section 3.3 describes the preparation of the participants. The tasks and the experiment execution are presented in section 3.4 and 3.5, respectively.

### 3.1 Questions and Metrics

For the applicability of the performance prediction models under study, two important factors are (1) the duration of a prediction and (2) the quality of the created models, which is reflected by the two questions presented below. Where appropriate, we compare Palladio to SPE as a baseline. For each metric, we have formulated hypotheses to support the evaluation of the metrics and answer the question. Due to space limitations, only informal explanations of the metrics are given here. The formal definitions can be found in [12, p.35]. Table 1 summarizes goal, question, metrics, and hypotheses.

**Q1: What is the duration of predicting the performance?** To evaluate the effort for making a prediction, we looked at the time needed, i.e. the duration, because time (in terms of person-days) is the major factor of effort and costs. For an empirical study of the effort of any software development technique, it is inevitable to include the used tools. Thus, here we measured the effort for the combination of applying the method (SPE and Palladio) and the corresponding tools (SPE-ED and PCM-Bench).

Metric 1.1 is the average duration of making a performance prediction. The duration includes reading the specification, modelling the control flow, adding resource demands, modelling the resource environment, modelling the usage profile, searching for errors, and analysing.

**Q2: What is the quality of the created performance prediction models?** First, a performance model should enable predictions that are similar to the reference performance model (i.e. the sample solution) when analysed. Here, the predicted response time was an important performance metric. Thus, we defined metric 2.1: *Relative deviation of predicted mean response times of the participants and of the reference model* (percentage).

To assess different design alternatives when designing or changing a system, the relation of the respective response times is also of interest. We let the participants evaluate several design alternatives and measured how many participants correctly identified the best design alternative in respect of its response time by stating metric 2.2: *Percentage of correct design decisions*.

As a software architect does not necessarily choose the design alternative with the best performance, but might consider other quality attributes or cost, the results for the performance-wise inferior design alternatives are also important. Thus, next to identifying the best design alternative, the participants had to rank all alternatives. The ranking of design alternatives by the participants was compared to the ranking of the design alternatives of the reference solution in metric 2.3: *Normalised deviation in design decision rankings*. For this metric, we counted how many ranks lie between the position of a design alternative in the ranking of a participant and the correct position of a this design alternative in the ranking for the reference solution. We normalised this metric so that a correct ranking has a deviation of 0% and the reversed ranking a deviation of 100%. Additionally, we recognised very similar response times as virtually equal design alternatives and did not punish rankings that permuted them.

Our hypothesis is that the created models are similar to the reference model. This can be broken down to (H2.1) that the average deviation as measured with metric 2.1 is not larger than 10%, (H2.2) that 80% of the participants can choose the correct design decision and (H2.3) that the rankings deviate no more than 10% in average for both Palladio and SPE.

### 3.2 Experiment Design

The study was conducted as a controlled experiment and investigated the effort with participants who are not the developers of the approaches. The participants of this study were students of a master's level course (see section 5 for the discussion of student subjects). In an experiment, it is desirable to trace back the observations to changes of one or more independent variables. Therefore, all other variables influencing the results need to be controlled. The *independent variable* in this study was the approach used to make the predictions. Observed *dependent variables* were the duration of making a prediction and the quality of the prediction to ensure a minimum quality.

The experiment was designed as a changeover trial as depicted in figure 2. The participants were divided into two groups, each applying an approach to a given task. In a second session, the groups applied the other approach to a new task. Thus, each participant worked on two tasks in the course of the experiment (inter-subject design) and used both approaches. This allowed to collect more data points and balanced potential differences in individual factors such as skill and motivation between the two experiment groups. Additionally, using two tasks lowered the concrete task's influence and increased the generalizability.

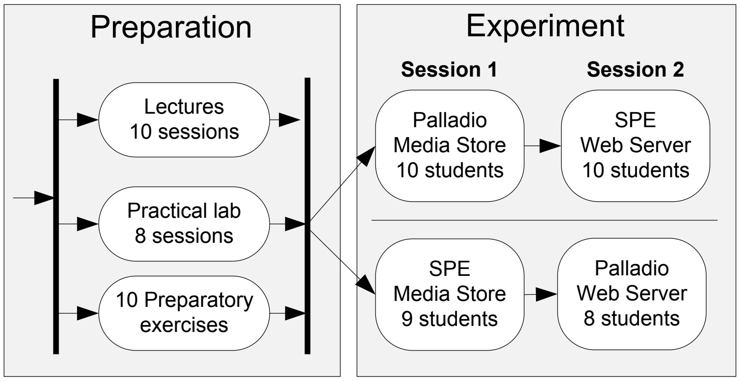

Preparation
Lectures 10 sessions
Practical lab 8 sessions
10 Preparatory exercises
Experiment
Session 1
Session 2
Palladio Media Store 10 students
SPE Web Server 10 students
SPE Media Store 9 students
Palladio Web Server 8 students

**Fig. 2.** Experiment design

We balanced the grouping of the participants based on the results in the preparatory exercises: We divided the more successful half randomly into the two groups, as well as the less successful half, to ensure that the groups were equally well skilled for the tasks. We chose not to use a counter-balanced experiment design, as we would need to further divide the groups, which would disturb the balancing between the groups. We expected a higher threat to validity from the individual participant's performance than from sequencing effects.

Before handing in, the participants' solutions were checked for minimum quality by comparing the created models to the respective reference model. This acceptance test included the comparison of the predicted response time with the reference model's predicted response time as well as a check for the models' well-formedness.

### 3.3 Student teaching

The 19 computer science students participating in the experiment were trained in applying SPE and Palladio during a one-semester course covering both theory and practical labs. For the theory part, there was a total of ten lectures, each of them took 1.5h. The first three lectures were dedicated to foundations of performance prediction and CBSE. Then, two lectures introduced SPE followed by five lectures on Palladio. The three additional lectures on Palladio in comparison to SPE were due to its more complex meta-model which allows for reusable prediction models. Note, that this also shows that reusable models require more training effort. In parallel to the lectures, eight practical labs took place, again, each taking 1.5h. During these sessions, solutions to the accompanying ten exercises were presented and discussed. Five of these exercises practised SPE and five Palladio.

The exercises had to be solved by the participants as homework. We assigned pairs of students to each exercise and shuffled frequently to get different combinations of students work together and exchange knowledge. This was assumed to lower the influence of individual performance in the experiment. Each exercise took the students 4.75h in average to complete.

Overall, the preparation phase was intended to ensure a certain level of familiarity with the tools and concepts, because participants who failed two preparatory exercises or an intermediate short test were excluded from the experiment.

### 3.4 Experiment tasks

To be applicable for both SPE and Palladio, the experiment tasks can only contain aspects that can be realized with both approaches. For example, the tasks cannot make use of the separate roles of Palladio, because these roles are not supported by SPE. Thus, each participant needs to fulfil all roles.

For reasons of compatibility, both experiment tasks had similar set-ups. The task descriptions contained component and sequence diagrams documenting the static and dynamic architecture of a CB system. The sequence diagrams also contained performance annotations. The resource environment with servers and their performance properties was documented textually. The detailed task description is available on-line in [12]. For each system, two usage profiles were given, to reflect both a single-user scenario (*UP1*) and a multi-user scenario leading to contention effects (*UP2*). Additionally, they differed in other performance relevant parameters (see below).

In addition to the initial system, several design alternatives were evaluated. This reflects a common task in software engineering. Four design alternatives were designed to improve the system's performance, and the participants were asked to evaluate which alternative is the most useful one. Three of these alternatives implied the creation of a new component, one changed the allocation of the components and the resource environment by introducing a second machine. With the final fifth alternative, the impact of a change of the component container, namely the introduction of a broker for component lookups, on the performance should be evaluated.

The two systems were prototypical systems specifically designed for this experiment. In the first session, a performance prediction for a web-based system called Media Store was conducted. This system stores music files in a database. Users can either upload or download sets of files. The size of the music files and the number of files to be downloaded are performance-relevant parameters. The five design alternatives were the introduction of a cache component that kept popular music files in memory ($v_1^{MS}$), the usage of a thread pool for database connections ($v_2^{MS}$), the allocation of two of the components to a second machine ($v_3^{MS}$), the addition of a component that reduces the bit rate of uploaded files to reduce the file sizes ($v_4^{MS}$) and the aforementioned usage of a broker ($v_5^{MS}$).

In the second session, a prototypical Web Server system was examined. Here, only one use case was given, a request of an HTML page with further requests of potential embedded multimedia content. Performance-relevant parameters were the number of multimedia objects per page, the size of the content and the proportion of static and dynamic content. The five design alternatives were the introduction of a cache component ($v_1^{WS}$), the aforementioned usage of a broker ($v_2^{WS}$), the parallelisation of the Web Server's logging ($v_3^{WS}$), the allocation of two of the components on a second machine ($v_4^{WS}$) and the usage of a thread pool within the Web Server ($v_5^{WS}$).

The participants using the Palladio approach were provided with the initial repository of available components without RD-SEFFs. It made the tasks for SPE and Palladio more comparable, because the participants still had to create the RD-SEFFs with the performance annotations, which is similar to the creation of an EG in SPE.

### 3.5 Experiment execution

The group of 19 computer science students was divided into two groups as shown in figure 2. We conducted two sessions, each with a maximum time constraint of 4.5 hours. One participant did not attend the second session due to personal reasons, thus, only 18 students took part. The participants were asked to document the duration of the activities given in metric 1.2 and to fill in a questionnaire with qualitative questions at the end of the session.

Four members of our chair were present to help with tool problems, the exercise, and the methods, as well as to check the solutions in the acceptance tests. This might have distorted the results, because they might have influenced the duration. The more problems were solved by the experimentators, the less time the participants might have spent on solving them themselves. To avoid this effect, the participants were asked to first try to solve problems on their own before consulting the experimentators. To be able to assess a possible influence of this help, we documented all help and all rejections in the acceptance test [12].

Because many participants did not finish the task within 4.5 hours in both sessions, the time restriction was loosened afterwards and they were allowed to work another 2.5 hours (session 1) and 2 hours (session 2). In both sessions, three (session 1) respectively two (session 2) participants were not properly prepared, as they needed a lot of basic help or were not able to finish even the initial system prediction. Thus, the results of these three / two participants could not be used. All other participants modelled the initial system and at least one design alternative. Because two participants failed using both approaches, omitting their results does not advantage one of the approaches. Additionally, the time constraints did not distort the results for the initial system prediction, because every remaining participant finished the initial prediction well before the end of the experiment.

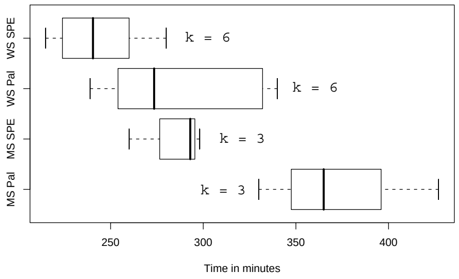
Overall, in session 1, three of the remaining seven participants using Palladio and seven of the nine participants using SPE were able to finish all design alternatives. In session 2, the eight participants using SPE finished all design alternatives, as well as six of the eight participants using Palladio. The acceptance test ensured that the created models were meaningful. As a result, the average deviation of the predicted response time from a reference solution was only about 10%.

## 4 Results

In the paper, we only present the evaluation of the metrics for Palladio. The results for SPE can be found in [12, p.83]. The metrics are evaluated for both tasks. Finally, the hypothesis of each question is checked based on the measured metrics.

### 4.1 What is the Duration of Predicting the Performance?

**Metric 1.1: Average duration of making a prediction** First, we evaluated metric 1.1 for the whole experiment task (=: scope $wt$), thus the duration $d_p$ includes the duration of analysing the initial system and all design alternatives. In neither session, all participants were able to finish the respective task within the extended time constraints, especially for Palladio. We first looked at those participants who finished the whole task with one approach $a$: Let $k_a$ be this number of participants. To not favour one approach, only the results of the $k = max(k_{Pal}, k_{SPE})$ fastest participants from both groups were evaluated for metric 1.1, so that for both groups, the slower participants were left out.



(a) Boxplot

| | MS | WS | Avg |
|---|---|---|---|
| $avd_{Pal}$ | 374 | 285 | 329.5 |
| $avd_{SPE}$ | 284 | 243 | 263.5 |
| $\frac{avd_{Pal}}{avd_{SPE}}$ | 1.32 | 1.17 | 1.25 |

(b) Average

**Fig. 3.** Metric 1.1: Duration of making a prediction in minutes

Figure 3(a) shows the results of metric 1.1 for the four combinations of approaches and systems in a boxplot, showing the minimum, the lower quartile, the mean, the upper quartile and the maximum for all groups and systems. The number of evaluated results is $k = 3$ for the Media Store (MS) and $k = 6$ for the Web Server (WS).

Table 3(b) shows the average metric 1.1. Additionally, we compared how much longer it takes in average to make the Palladio prediction compared to making the respective SPE prediction. These factors are shown as $avd_{Pal}/avd_{SPE}$.

We tested our initial hypotheses using Welch's t-test [22], as we cannot assume identical variances for the distributions, and chose a significance level of 0.05. Hypothesis 1.1 is rejected (p=0.009). Students using Palladio needed significantly less than 1.5 the time than students using SPE, as the opposite is rejected with p=0.004. The statistical power of these two tests is 0.65 and 0.78, respectively, and barely sufficient [18]. Still, students using Palladio needed significantly more time than students using SPE for the whole task as well, as the opposite is rejected (p=0.01, power 0.78).

### 4.2 What is the Quality of the Created Performance Prediction Models?

**Metric 2.1: Relative deviation of predicted mean response times between the participants and the reference model.** Table 2 shows the results of metric 2.1 for Palladio.

| | $v_0^s$ | $v_1^s$ | $v_2^s$ | $v_3^s$ | $v_4^s$ | $v_5^s$ | Avg |
|---|---|---|---|---|---|---|---|
| Media Store $UP1$ | 1.93% | 0.90% | 0.49% | 20.08% | 3.02% | 1.69% | 4.69% |
| $(s = MS)$ $UP2$ | 13.21% | 2.20% | 4.15% | 13.23% | 4.42% | 3.51% | 6.79% |
| Web Server $UP1$ | 1.00% | 11.07% | 1.94% | 4.23% | 4.55% | 9.40% | 5.47% |
| $(s = WS)$ $UP2$ | 15.92% | 20.35% | 10.87% | 10.67% | 2.57% | 3.64% | 10.67% |
| Overall $propDevMeanResp_{Pal}$ | | | | | | | 6.90% |

**Table 2.** Metric 2.1: Relative deviation of the predicted response times for Palladio

We first consider the average deviation for each task. Overall, the deviation is lower using the Media Store and for $UP1$. The overall average is low with 6.9%. Interestingly, the deviation varied a lot between the different design alternatives. For the Media Store and Palladio, the alternative $v_3^{MS}$ (second server), has a high deviation, and $v_0^{MS}$ for the $UP2$, too. For the Web Server and Palladio, the deviations for the $v_2^{WS}$, the broker alternative, $v_0^{WS}$, $v_1^{WS}$ (Cache), and $v_3^{WS}$ (Logging) are also high.

For SPE, we measured a slightly higher average deviation of 8.3% and also strong variations for the different design alternatives.

**Metric 2.2: Percentage of correct design decisions.** For metric 2.2, we compared the results of the reference model (cf. section 3.1) with the participants rankings and assessed the percentage of correct identification of the performance-wise best design alternative. Some participants did not manage to model all alternatives in the given time and thus, their rankings were incomplete and their results cannot be used (see fig. 2 for the total numbers of participants).

As the predicted response time of the best and second-best alternatives of the Media Store were close to each other, we made no distinction between these two. Thus, all participants chose right, because all of them identified either the bit rate ($v_4^{MS}$) or the cache option ($v_1^{MS}$) as the best design alternative and ranked the respective other one second-best.

For the Web Server, $UP1$ and Palladio, 4 out of 6 participants who ranked all alternatives identified the second server $v_4^{WS}$ as the best alternative. Of the two others, one actually predicted a lower response time for the cache ($v_1^{WS}$), the other one seemed to have other reasons or could not correctly interpret the CDF, as the second server $v_4^{WS}$ is faster for his model, too. We get $perc_{WS,UP1,Pal} = 0.67$. All eight SPE participants chose the right alternative: $perc_{WS,UP1,SPE} = 1$.

For usage model 2, all five Palladio participants who ranked all alternatives identified the second server $v_4^{WS}$ as the best alternative. For SPE, 7 out of 8 participants who ranked all alternatives did so: $perc_{WS,UP2,SPE} = 0.88$.

Combined[1] we get $perc_{SPE}$ = 0.97 and $perc_{Pal}$ = 0.85.

**Metric 2.3: Normalised deviation in design decision rankings** Not all participants ranked all alternatives, because they did not complete all predictions or missed the time to complete the ranking, even if they completed the predictions. We still used the incomplete rankings for the evaluation of the metrics, but were careful to weight complete rankings stronger (cf. [12, p.86f]).

For Palladio, the ranks were wrong by 6.5% of the maximum possible permutation. For SPE, the ranks were wrong by 7.3% of the maximum possible permutation. Thus, SPE rankings were more permuted by factor 0.12 compared to Palladio rankings.

**Hypothesis 2** With both approaches, the mean response time predicted by the participants only deviates in average 6.9% (Palladio) and 8.3% (SPE) from the mean response time predicted for the reference model. Thus, the deviation of the average is within the limit of 10%. However, for single alternatives, the deviation was higher (see table 2). These pose a threat to hypothesis 2.

Most participants also were able to identify the correct design decisions, in particular 85% for Palladio and 97% for SPE, both is within the bounds of 80%. Finally, the deviation of the ranking is also low (not more than 10% in average).

Overall, the results indicate that hypothesis 1 cannot be rejected for the average case. However, the high variation of the deviation of the predicted mean response time between the different design alternatives hampers assessing hypothesis 1. As the alternatives have differing results, it is unclear how the metrics would be evaluated for different design alternatives.

## 5 Threats to Validity

To enable the reader to assess our study, we list some potential threats its validity in the following. We look at the internal, construct, and external validity (a more thorough discussion can be found in [12]).

[1] Note that the percentages for the two systems do not equally influence the results, but are weighted by the number of decisions by definition of the metric (cf. [12, p.41])

The *internal validity* states whether changes of an experiment's independent variables are in fact the cause for changes of the dependent variables [23, p.68]. Controlling potential interfering variables ensures a high internal validity. In our experiment, we evaluated the pre-experiment exercises and assigned the students to equally capable groups based on the results to control the different capabilities of the participants. A learning effect might be an interfering variable in our experiment, as the students finished the second experiment session faster than the first one.

A potential bias towards or against Palladio was threatening the internal validity in our experiment, as the participants knew that the experimenters were involved in creating this method. However, we did not notice a strong bias from the collected data and the filled-out questionnaires, as the participants complained equally often about the tools of both approaches.

The *construct validity* states whether the persons and settings used in an experiment represent the analysed constructs well [23, p.71]. Palladio and SPE are both typical performance prediction methods involving UML-like design models. The SPE approach has no special support for component-based systems, and was chosen for the experiment due to its higher maturity compared to existing CBSPE approaches. To allow a comparison, we designed the experimental tasks so that not all specific component-based features of Palladio (e.g. separation of developer roles in component-based development, performance requirements using quantiles) were used.

While our experiment involved student participants, we argue that their performance after the training sessions was comparable to the potential performance of practitioners. Most of the students were close to graduating and will become practitioners soon. Due to the training sessions, their knowledge about the methods was more homogeneous than the knowledge of practitioners with different backgrounds. With a homogeneous group of participants, the significance of the results is even improved. Studies, such as [10], suggest the suitability of students for similar experiments.

The *external validity* states whether the results of an experiment are transferable to other settings than the specific experimental setting [23, p.72]. While we used medium-sized, self-designed systems for the students to analyse, we modelled these system designs and the alternatives after typical distributed systems and commonly known performance patterns [20], which should be representative for the usually analysed systems.

We tried to increase the external validity of our study by letting the participants analyse two different systems, so that differences in the results could be traced back to the systems, and not the prediction methods. Effects that are observed for both tasks are thus more likely to be generalizable to other settings.

Still, the systems under study were modelled on a high abstraction level due to the time constraints of such an experiment. More complex systems would increase the external validity, but would also involve more interfering variables thus decreasing the internal validity. Furthermore, the available information at early development stages is usually limited, which would be reflected by our experimental setting.

## 6 Related Work

Basics about the area of *performance prediction* can be found in [20, 16]. Balsamo et al. [1] give an overview of about 20 recent approaches based on QN, SPN, and SPA. Becker et al. [5] survey performance prediction methods specifically targeting component-based systems. Examples are CB-SPE [7], ROBOCOP [8], and CBML [24].

*Empirical studies* and controlled experiments [23] are still under-represented in the field of model-based performance predictions, as hardly any studies comparable to ours can be found. Balsamo et al. [3] compared two complementary prediction methods (one based on SPA, one on simulation) by analysing the performance of a naval communication system. However, in that study, the authors of the methods carried out the predictions themselves. Gorton et al. [9] compared predicted performance metrics to measurements in a study, but only used one method for the predictions.

Koziolek et al. [11] conducted a study similar to this one. They compare predictions with SPE [20], Capacity Planning [16], and umlPSI [2] with measurements of an implementation. It attested SPE the most maturity and suitability for early performance predictions and influenced our decision to compare Palladio to SPE.

## 7 Conclusions

We have conducted an empirical investigation to quantify the higher effort for creating reusable, component-based models for performance prediction in relation to create throw-away models. After substantial training, we let 19 computer science students apply the SPE method and the Palladio method to predict the response times of two example systems. We found that the effort for applying Palladio on the whole task was in average 1.25 times the effort for applying SPE. Our results indicate that in some cases, the effort of creating reusable models for performance prediction can already be justified if the models are reused at least once, if the costs for the reuse itself are low. If the models are reused more often, the additional upfront effort pays off even more. A more recent study [15] indeed confirms that reusing Palladio models can save time, because effort to reuse can be explained by a model that is independent of the inner complexity of a component. Furthermore, we found that the quality of the models and predictions created by the students deviated less than 10 % from the predictions achieved with a reference model created by the experimentators. We learned that more than 80% of students were able to rank the given design alternatives correctly.

The results are useful for both practitioners and researchers. Practitioners, such as software architects and performance analysts, get a first quantification of the higher effort to create reusable, component-based models, which they could use in front of management to justify higher upfront costs for modelling. Researchers obtain a reusable experimental setting, which is the basis for future replications of the experiment. The results suggest that it is worthwhile to put more research effort into creating reusable models, because their creation can quickly pay off. However, our study cannot give a definite, overall answer to the questions raised, as the results are also confined to our specific experimental setting.

Our investigation opens up future directions for research. The study could be repeated with a larger sample size to allow a better quantification of the additional effort as well as a validation of the results. Moreover, the analysis of factors influencing the effort, especially the nature of the systems under study, is an issue for future research.

**Acknowledgements** We would like to thank Walter Tichy, Lutz Prechelt, and Wilhelm Hasselbring for their kind review of the experimental design and fruitful comments. Furthermore, we thank all members of the SDQ Chair for helping prepare and conduct the experiment. Last, but not least, we thank all students who volunteered to participate in our experiment.

## References

1. S. Balsamo, A. Di Marco, P. Inverardi, and M. Simeoni. Model-Based Performance Prediction in Software Development: A Survey. *IEEE TSE*, 30(5):295–310, May 2004.
2. S. Balsamo and M. Marzolla. A Simulation-Based Approach to Software Performance Modeling. In *Proc. of ESEC/FSE*, pages 363–366. ACM Press, 2003.
3. S. Balsamo, M. Marzolla, A. Di Marco, and P. Inverardi. Experimenting different software architectures performance techniques. In *Proc. of WOSP*, pages 115–119. ACM Press, 2004.
4. V. R. Basili, G. Caldiera, and H. D. Rombach. The Goal Question Metric Approach. In J. J. Marciniak, editor, *Encyclopedia of Software Engineering - 2 Volume Set*, pages 528–532. John Wiley & Sons, 1994.
5. S. Becker, L. Grunske, R. Mirandola, and S. Overhage. Performance Prediction of Component-Based Systems: A Survey from an Engineering Perspective. In *Architecting Systems with Trustworthy Components*, volume 3938 of *LNCS*, pages 169–192. Springer, 2006.
6. S. Becker, H. Koziolek, and R. Reussner. Model-based Performance Prediction with the Palladio Component Model. In *Proc. of WOSP*, pages 54–65. ACM Sigsoft, February5–8 2007.
7. A. Bertolino and R. Mirandola. CB-SPE Tool: Putting Component-Based Performance Engineering into Practice. In *Proc. of CBSE*, volume 3054 of *LNCS*, pages 233–248. Springer, 2004.
8. E. Bondarev, J. Muskens, P. H. N. de With, M. R. V. Chaudron, and J. Lukkien. Predicting real-time properties of component assemblies: A scenario-simulation approach. In *Proc. 30th EUROMICRO-Conference*, pages 40–47, 2004.
9. I. Gorton and A. Liu. Performance Evaluation of Alternative Component Architectures for Enterprise JavaBean Applications. *IEEE Internet Computing*, 7(3):18–23, 2003.
10. M. Höst, B. Regnell, and C. Wohlin. Using students as subjects - A comparative study of students and professionals in lead-time impact assessment. *Empirical Software Engineering*, 5(3):201–214, 2000.
11. H. Koziolek and V. Firus. Empirical Evaluation of Model-based Performance Predictions Methods in Software Development. In *Proc. of QoSA*, volume 3712 of *LNCS*, pages 188–202, Erfurt, Germany, Sep. 2005.
12. A. Martens. Empirical Validation of the Model-driven Performance Prediction Approach Palladio. Master's thesis, Carl-von-Ossietzky Universität Oldenburg, Nov. 2007.
13. A. Martens, S. Becker, H. Koziolek, and R. Reussner. An Empirical Investigation of the Applicability of a Component-Based Performance Prediction Method. In N. Thomas and C. Juiz, editors, *Proceedings of the 5th European Performance Engineering Workshop (EPEW'08), Palma de Mallorca, Spain*, volume 5261 of *Lecture Notes in Computer Science*, pages 17–31. Springer-Verlag Berlin Heidelberg, 2008.

14. A. Martens, S. Becker, H. Koziolek, and R. Reussner. An Empirical Investigation of the Effort of Creating Reusable Models for Performance Prediction. In *Proceedings of the 11th International Symposium on Component-Based Software Engineering (CBSE'08), Karlsruhe, Germany*, volume 5282 of *Lecture Notes in Computer Science*, pages 16–31. Springer-Verlag Berlin Heidelberg, 2008.
15. A. Martens, H. Koziolek, L. Prechelt, and R. Reussner. From monolithic to component-based performance evaluation of software architectures. *Empirical Software Engineering*, 16(5):587–622, 2011.
16. D. A. Menascé, V. A. F. Almeida, and L. W. Dowdy. *Performance by Design*. Prentice Hall, 2004.
17. D. C. Petriu and X. Wang. From UML description of high-level software architecture to LQN performance models. In M. Nagl, A. Schürr, and M. Münch, editors, *Proc. of AGTIVE'99 Kerkrade*, volume 1779. Springer, 2000.
18. L. Sachs. *Applied Statistics: A Handbook of Techniques*. Springer-Verlag, New York, USA, 1982.
19. C. U. Smith. *Performance Engineering of Software Systems*. Addison-Wesley, Reading, MA, USA, 1990.
20. C. U. Smith and L. G. Williams. *Performance Solutions: A Practical Guide to Creating Responsive, Scalable Software*. Addison-Wesley, 2002.
21. C. Szyperski. *Component Software: Beyond Object-Oriented Programming*. ACM Press, Addison-Wesley, Reading, MA, USA, 1998.
22. B. L. Welch. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.
23. C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering: an Introduction*. Kluwer Academic Publishers, Norwell, MA, USA, 2000.
24. X. Wu and M. Woodside. Performance Modeling from Software Components. *SIGSOFT SE Notes*, 29(1):290–301, 2004.