



Quality Testing Concepts

Paul Erlenwein

paul.erlenwein@tu-dresden.de

Born on: 31st December 1996 in Ludwigshafen

Course: Distributed Systems Engineering

Matriculation number: 4609464

Matriculation year: 2016

Bachelor Thesis

to achieve the academic degree

Bachelor of Science (B.Sc.)

Supervisors

Dr.-Ing. Birgit Demuth

Markus Hamann

Dr. Sebastian Götz

Supervising professor

Prof. Dr. rer. nat habil. Uwe Aßmann

Submitted on: 5th February 2021



Contents

1. Introduction	5
1.1. Motivation	5
1.2. Research Questions	6
2. Related Work	7
2.1. INLOOM	7
2.2. Quality-Testing in existing Model Evaluation Systems	8
2.2.1. Grade Quotient Strategy	10
2.2.2. Grace Points and Clean Scores	10
2.2.3. Inter-Rater-Reliability	11
2.2.4. Informedness, Markedness	11
3. INLOOM QT: A facility for quality testing INLOOM	12
3.1. Software Requirements	12
3.2. Available Data	12
3.2.1. INLOOM Result Files	13
3.2.2. Manually Evaluated Student Solutions	15
3.2.3. Data Driven Data Structure	16
3.2.4. Digitizing Manual Evaluations	17
3.3. KPI for Testing	18
3.3.1. Comparison Detail Levels	19
3.3.2. Categories of Comparison	20
3.3.3. Visualizing Validation Results	20
3.4. Summary	21
Bibliography	23
A. Weitere Latex-Dokumentation	26

List of Figures

- 3.1. Abstracted workflow of the creation of manual and automatic evaluations. Rectangles mark data elements, while ellipses represent process steps. The rectangles marked in green, represent the data available for testing purposes. 13
- 3.2. Data model, that results from the structural analysis of the available data. 17

List of Tables

2.1. Quality validation performed in systems for student model evaluation.	10
--	----

1. Introduction

1.1. Motivation

2020 was jinxed. The Covid-19 pandemic changed how our life's work. It has presented the globalized world with little anticipated challenges and we can feel its influence almost every aspect of our everyday life. In order to reduce the amount of human contacts as much as possible, every aspect of human interaction was evaluated for its digitizeability. However: We were all forced to witness, that our digital infrastructure is most obviously not yet up to the task of enabling us to *live the remote life*.

Most of the fundamental problems were trivialities, like missing webcams or a too slow internet connection. Where such were taken care of, the harder-to-fix problems came to light [29]. Problems like inadequately educated overtaxed personnel and missing software solutions, that comply with European digital privacy regulation. At first glance the personnel problems don't seem to matter much to software developers, it still are problems, which, I firmly believe, can and will, at least in part, be resolved by them.

I don't want to claim, that the digitization of everyday life was a complete failure though. Like me, most office workers were able to migrate to home office without much fuss. Still: Living the life of a remote student for half a year, definitely motivates me, to spend some thought on how to make e-teaching a little better.

Even though it proved hazardous sometimes, working with existing e-teaching tools made me realize, what huge potential lies within a properly digitized higher education. Such would not only help temper the effects emergencies, like the Covid pandemic, have on university life, but will also be a powerful tool in futureproofing universities [12] for the challenges of rising student numbers [33] in the years to come.

Intelligent-Tutoring-Systems like INLOOM, an ITS under active development at TU Dresden will make the increased workload manageable for university personnel. Developers are required to produce software that is as intuitive as possible, provides a decent grade of digital security, complies with privacy regulations, handles high traffic without complain and all that, while providing an unquestionably accurate and fair environment for everyone involved.

Integrating digital resources into their workflow seamlessly, will enable teaching personnel, to still be able to focus on the individual student, when the student groups they teach become way bigger, than they are today.

1.2. Research Questions

The goal of this thesis is finding or developing a concept for validating the quality of automatically generated evaluations of student created models. The result should be the functioning prototype of an application for quality testing INLOOM [14]. Quality testing is necessary, to ensure the correctness and fairness of the evaluations, INLOOM generates for student created solution-models.

Since not all mistakes a student could make can be predicted, it will not be possible to ensure, that the automatic grading INLOOM performs, finds all errors a student solution contains. Therefore it will not be achievable to perform an objective evaluation of INLOOMs results. Evaluating model evaluations is an ill-defined problem.

The only remaining option is performing a relative evaluation, a comparison of INLOOMs results with the best evaluation of the same model we know. These best evaluations, in INLOOMs case, are manual ones of the same student solution, created by human tutors (manEval). This in turn means, that the possibilities for validating INLOOMs results, are severely limited by the availability and ascertainability of the underlying data.

For the purpose of validating INLOOMs evaluations I will aim to answer the following research questions:

- RQ1 Which values can be extracted from the manual and automatic evaluations?
- RQ2 What comparative scale is qualified to provide a conclusive impression on the quality of the evaluations INLOOM generates?
- RQ3 What methodologies are employed by existing ITS to validate their results?
- RQ4 How can the developer/tutor/instructor best be assisted in collecting and pre-processing the evaluation data, required for the quality validation?
- RQ5 How can the results of the comparison of man- and autoEval be presented to greatest effect?

The answer to RQ1 will determine which values are available to facilitate a comparison between man- and autoEval. The question is approached with an analysis of the available datasets. The second research question focusses on what to do with the data, once it is collected. Assuming, that comparable data can be extracted from both man- and autoEval (so basically that RQ1 can be answered successfully), a way to combine the found relative differences between the two evaluations, into a conclusive quality score, is still required. Answering the question will be approached by performing a literature survey, concerning the quality validation methodologies of existing ITS and automatic grading methods. Under any envisionable circumstances, it will be necessary, to digitize manEvals a tutor created for student solutions. This digitization process entails a high amount of overhead for the creation of testcases, that should be reduced as much as possible. Research Question three (RQ3), aims to resolve this problem and will be tackled by an analysis of the test-creation process, which will determine the, most workload intensive, steps of the process. These steps can then be considered in the design of the software solution proposed in this thesis. Lastly, it remains to be determined how to best present or visualize the results of the meta evaluation (the evaluation of the evaluations) to the developer/tutor/instructor. This is the reason why RQ4 is listed among the research questions. Answering it, will depend greatly on the answer to RQ2.

2. Related Work

2.1. INLOOM

This work aims to validate the quality of the results, the INLOOM software [14] produces. For that reason it is inevitable, to take a look into what the software does and how it works.

INLOOM is an acronym for *INteractive Learning center for Object-Oriented Modelling*. The software, as the name suggests, is intended to be employed in a learning environment. It is used to evaluate student solutions to modelling tasks, the students have to work on, as part of the mandatory beginner software engineering course, at TU Dresden and is specifically designed to aid in teaching *Object-oriented analysis* and *Object-oriented design*.

INLOOM was originally developed as an extension to the existing INLOOP[19] Software, which allows students to solve programming tasks online and to evaluate their results using supplied JUnit testcases. While in evaluating code, the task of judging whether or not the entered solution is correct, is rather easy, for software models the same task becomes way more complex. That is because, for a given modelling task, there may exist multiple correct solutions, which makes defining one optimal solution, that can be used to compare student solutions against, much harder, if not impossible. Additionally, the comparison process itself is much more challenging, than to simply check, whether or not a supplied code fragment passes a test or not. This is again due to the fact, that whoever creates a model, enjoys a wide range of freedom in solving the supplied task.

The student solution can for example use a different naming scheme than the expert solution it is compared against, or model a property using aggregation or composition, where the expert solution does not and still be correct.

INLOOM evaluates supplied EMF (*Eclipse Modelling Framework*) models, by performing a number of *constraint tests*. In a first step, a *Constraint-based Test Generator* generates a number of constraint files, from an expert solution. Each of those holds constraints INLOOM later applies to the input student model. One constraint file is generated per found element and holds all constraints that are to be applied to this element.

A constraint, in this context, means a *feature requirement* applied to the student model. If, for example, the expert solution to the given modelling task, contains a class called "Student", the student solution is expected to contain an equivalent element. This expectation is expressed by the existence of a constraint, that checks for the presence of the class "Student" in the student solution.

By extracting the constraints from an existing UML model automatically, the instructor, who wants to create a new task, is not required to have any deeper understanding of how the constraints work or how they are implemented. The instructors work is reduced to supplying an expert solution for the modelling task he creates.

The constraint generation is made possible by the existence of a *Master Constraint-Set*, which is basically a collection of constraint templates. The *Master Constraint Set* is individually compiled per diagram type, since the features, one wishes to check for, will inevitably vary, depending on the input type of UML-model. What constraints are employed, when evaluating a student solution, depends on the type of the model in the student solution and the *Master Constraint-Set*.

For each constraint, INLOOM generates an output, identifying the constraint used and containing information about, if and how the constraint was satisfied by the student solution. The constraint result stores, how many points were awarded for the feature checked by the constraint. It also identifies that checked feature, stores which element of the solution was checked and assigns a category flag to the result. Each such category flag relates to a number of points awarded. All the constraint results, thus created for the student solution, are collected in a common output XML file.

In addition to the constraint results, this XML also contains some meta data, like the identification of the student who produced the evaluated solution, which expert solution it was compared to and the exercise the student tries to solve with the supplied diagram. Additionally, the total points achieved, as well as the maximum points that can be achieved, are stored in the XML.

2.2. Quality-Testing in existing Model Evaluation Systems

Due to rising student numbers and the availability of modern interfacing technologies, the interest, in the automatic evaluation of student modelling work, has increased in recent years. Even though: Functioning evaluation tools and methods remain scarce [14].



In this section, the results of a literature research, into concepts for validating the quality of evaluations, automatic grading systems produce, are presented. Starting point for the research, was a collection of such systems, referenced in [14]. Since the design of the listed systems influenced decisions, made during the design of INLOOM, it is only natural to also focus on them in a pre-study, that aims to identify possibilities to validate an alike system. The collection is also rather recent and quite complete, since an accompanying research into automatic grading systems and intelligent tutoring systems (ITS), did not turn up any software solutions, that were not listed.



The listing in [14] differentiates between types of evaluation systems. There are the two classes: *System* and *Method* - as well as two classes for the systems input: *Web* and *Tool*. This differentiation will not be made here. The undertaken research showed, that it does not play a major role, in terms of the applied strategy for quality validation, of which of the classes the described system is.

The systems were examined with regard to their quality assurance measures. Table ?? lists the analyzed systems and indicates the kind of quality validation they perform. In some cases the undertaken quality assurance measures were described in a separate paper. In these cases, the system, along with the additional literature is consolidated in one table entry.

In order to avoid repeating the description of a frequently employed concept for quality validation, it is opportune, to list such first. There are three trends, in terms of quality validation in student model evaluation. One of the most common strategies seems to be, not to validate the produced evaluations at all. At least no publication,

of a description of the validation process, could be found for six of sixteen, of the examined systems [3, 11, 17, 23, 27] (*No Validation*). This is not to mean, that they do not describe *any* testing process! It is meant to say, that these systems do not describe a process to test the *quality* of their results.

Quality, in the context of this paper, can be equated with the confidence an instructor, using a tool to automatically evaluate student solutions, has in that tool. Angry and too happy students indicate a low *quality* and are to be avoided. The tool does a *good job* if it evaluates the student created models *correctly*. It awards points exactly where due. If that is the case, it is very unlikely, that the instructor will hear *any grumbling from his students* and he can use the tool with a high amount of confidence. If, however, the tool he employs, does a *bad job* and for example, does not recognize valid inputs as such or finds errors where there are none, his students will be fast to complain and the instructors confidence in the tool will decrease.

Maybe not as bad, but also to be avoided, are *too happy* students. If the tool awards points where none are due, it is not doing a *good job* and the instructor again will not be able to employ it with much confidence.¹

Seven of the publications describe a validation that is based on a comparison of the *grade equivalent of the respective method* [4–6, 26, 28, 32]. Such strategies will be discussed in their own subsection and are combined under the umbrella term *Grade Quotient Strategy*. This kind of quality validation is also described for INLOOM in [14].

For the rest of the systems *no strategy, regarding a validation* of their evaluations quality, is described. The authors do, however, detail strategies for performing an evaluation, regarding the *didactic* use of their respective tools [2, 21, 24] (*Didactic Evaluation*). Such an evaluation might be interesting in the future, but is not within the scope of any of the questions this thesis is intended to answer. These approaches were not examined any closer.

It must be mentioned, that there do exist quite a few more systems for evaluating student created UML models [1, 18, 20]. These are not listed and were not examined closer, because they do not employ a *constraint-based approach* or at least result in a grade. The research *was thus* limited for the following reason: The data available for quality-testing INLOOM is limited, as will be elaborated in a later section, to the results a constraint-based system *can* produce. A constraint-based approach obviously results in data that details, which constraints were met and which were not.

A superficial analysis of systems, that approach the evaluation differently, was performed. It was found, that evaluation-strategies, that are not constraint-based, or do at least result in a grade, will approach validating the quality of their results, with a completely different focus, than it is required in the context of INLOOM. Any strategy, for evaluating the quality of a fundamentally different result, will ultimately not be applicable, since it is not based on data anything alike the results, INLOOM produces.

¹This paragraph might appear trivial and unnecessarily prosaic, but it actually took me a while to define the term for myself. Quality implies a lot of things, but actually describing what it *is*, turned out to be rather difficult. Quality is a measure between perfect and really bad. What is either good or bad however, can depend on any number of arbitrary factors. The metaphor describes my understanding of *quality* accurately and is intended to convey it.

Table 2.1. Quality validation performed in systems for student model evaluation.

	System	Source(s)	Quality Validation Performed
💬	INLOOM	[14]	Grade Quotient Strategy
	Bian	[5, 7]	Grade Quotient Strategy
	Schramm	[24]	Didactic Evaluation
💬	UML GRADER	[15]	Grade Quotient Strategy
	Striewe	[28]	Grade Quotient Strategy
	Demuth	[11]	No Validation
	Baghaei	[2]	Didactic Evaluation
	Le	[21]	Didactic Evaluation
	CourseMaster	[16, 17]	No Evaluation
	Artemis	[4, 20]	Grade Quotient Strategy
	Beck	[3]	No Evaluation
	Sousa	[27]	No Evaluation
	Smith & Thomas	[26, 30, 31]	Grade Quotient Strategy
	Prados	[23]	No Evaluation
	Tselonis	[32]	Grade Quotient Strategy

2.2.1. Grade Quotient Strategy

A *Grade Quotient Strategy* describes a strategy, that is based on calculating the percentage deviation of the automatically generated evaluations result, from a manually created one. The literature agrees, that manual evaluations of student models are - speaking in terms of quality - the *best* evaluations known. Therefore, they are the measure of quality applied to automatic evaluations. Generally, in order to calculate some kind of *Grade Quotient* the authors collect as many real live student solutions, as they can and evaluate them. Once, using their tool (or tools[8]) and once again, manually. Two *grades* are extracted from each student solution this way. The difference between the two can now be expressed by a quotient. The quotient is usually presented as a difference-percentage.

An especially mentionable paper, whos authors employed a *Grade Quotient Strategy*, was [8]. An extensive effort to compare different evaluation methodologies is described. However, the described validation process is not specialized in validating the results of a constraint-based approach, as the described process' is designed to be general, since it must consider a mutable evaluation format for both the manual and the automatic evaluation.

2.2.2. Grace Points and Clean Scores

The literature points out that the ratings of different reviewers can differ significantly. [14] describes the concept of *grace points*. Points that should not have been awarded if the instructor had followed the evaluation scheme exactly during the correction. It is the instructors prerogative to turn a blind eye, if the students solution is *close enough*. If a constraint-based evaluation tool was to make such a judgement call however, it would indicate something is broken.

Since every evaluator has $(.*|their)^2$ own style and preferences, this can lead to a measurable deviation between two manual evaluations of the same student solution, that were created by different reviewers. *Clean scores* and other concepts like *moderated human marks*[30] are designed to mitigate the negative effects the described deviations

²Ha. A regex Joke.

have on the applicability of the manual evaluations as a quality standard. By averaging multiple manual evaluations of the same student solution, an even *better* quality standard, the automatic evaluation must measure up against, is created.

In order to decrease the influence, the differing preferences of evaluators can have on the value of their evaluation, as a benchmark, even further, [30] employs *Inter-Rater-Reliability* statistics.

2.2.3. Inter-Rater-Reliability



Statistics like Cohens Kappa[9], Scott's Pi[25], Fleiss Kappa[13] and Gwets AC1-Statistic [10] are measurements for the likeness of the work, of two or more evaluators. How many evaluators can be compared at once varies between the listed statistics. All of these values originate from a psychological or sociological background. They are designed to compare evaluators whos evaluation consists of categorizing a person in a number of categories, based on the answers to arbitrary questions and chance observations.

In this kind of scenario, one has to factor in, or rather out, a chance agreement of the evaluators, who all have their individual practices, yet have to use the same form. Inter-Rater-Reliability statistics aim to remove the habitual or by chance agreement, which results, from the equation.

[30] employed both Fleiss Kappa and the AC1-Statistic, to compare the results of their evaluation tool with manual evaluations. Fleiss Kappa is the more general of the two approaches and most Inter-Rater-Reliability statistics are calculated in an, at least similar, fashion. All Inter-Rater-Reliability statistics have in common, that they are intended to compare evaluation processes that happen on a nominal scale.

2.2.4. Informedness, Markedness



A popular and very well known technique to validate the results of machine learning algorithms, is to calculate *Precision and Recall*. It is commonly used in scenarios, where an algorithm has to make a binary decision. *Precision* specifies for how many of the items, the algorithm chose a certain option for, that decision was correct. *Recall* specifies how many of the items, for that a certain option would have been correct, were categorized correctly.

The concept is extended, by *Informedness and Markedness* [22]. Unlike *Precision and Recall* these consider the *true negatives* in the calculation of the benchmark and thus factor in the general likelihood of an item, to be of a certain category.

3. INLOOM QT: A facility for quality testing INLOOM

This chapter aims to present the design for a software, that allows for the validation of the evaluations INLOOM [14] produces. For this purpose, the requirements that are placed on the intended application are first compiled. The datasets that are available for the validation are analyzed and a data structure, to be employed by the application, is derived. Continuing to follow this data-driven approach, a possible design for the intended application is presented.

3.1. Software Requirements

INLOOM is supposed to evaluate student submissions to modelling tasks. Since this evaluation is to be done without further human supervision, the system must be tested thoroughly, to avoid grading students unfairly or in error. The maintainer of INLOOM must be enabled to get an insight into the current quality of INLOOMs results and to quickly react to newly encountered sources of error.


This leads directly to two leading requirements (RQ) for the software proposed here.

RQ1 The software must be able to automatically test the *quality* of the results INLOOM generates.

RQ2 It must be possible to present the results of the performed validation in an easily comprehensible way, to quickly gain insight about the current state of affairs.

3.2. Available Data

The biggest limitation for the test system is the availability of test data. Since the complexity of solutions to modelling tasks cannot be predicted and there can be multiple correct solutions to the same task, the literature agrees, that the only feasible method of validating the evaluations, an automatic grading tool like INLOOM creates, is comparing the automatic evaluation to a manual one, which was created for the same student solution.

The usefulness of faking student solutions for this purpose is limited. Any testing done, using faked up data, would ultimately result in unit tests for the constraints. Thus, the only way to gain a reliable impression of the quality of INLOOMs results, is to use it in a live scenario or to at least use real data for testing. 

Since every other part of the design, depends on the underlying data structure, which in turn heavily depends on the data available, this leads to an obvious third requirement or rather an important limitation for the software proposed here.



RQ3 All tests must be performed using the data available.

As mentioned before, the data available consists of automatically generated evaluations for already manually graded student solutions, as well as the respective manual evaluations, in analog form.

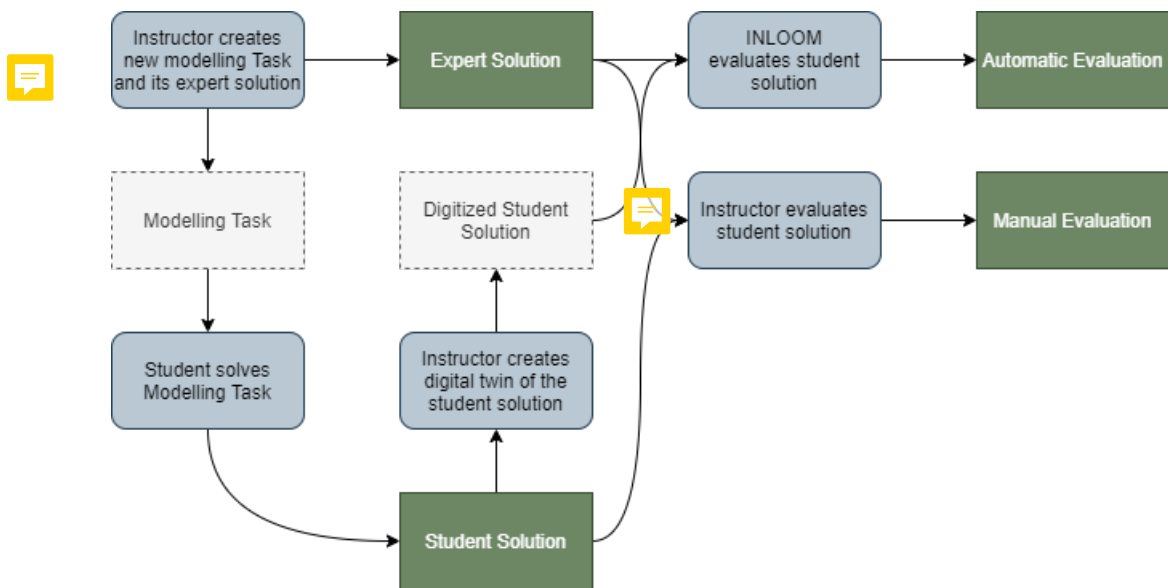


Figure 3.1. Abstracted workflow of the creation of manual and automatic evaluations. Rectangles mark data elements, while ellipses represent process steps. The rectangles marked in green, represent the data available for testing purposes.



RQ3.1 The software must employ a data structure, that is able to hold information collected on automatic and manual evaluations and enables comparing the two.

3.2.1. INLOOM Result Files

The first part of the available data are the results produced by INLOOM. INLOOM persists its results in form of XML files. Each of the XML files, contains the results of the constraints applied to one student solution, as well as some meta data. The format used in these files, has changed since the publication of [14] and is now described by the following DTD:

```

<!ELEMENT TestResult (TestData , Results , ResultPoints)>
<!ELEMENT TestData (ExpertModel , TestModel)>

<!ELEMENT ExpertModel EMPTY>
<!ATTLIST ExpertModel id CDATA #REQUIRED>

<!ELEMENT TestModel EMPTY>
<!ATTLIST TestModel id CDATA #REQUIRED>

<!ELEMENT MetaModel EMPTY>
<!ATTLIST MetaModel type CDATA #REQUIRED>

<!ELEMENT MCSIdentifier EMPTY>
<!ATTLIST MCSIdentifier id CDATA #REQUIRED>

<!ELEMENT MCSVersion EMPTY>
<!ATTLIST MCSVersion value CDATA #REQUIRED>

<!ELEMENT Results (CResult+)>


<!ELEMENT CResult (
    ExpertObject , ExpertType , TestObject , TestType
    Rule , Category , Points , Msg
)>
<!ELEMENT ExpertObject (#PCDATA)>
<!ELEMENT ExpertType (#PCDATA)>
<!ELEMENT TestObject (#PCDATA)>
<!ELEMENT TestType (#PCDATA)>
<!ELEMENT Rule (#PCDATA)>
<!ELEMENT Category (#PCDATA)>
<!ELEMENT Points (#PCDATA)>
<!ELEMENT Msg (#PCDATA)>

<!ELEMENT ResultPoints (MaxPoints , TestPoints)>
<!ELEMENT MaxPoints (#PCDATA)>
<!ELEMENT TestPoints (#PCDATA)>

```

Listing 3.1 XML format, currently used to persist the results the INLOOM software generates.

The root of the XML files is the "TestResult". All meta data is stored in "TestData", while the individual constraint results are persisted as a list of "Result" under "Results".

Each such "Result" identifies the element, used during the constraint generation from the expert solution, in "ExpertObject" and "ExpertType". The matching element of the student solution is stored in "TestObject" and "TestType". The "Object" Part, holds the label or name of the used element. The "Type" Part stores the type of the element in the diagram. What types INLOOM is able to detect and grade, depends on the meta model used for the evaluation [14]. 

In the "TestData" branch, information about the evaluation is available. The "id" contained in "ExpertModel" references the expert solution the students work was compared to. "TestModel" identifies the evaluated student solution. "MetaModel", "MCSIdentifier" and "MCSVersion" contain versioning information about the created evaluation. These tags are interesting for making sure, that only evaluations, that were created under the same circumstances are compared.

The amount of test data sets will most probably not increase dramatically in the near future, so there is no reason, to reduce the result data in any way before using it for testing. All information required can automatically be extracted from the automatic evaluation result XML files.

3.2.2. Manually Evaluated Student Solutions

The second part of the test data are manual evaluations of student solutions. Thirty already graded pen-and-paper student solutions to exam tasks were digitally reproduced by [14], to evaluate them using INLOOM. Of these thirty solutions, ten are solutions to exam tasks of the summer term exams 2017, 2018 and 2019 respectively. Of the evaluations to this student solutions, only an analog version exists. It should be noted that, that only the evaluations of the summer term exam 2019 used the same uniform grading scheme for both the manual and automatic evaluation by default [14]. Every meaningful comparison of two evaluations requires the two to be based on the same grading scheme. This is the first formal limit (L) to the application.

L1 The system requires manual and automatic evaluations as input, that were created, using the same grading scheme.

The literature agrees, that the manual evaluations of these digitized student solutions are the best evaluations known for the specific solution and therefore, are the only measure of quality one can apply to INLOOM. It can be assumed that the automatic evaluations quality is sufficient, if it reaches the same result as the manual evaluation.

In order to compare these manual evaluations to the ones automatically generated by INLOOM however, they need to be digitized. Right now, the available manual evaluations consist of a number of handwritten annotations in the student solutions. The annotations are mostly checkmarks and points awarded for elements of the model. What feature the annotation references is indicated by its position in the student solution. Due to that format and the fact that the student solutions were stored as black and white scans, it is unlikely that the evaluation data can be automatically extracted. Therefore it is necessary to provide an evaluation digitization facility.

RQ4 The software must include a UI facility to digitize manually created evaluations of student solutions.

3.2.3. Data Driven Data Structure

There are some elements, automatic and manual evaluations obviously have in common. Others are more oblique and some transformation is required, before they can be assumed present in both.

Each of the evaluations, was made for exactly one *student solution* to solve exactly one *exercise*. Each evaluation can only ever be created by one *evaluator* and using one *expert solution* for reference.

A comparison of an automatic and manual evaluation can thus be identified by a key, that consists of the *student id*, the *exercise id* and the *expert solution id*. There is no point in comparing an automatic evaluation to a manual one, if they differ in one of these attributes. Such a comparison will later be called *TestDataSet*. Both kinds of evaluation need to contain these key attributes and rather obviously do.

For one student solution, there can exist multiple automatic and manual evaluations. Multiple manual evaluators or different versions of the *INLOOM software* can create evaluations for the same student solution. The *literature research showed*, that it can be interesting to inspect multiple manual evaluations of the same student solution, created by different evaluators. Each evaluator has his/her own style and preferences, which will be reflected in the evaluation. Enabling the application to store multiple manual corrections of the same student solution allows for easy calculation of previously described *clean scores* and any alike values.

RQ3.2 The software must be able to persist multiple evaluations for the same student solution.

Every expert solution is basically a collection of elements and features the student solution needs to contain, in order to be deemed correct. Both kinds of evaluation use an expert solution for reference. For each expected feature the automatic evaluation adds a *Result* to its *Results*-list.

The equivalent in the manual evaluation are the point annotations. Each of those awards points for a feature of the solution evaluated. A feature, that has to be part of an expert solution in order for it to be correct. Each of the point annotations can thus be transformed into a "result" of the manual evaluation.

From the information collected about both the automatic and manual evaluations, using a data driven approach, a data structure can be inferred. Of the entities relevant to the system, only two can exist without any dependencies. The *Evaluator* and the *Exercise*.

For each exercise there can exist multiple *ExpertSolutions*. As previously described, a *TestDataSet* must reference an *Exercise*, an *ExpertSolution* and a student, for it to be uniquely identifiable.

For each *TestDataSet*, or rather each combination of keys, that identify a *TestDataSet*, there can exist multiple automatic and manual Evaluations: *AutoEvals* and *ManEvals*. Both of those are *Evaluations* and generally follow the format introduced by INLOOMs result XML files. The only difference between the two is, that a manual evaluation must have been created by an *Evaluator*, while for the creation of an automatic evaluation the attributes *MCSIdentifier* and *MCSVersion* are required.

The following model results from combining all of these required features.

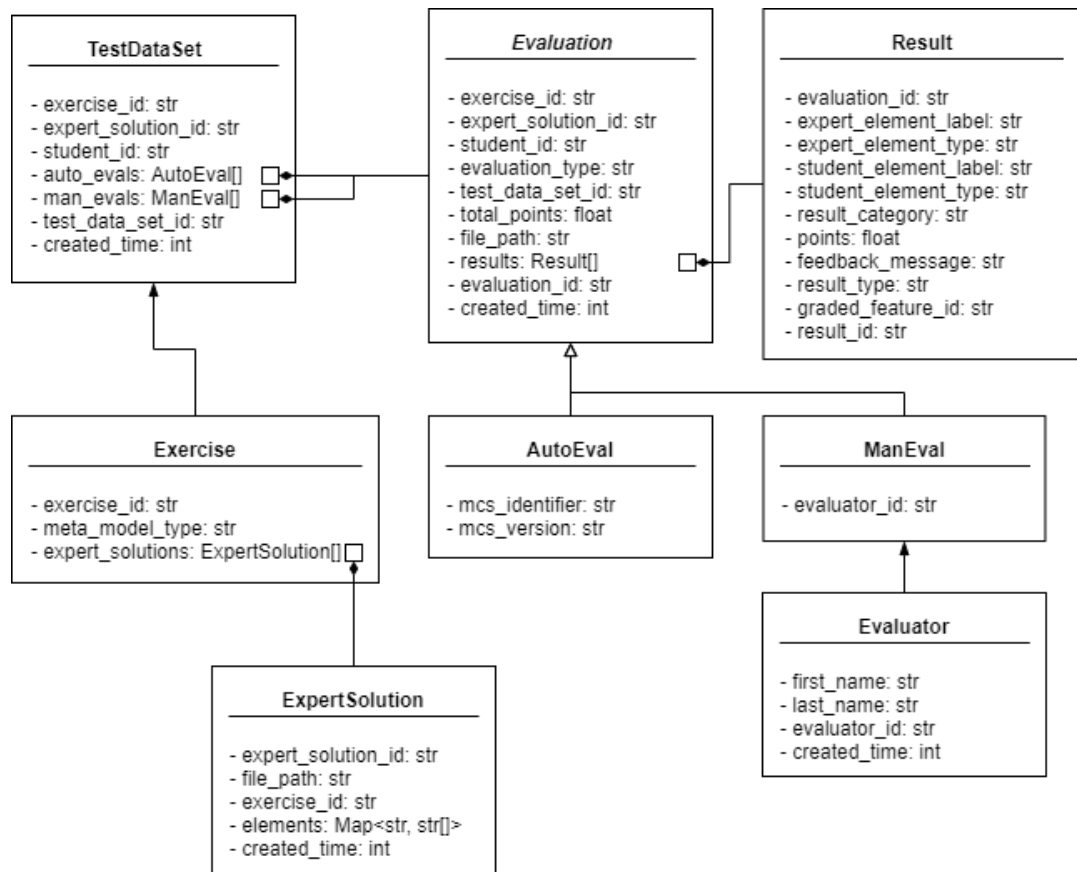


Figure 3.2. Data model, that results from the structural analysis of the available data.

3.2.4. Digitizing Manual Evaluations

The need to digitize the manual evaluations before being able to compare them automatically to the evaluations INLOOM produces, means a huge overhead for the testing process. Digitizing the manual evaluations is inevitable however. This is reflected by RQ4. The effort required for digitizing manual evaluations must be reduced as much as possible.

The process of extracting the data from the manual evaluations into a data structure like the one presented above, can roughly be separated into three steps and is the same for each manual evaluation.

1. Supplying identifying attributes (exercise, student, expert solution).
2. Supplying metadata (total points, evaluator).
3. Transforming point annotations into results and adding those results to the evaluation until the point total is accounted for.

Since many of the attributes, one needs to enter are repetitive and limited to a number of options, the UI facility must aim to provide selections rather than inputs for as many of them as possible. This can obviously be done for both the exercise and expert solution, since only a limited number of those are known to the application. The same holds true for the evaluator.

More interesting however, is easing the inputs required in the third step of the digitization, since it is the only one, that will have to be repeated multiple times. In

each repetition of step three, six attributes need to be entered, in order to register a new *Result* of the manual evaluation one is digitizing.

1. Expert Element Label
2. Expert Element Type
3. Student Element Label
4. Student Element Type
5. Result Category
6. Points

Of these only the student element label, student element type and points are not generic. The rest of the attributes is limited to a number of options. The options, except the the ones for the result category, are defined by the employed expert solution. The options valid for use as a result category are defined by INLOOM and are immutable. Adding a new *Result* can thus be reduced to entering three values and selecting the rest from predefined options.

3.3. KPI for Testing

The **literature research** showed quite clearly, that a *Grade Quotient Strategy* is the only real contender for any kind of statistic about the likeness of two evaluations. Since INLOOM is a constraint based system and its results need not be compared on a nominal scale, it is not deemed necessary, to calculate any of the Inter-Rater-Reliability (IRR) statistics encountered in the literature. Assuming a nominal categorization, would disregard a big amount of the data available.

Even when using a percentage difference for categories, artificially creating a nominal categorization, as was done in [30] and thus enabling their employment, these statistics do not add any significant new information. Also, their results are often less unintuitive than a simple quotient and are not deemed adequate, to provide a first glance impression of the current state of affairs.

IRR are designed to mitigate, the effects of habitual and random decision making, when validating evaluations, that employ categories with ill defined options. A big part of the randomness, in the typical use cases of IRR, results from these ill defined options. Since the evaluator is required to make a judgement call, it is possible, that not even the same evaluator will be able to repeat this decision, should **(.*|they)** be confronted with an alike case. This makes for grade equivalents in these evaluation, that can, by nature, not be vindicated after the fact, since they are made up of partial decision, that can not be repeated with any amount of certainty.

The use case of INLOOM is quite different. As previously described the XML results of the system, list all the constraint results, the final grade is made up of. Thus, each step that led to the final grade can be examined and it is most definitely possible to justify the final grade and to repeat the evaluation. For INLOOM one can conclude from its results, that the tool made an objectively wrong decision during its evaluation. This is not the case in IRRs typical use cases and in those, one would usually not be able to define *objectively wrong*.

Calculating IRR might still be interesting at a later point, since these statistics are employed in the literature and it might prove informative to be able to compare INLOOMs quality to the quality of other systems.

IRR might also become interesting for INLOOMs use case, when more than one manual evaluation should be the norm rather than the exception, at some point in time. Under that circumstances IRR will become useful to compare multiple manual evaluations to one another and create something alike a clean score automatically.

Since the software proposed here is required by RQ3.2, to be able to store multiple evaluations of the same student solution, it is opportune to add another requirement, that ensures the possibility to calculate additional KPI later on.

RQ3.3 The data collected by the proposed application must be easily accessible to programmatic analysis, one might wish to perform on it in the future.

Since this work aims to validate the quality of a constraint based system, with a uniform output format, no meta analysis of the quality validation is required, to extract additional information from a single value available. Instead of using meta statistics to extract more information from the *Grade Quotient*, due to the uniform output format of INLOOMs result XML files, a more detailed analysis can be performed. As described earlier, data about each constraints result is available. These results can be numerically analyzed in a number of ways and on different levels of detail. A grade quotient equivalent can be calculated and compared in each category and for each level of detail. Such an examination will enable the user of the proposed application not only to gain a quick overview over INLOOMs current performance, but also enable (.*|them) to quickly identify likely sources of error.

3.3.1. Comparison Detail Levels

A comparison of a manual and an automatic evaluation, collected and persisted as described, can happen on a number of different levels of detail, or rather with a number of different scopes. On each of these detail levels a number of categories has to be considered for comparison.

The final grade or exam score is the first and broadest category, two evaluations can be compared in. The final grade is a rating, made up from a number of more detailed ratings and does by itself not allow for an inspection of its composition. Each of the more detailed ratings is performed on, what in the literature is described as a *MMU*, a *minimal meaningful unit* [30]. In INLOOMs case, these MMU are identified by the *label* and *type* keys of the result. INLOOM matches an MMU in the expert solution to one, identified in the student solution and, using constraints to compare their features, grades their likeness. Each *result* listed in INLOOMs result XML files, represents the result of one such rating.

To evaluate the composition of the final grade, in a secondary analysis the results can be grouped by one of the available keys. By summing up the points of a group or evaluating the frequency of a certain result category within that group, one can calculate sub quotients of the grade quotient. These values specify the extent of agreement of two evaluations within a certain limit and are the most fine grained category for comparing two evaluations.

Since both the automatic and the manual evaluation use the same expert solution to compare the student solution to, the expert elements expected from the student solution are the same for both. Thus, it makes sense to use the expert elements as a key, for collecting data on, since they are common to both the manual and automatic evaluations.

Values collected for each evaluation, can be averaged over a group of evaluations. Such is interesting for both, combining multiple manual evaluations in a moderated evaluation and for inspecting a group of related evaluations. Collections of evaluations

for which an average KPI can be evaluated, represent the third level of detail one can inspect INLOOMs results on.

3.3.2. Categories of Comparison

The main category of comparison, of the two evaluations will, as previously stated, be a regular grade quotient. For each student solution, by the application identified using a key that consists of the student id, the exercise id and the expert solution id, such a quotient will be calculated as a result of the meta evaluation. Obviously such a grade quotient can be averaged over an arbitrary collection of evaluations later on.

Although more elaborate evaluations might prove interesting in the future, the application proposed here will only perform the most basic evaluation on the result level. In compliance with RQ3.3 it will later easily be possible to extend the application by any evaluations that are feasible on the available datasets.

In order to enable the user, to quickly identify sources of error, the application will group found results by element type and sum up the points awarded to each group (*points-per-element-type*). This way the user gets an easy grasp of where discrepancies in the grade quotient originate. The comparison of the points-per-element-type rating, just as the grade quotient, can be expressed by a quotient per element type.

Evaluations will be grouped by evaluator and expert solution id. The grade quotient and points-per-element-type rating will be average over all evaluations within a group. The evaluator and expert solution id are obvious candidates for a comparison on this level of detail.

As is expressed by the perceived need to employ IRR [30], evaluations by different reviewers can differ significantly. The application must take this into consideration and give the user the means to check for any negative effects using manual evaluations by multiple evaluators might have on the quality of the systems results. This is remedied by calculating an per evaluator average of the likeness ratings awarded to pairs of manual and automatic evaluations.

By calculating an average value for the evaluations of a specific expert solution, the user can examine INLOOMs results for error trends, that arise from the structure of a specific expert solution.

3.3.3. Visualizing Validation Results



Visualizations are unquestionably one of the most effective ways of presenting data. For visualizing the results of the validation, the proposed software is supposed to perform, several types of visualization are available for presenting results on each level of detail.

The Grade Quotient is the most comprehensive of the values calculated by the proposed software. Presenting it is the main task of the software's frontend. For that reason it should be presented to the user as often as possible. However, visualizing the Grade Quotient is only interesting in contexts, where there is more than one grade quotient available. Visualizing a single quotient would not gain anything. Individual Grade Quotients will therefore be presented to the user as they are.

Multiple grade quotients are available when grouping test data sets. This, as described in 3.3.2, will be done using the evaluator id and the expert solution id as keys. In these cases multiple grade quotients have to be presented.

The purpose of this kind of presentation is identifying unusual elements in the groups. Bar charts are suitable for this use case. Since quotients are being visualized, the range of possible values is limited and the individual values can easily be compared

without any further need for preprocessing the data. A visualization of the values in form of a bar chart will enable easy comparison between elements of the groups and thus help identify unusual elements.

Since the software is able to store multiple manual and automatic evaluations for one student solution, it might be interesting to evaluate the development of the quality of INLOOMs evaluations for one student solution, over a number of automatic evaluations. Since one `TestDataSet` only ever references evaluations, that were created using the same expert solution, each automatic evaluation available in the `TestDataSet` will represent one *generation* of INLOOM. Since the quality of the ratings should be prevented from deteriorating over time, the quality history should be made available to the user of the software. Using a line chart is appropriate for this task.

The last evaluation whose results remain to be presented is the comparison on the result level called *points-per-element-type*. Since the results of this validation are not very complex, but consist only of a number of points, awarded for a specific element type, by one of the compared evaluations, a simple bar chart is again suitable for their visualization. To enable easy comparison between the points awarded to a type by the automatic and manual evaluation, the points can be presented as adjacent bars. This will allow to compare the two evaluations to one another while also being able to assess the relative relevance of the inspected element type for the final score.

3.4. Summary

In this section the requirements collected are summarized. For each requirement, it is checked whether the proposed design is capable of satisfying it.

RQ1 The software must be able to automatically test the *quality* of the results INLOOM generates.

Result: The software will be able to automatically validate the quality of INLOOMs result files, employing a strategy, that is an extension of the *Grade Quotient Strategies* described in the literature.

RQ2 It must be possible to present the results of the performed validation in an easily comprehensible way, to quickly gain insight about the current state of affairs.

Result: The software will visualize the results of the validations it performs using bar charts and line charts as well as usual lists and detail views of the stored data.

RQ3 All tests must be performed using the data available.

RQ3.1 The software must employ a data structure, that is able to hold information collected on automatic and manual evaluations and enables comparing the two.

Result: Based on a data-driven approach, such a data structure was derived from the structures of the data available for quality testing.

RQ3.2 The software must be able to persist multiple evaluations for the same student solution.

Result: The data structure described in 3.2.3 is up to this requirement.

RQ3.3 The data collected by the proposed application must be easily accessible to programmatic analysis, one might wish to perform on it in the future.

Result: This requirement will be trivially tackled by storing all collected data in an independent database and providing all required interfacing functionalities.

RQ4 The software must include a UI facility to digitize manually created evaluations of student solutions.

Result: The software will incorporate an adequate facility, as was described in 3.2.4.

L1 The system requires manual and automatic evaluations as input, that were created, using the same grading scheme.



Bibliography



- [1] N. H. Ali, Z. Shukur, and S. Idris. "A Design of an Assessment System for UML Class Diagram". In: *2007 International Conference on Computational Science and its Applications (ICCSA 2007)*. 2007 International Conference on Computational Science and its Applications (ICCSA 2007). Aug. 2007, pp. 539–546.
- [2] Nilufar Baghaei, Antonija Mitrovic, and Warwick Irwin. "Supporting collaborative learning and problem-solving in a constraint-based CSCL environment for UML class diagrams". In: *International Journal of Computer-Supported Collaborative Learning* 2.2 (Sept. 1, 2007), pp. 159–190.
- [3] Philip-Daniel Beck et al. "COCLAC - Feedback Generation for Combined UML Class and Activity Diagram Modeling Tasks". In: (), p. 8.
- [4] Jan Philip Bernius and Bernd Bruegge. "Toward the Automatic Assessment of Text Exercises". In: (), p. 4.
- [5] Weiyi Bian, Omar Alam, and Jörg Kienzle. *Automated Grading of Class Diagrams*. Sept. 11, 2019.
- [6] Weiyi Bian, Omar Alam, and Jörg Kienzle. *Automated Grading of Class Diagrams*. Sept. 11, 2019.
- [7] Weiyi Bian, Omar Alam, and Jörg Kienzle. "Is automated grading of models effective? assessing automated grading of class diagrams". In: *Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems*. MODELS '20. New York, NY, USA: Association for Computing Machinery, Oct. 18, 2020, pp. 365–376.
- [8] Younes Boubekour, Gunter Mussbacher, and Shane McIntosh. "Automatic assessment of students' software models using a simple heuristic and machine learning". In: *Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*. MODELS '20. New York, NY, USA: Association for Computing Machinery, Oct. 16, 2020, pp. 1–10.
- [9] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (Apr. 1, 1960), pp. 37–46.

- [10] *Computing inter-rater reliability and its variance in the presence of high agreement - Gwet - 2008 - British Journal of Mathematical and Statistical Psychology - Wiley Online Library*. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/full/10.1348/000711006X126600?%20casa_token=5iF-i8Qxc_MAAAAA%3AyH8jCJPXFxwcVx1sMkF-Y--pUY_OV6JtT_mrpWzuBArpfyrWidFIVq0WoSIEEa%20Z1ftUiczdqLTMkAvR9 (visited on 12/29/2020).
- [11] B. Demuth and D. Weigel. "Web Based Software Modeling Exercises in Large-Scale Software Engineering Courses". In: *2009 22nd Conference on Software Engineering Education and Training* (2009).
- [12] James DeVaney et al. "Higher Ed Needs a Long-Term Plan for Virtual Learning". In: *Harvard Business Review* (May 5, 2020).
- [13] Joseph L. Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological Bulletin* 76.5 (1971), pp. 378–382.
- [14] Markus Hamann. "Automatic Feedback for UML Modeling Exercises as an Extension of INLOOP". TU Dresden, 2020.
- [15] Robert W. Hasker. "UMLGrader: an automated class diagram grader". In: *Journal of Computing Sciences in Colleges* 27.1 (Oct. 1, 2011), pp. 47–54.
- [16] Colin Higgins and Brett Bligh. "Formative computer based assessment in diagram based domains". In: *ACM SIGCSE Bulletin*. Vol. 38. June 26, 2006, pp. 98–102.
- [17] Colin Higgins, Pavlos Symeonidis, and Athanasios Tsintsifas. "The marking system for CourseMaster". In: *ACM Sigcse Bulletin*. Vol. 34. Sept. 1, 2002, pp. 46–50.
- [18] Gil Hoggarth and Mike Lockyer. "An automated student diagram assessment system". In: *ACM SIGCSE Bulletin* 30.3 (Aug. 1, 1998), pp. 122–124.
- [19] *INLOOP: interactive learning center for object-oriented programming*. Github. URL: <https://github.com/st-tu-dresden/inloop> (visited on 12/20/2020).
- [20] Stephan Krusche and Andreas Seitz. "ArTEMiS: An Automatic Assessment Management System for Interactive Learning". In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. SIGCSE '18: The 49th ACM Technical Symposium on Computer Science Education. Baltimore Maryland USA: ACM, Feb. 21, 2018, pp. 284–289.
- [21] Nguyen-Thanh Le. "A Constraint-based Assessment Approach for Free Form Design of Class Diagrams using UML". In: (Nov. 27, 2020).
- [22] David Powers. "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation". In: *Mach. Learn. Technol.* 2 (Jan. 1, 2008).
- [23] Ferran Prados et al. "An automatic correction tool that can learn". In: *Proceedings - Frontiers in Education Conference* (Oct. 1, 2011).
- [24] Joachim Schramm et al. "Teaching UML Skills to Novice Programmers Using a Sample Solution Based Intelligent Tutoring System". In: May 25, 2012.
- [25] William A. Scott. "Reliability of Content Analysis: The Case of Nominal Scale Coding". In: *The Public Opinion Quarterly* 19.3 (1955), pp. 321–325.
- [26] N. Smith, P. Thomas, and K. Waugh. "Automatic Grading of Free-Form Diagrams with Label Hypernymy". In: *2013 Learning and Teaching in Computing and Engineering*. 2013 Learning and Teaching in Computing and Engineering. Mar. 2013, pp. 136–142.
- [27] Rúben Sousa and José Leal. "A Structural Approach to Assess Graph-Based Exercises". In: June 18, 2015, pp. 182–193.

- [28] Michael Striewe and Michael Goedicke. "Automated checks on UML diagrams". In: *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*. ITiCSE '11. New York, NY, USA: Association for Computing Machinery, June 27, 2011, pp. 38–42.
- [29] *The COVID-19 pandemic has changed education forever. This is how*. World Economic Forum. URL: <https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/> (visited on 12/03/2020).
- [30] P. Thomas, N. Smith, and Kevin G. Waugh. *Automatic assessment of sequence diagrams*. undefined. 2008. URL: </paper/Automatic-assessment-of-sequence-diagrams-Thomas-Smith/%20ef570840cbb182d6e8f861ced321992e20b94f93> (visited on 11/27/2020).
- [31] Pete Thomas, Neil Smith, and Kevin Waugh. "Automatically assessing graph-based diagrams". In: *Learning Media and Technology* 33 (Sept. 1, 2008).
- [32] Christos Tselonis, John Sargeant, and Mary McGee Wood. "Diagram matching for human-computer collaborative assessment". In: (Jan. 1, 2005).
- [33] *Zahl der Studierenden erreicht im Wintersemester 2019/2020 neuen Höchststand*. Statistisches Bundesamt. URL: https://www.destatis.de/DE/Presse/Pressemitteilungen/2019/11/PD19_453_213.html (visited on 11/24/2020).

A. Weitere Latex-Dokumentation

Statement of authorship

I hereby certify that I have authored this Bachelor Thesis entitled *Quality Testing Concepts* independently and without undue assistance from third parties. No other than the resources and references indicated in this thesis have been used. I have marked both literal and accordingly adopted quotations as such. There were no additional persons involved in the intellectual preparation of the present thesis. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.

Dresden, 5th February 2021

Paul Erlenwein