University of Zurich[UZH]

# PaperValidator - Towards the Automated Validation of Statistics in Publications

Thesis September 7, 2016

**Manuel Rösch**
of Zurich ZH, Switzerland

Student-ID: 10-918-332
manuel.roesch@uzh.ch

Advisor: **Patrick de Boer**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

# PaperValidator - Towards the Automated Validation of Statistics in Publications

**Manuel Roesch**
Department of Informatics
University of Zurich
Zurich, Switzerland
manuel.roesch@uzh.ch

**Patrick de Boer**
Department of Informatics
University of Zurich
Zurich, Switzerland
pdeboer@ifi.uzh.ch

**Abraham Bernstein**
Department of Informatics
University of Zurich
Zurich, Switzerland
bernstein@ifi.uzh.ch

## ABSTRACT

The validity of statistics in scientific publications is crucial for accurate and reliable results. However, there are many publications, that are not acceptable in this regard. This work confronts this problem by proposing a tool that allows for the automated validation of statistics in publications, focusing mainly on statistical methods and their assumptions. The validation process is rule-based using partially crowd-sourced workers hired from the Amazon Mechanical Turk (MTurk) platform. The tool and the validation process, were successfully tested on 100 papers from the ACM Conference on Human Factors in Computing Systems (CHI) and further applied to examine the usage of statistics over the years in CHI papers.

### Author Keywords

Verification of Statistics; Automated Reviewing; Crowd Sourcing; Human Computation

## INTRODUCTION

The quality of statistics is questioned by several studies in various research fields. Altman [1], for example, summarized 13 studies from 1966 to 1996 about statistical quality in medical papers, with the conclusion that the statistics in one out of three papers contains flaws. Another example is the work of Veldkamp et al. [17], who examined the statistics in 697 articles published in six renowned psychology journals with the result that 20% of the papers contain severe statistical errors.

To get some insight into the present situation concerning statistics and publishing, we asked twelve publishers[1] of big journals how they validated statistics. The answers were similar: The journals mainly rely on peer-reviewing, and most of them have introduced a catalogue of statistical guidelines, which are supposed to be implemented by the authors and checked by the reviewers. Such a process can be insufficient, considering that there are numerous human biases in the interpretation of research results; e.g. the bias towards statistically significant results [13, 6]. Besides, as we learned from reading the journals' guidelines and asking statisticians from different research areas, there are widely varying opinions about how statistics should be reported.

[1]Nature, PLOS, AAAS, BMJ, The Royal Society, NAS, BioMed Central, JCI, Rockefeller, Hindawi, Feinstein and The Lancet

The publication at hand confronts this problem by proposing an open-source online tool called *PaperValidator*[2], which allows the automated validation of statistics in publications. Such a tool could encourage authors and reviewers to comply more strictly with important statistical rules by automatically validating them. The focus of the *PaperValidator* is mainly on the correct usage of statistical methods through verification if the author has checked the assumptions before applying a particular method. We decided to focus on that aspect of statistics because it is testable, influential, and a common source of error [8, 3].

The functionality of *PaperValidator* has been tested on 100 papers taken from different CHI conferences, and we could show that the verification process works with sufficiently high accuracy to justify its application for further analysis on the usage of statistics in CHI over the past years. This analysis showed that the usage and reporting of statistics have been improved, but they are, nevertheless, still at an insufficient level.

## RELATED WORK

Statistics are a crucial part of research, for this reason many publications from various academic fields are concerned with this topic. Most of the papers that we looked at either had statistical reviews of other authors and/or provided statistical guidelines to improve statistical reporting quality. Altman [1], for example, contains a meta-review of 13 papers reviewing a total of 1667 conference papers altogether; Kaptein et al. [9] reviewed the submitted papers presented at CHI 2010 and provided some best practice recommendations. The guideline papers often come directly from the journals; e.g. the guideline from the *American Physiological Society* [4] or from statistics institutes such as the one from *Oxford Centre for Statistics in Medicine* [10].

### Statsplorer

The reviews and guidelines certainly have a positive influence on how statistics are performed by pointing out errors and guiding authors through the analysis of data. However, the problem of bad statistics is not yet solved. An approach that takes the concept of statistical guidelines one step further is the work of Wacharamanotham et al. [18]. He provides a tool called *Statsplorer*, which guides an author step-by-step through the data analysis and hence ensures that all crucial

[2]https://github.com/manuelroesch/PaperValidator

steps are performed. *Statsplorer* thereby automatically generates some text snippets, which contain all the necessary information that should be reported in a publication. Although the usefulness of such a tool is clear, the limited functionality is probably insufficient for certain applications, and the strict guidance through the data analysis process may bias or constrain authors. Better in this regard is the *Statcheck* tool from Nuijten et al. [15], which works directly on the finished publications, as our *PaperValidator*, keeping all options open.

### Statcheck

The tool *Statcheck* is not a standalone application like our *PaperValidator* or the *Statsplorer* [18] presented in the previous section, but it is an extension to the *R programming language*[3] and does not have a graphical user interface.

The idea underlying *Statcheck* is the automatic validation of common statistical tests, such as f-tests, t-tests, Z-tests or chi-square tests. The tool realizes this in two steps. First, all the reported tests are extracted using text search with regular expressions. Second, the extracted tests are recalculated and validated for whether the stated p-values are in compliance with the recalculated p-values. It is thereby important that the tests are reported using the *American Psychological Association (APA)* reporting standard[4], because otherwise, the tool is not able to extract the test. [15]

Nuijten et al. [15] used the *Statcheck* package to determine the prevalence of statistical reporting errors in psychology from 1985 to 2013, and could determine that one in eight papers contains p-values that are inconsistent and may affect statistical conclusions.

Our approach is similar, at least for the first part: We also worked directly on the publications, extracting parts from the text using regular expressions. However, our focus was not on p-value recalculation, but on statistical methods and their assumptions, because we think that the *Statcheck* tool only encourages authors to copy their calculations more carefully from their statistics program to their reports so that the recalculation is correct. Meanwhile, our *PaperValidator* can improve the quality of statistics by motivating authors to check and report all required assumptions of the methods used.

### Statistical Methods and Assumptions

The work of Hoenkstra et al. [8] shows that assumptions are rarely reported, and there is an issue not only of a lack of reporting but also a lack of statistical knowledge, as have discovered during interview sessions with the participants of their study. This leads to severe validity problems in statistics, since many statistical methods require one or more assumptions to be met in order to produce correct and reliable results. Having unchecked or violated assumptions for a statistical method can seriously influence Type I and Type II errors, and it causes the problem of over- or underestimation of the inferential measures and effect sizes. [8]

For every assumption, there are several tests. Some of them are graphical; e.g. checking normality by making a normal quantile plot. Others are numerical; e.g. the Levene's test that checks whether the variances are distributed equally (homoscedasticity) [8]. The mapping of which assumptions belong to which statistical method, as well as how the assumptions of a particular method can be checked, is stated in many statistics books, such as the one from Field [7], who provides a step-by-step guide for most statistical methods and how they can be performed using the statistic software SPSS[5], including which assumptions must be checked and how.

[3]https://www.r-project.org/
[4]http://www.apastyle.org/

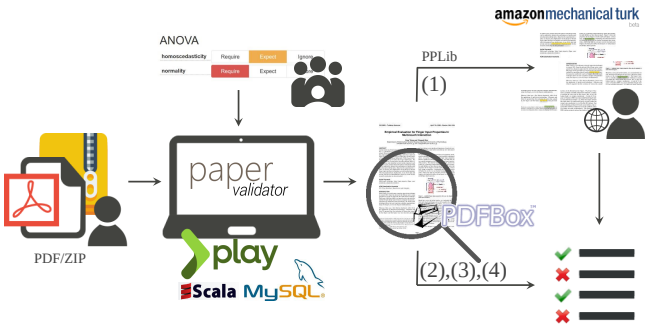


**Figure 1. An overview of the most important parts and functionalities of the *PaperValidator* tool.**

Although the necessity of checking assumptions exists and guidance on how to check them can be found commonly in literature, authors frequently neglect to check them. In the following section, the tool *PaperValidator*, our approach to face this problem, is presented.

## APPROACH

The following section consists of two parts. In the first subsection, the *PaperValidator* tool and its functionality are presented, followed by some details on how a central part of the tool, the method-assumption template, was determined.

### PaperValidator

The *PaperValidator* consists of different parts, that are based on different frameworks and libraries. Figure 1 presents an overview of this system and in the following; each part is described in more detail.

#### *Technical Details*

The *PaperValidator* system builds on the *Play! Framework*[6], which is a web framework facilitating the creation of web applications. The system is mainly written in *Scala*[7] and partially in *Java*[8]. As storage, we use a *MySQL*[9] database, which runs with our web application on a server at the University of Zurich with an *Intel(R) Xeon(R) CPU X5570 @ 2.93GHz* and 80 GB RAM. The PDF processing relies on *Apache PDFBox*[10], an open-source Java tool, which allows the extraction

[5]http://www.spss.com
[6]https://www.playframework.com/
[7] http://www.scala-lang.org/
[8]http://java.com
[9]https://www.mysql.com/
[10]https://pdfbox.apache.org/

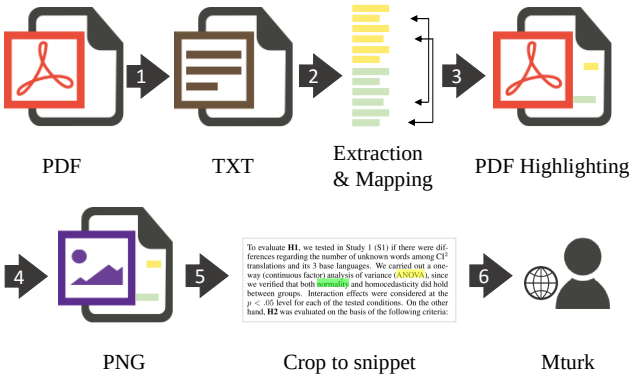


**Figure 2. The different steps of the *PaperValidator*'s method-assumption analysis.**

of content from PDF documents or the conversion of a PDF document into an image.

For the crowd-sourcing component, which is used during the statistics validation process, the system makes use of the *PPLib* [2], a library, that facilitates the creation of crowd-sourcing tasks. This library was used to send validation tasks to *Amazon Mechanical Turk*[11] *(Mturk)*, a popular crowd-sourcing platform.

### *Functionality Overview*

The target users of *PaperValidator* are authors, reviewers, as well as conference chairs. For each of these users, the tool provides a different functionality. In addition, there are also the Mturk crowd workers who access the tool. All these different users are presented in Figure 1, where authors/reviewers are on the left, Mturk workers are on the right, and conference chairs are at the top.

### *Functionality for Authors*

As can be seen in Figure 1 on the left, an author starts the process by uploading his publication to *PaperValidator* using the provided upload form. In doing so, he has to select the conference to which he wants to upload the paper. The conference was previously created by a conference chair (Figure 1 at the top), which will be explained later. It is worth mentioning that the system supports the upload of a single PDF file as well as the upload of multiple PDF documents compressed in a ZIP file.

After the upload, the system analyses the paper using validation algorithms partially based on crowd workers. The *PaperValidator* performs an analysis consisting of four different parts: (1) There is the method-assumption part, which validates methods and assumptions; (2) the *Statchecker* part, which implements the functionality as provided by the *Statchecker* tool [15] as presented in the Related Work section; (3) a part that validates some basic statistical rules; and (4) a part that performs some basic layout inspection. The parts are marked with brackets and numbers in Figure 1.

[11] https://www.mturk.com

To evaluate **H1**, we tested in Study 1 (S1) if there were differences regarding the number of unknown words among $CI^2$ translations and its 3 base languages. We carried out a one-way (continuous factor) analysis of variance (ANOVA), since we verified that both normality and homocedasticity did hold between groups. Interaction effects were considered at the $p < .05$ level for each of the tested conditions. On the other hand, **H2** was evaluated on the basis of the following criteria:

**Figure 3. Exemplary snippet of a method-assumption pair as generated by the *PaperValidator*.**

Part (1), as summarized in Figure 2, is the most central and relevant part in this work. For this method-assumption part, the text is first extracted from the uploaded PDF and further processed using regular expressions search for a predetermined set of methods, assumptions, and their synonyms. After having determined all the methods and assumptions in the text, a matching algorithm determines, which methods and assumptions fit together by using a predefined list containing the method-assumption allocation.

The next step is the creation of method-assumption snippets, which are later sent to Mturk for validation. The creation of such snippets is necessary because the copyrights of the papers often prohibit papers be distributed as a whole. The creation of a snippet works as follows: First, a method-assumption pair, which has been extracted previously, is annotated in a copy of the uploaded PDF file. The method is annotated in yellow, the assumption in green. In the next step, the PDF file is converted to a PNG image and cropped so that both the method and assumption are visible. In case they are on different pages, the pages are put together into one image, and the page break is indicated by a page break symbol. An example of such a snippet is shown in Figure 3.

The last step in part (1) of the analysis is the validation of the snippet using crowd-sourcing. For this, a question is generated on Mturk, as shown in Figure 4. The Mturk worker (Mturker) then decides whether the method-assumption pair is related, and if the author has checked the assumption before applying the method. Thereby, we do not only ask one Mturker, but several of them with the stopping rule that the final answer must win with at least three more votes than the second most voted answer. To increase the reliability of the answers, we also introduced two further measures. First, we let the Mturker report their thoughts during the decision-making process and write them down. This should encourage them to think more deeply and elaborately. Second, we let them report their confidence from one to seven on a slider (see Figure 4 at the bottom) and eliminate all answers with a confidence lower than five from further analysis. The threshold of five was determined empirically by a couple of initial test runs and is also confirmed by the work of Lessel et al. [12], who also uses a seven-point confidence scale with a threshold of 5.

Part (2) of the analysis, the *Statchecker* part, first converts the PDF to text and performs a validation equivalent to the functionality of the *Statchecker* R package presented in the


**Figure 4. The interface as seen by a crowd worker on Mturk when solving a method-assumption classification task generated by the *PaperValidator*.**

relation work section. This means that *PaperValidator* extracts common statistical tests like f-tests, t-tests, Z-tests or chi-square tests reported in the APA format using text search with regular expression from the converted text. All the extracted tests are, in the next step, recalculated and compared with the reported p-values. If such p-values are not in compliance with the recalculated p-values, they are saved as an error in the database.

Part (3) deals with basic statistical rules as reported in [16]. Here, the first step is once again the conversion of the PDF into text. Next, *PaperValidator* performs a text search using regular expressions to answer the following questions:

- Is the sample size stated in the text?
- Is there any incorrect statistical terminology in the text?
- Does the PDF contain any p-values? Are they in the correct range and precision?
- Is there a mean without variance reported in the text?
- Has the author performed a statistical test without stating effect size or power of the test?

In part (4), a simple layout analysis is performed. For that, the PDF is converted into a PNG image, which is analyzed by *PaperValidator* considering the following questions:

- Does the paper have a certain distance between content and border so that it can be printed properly?

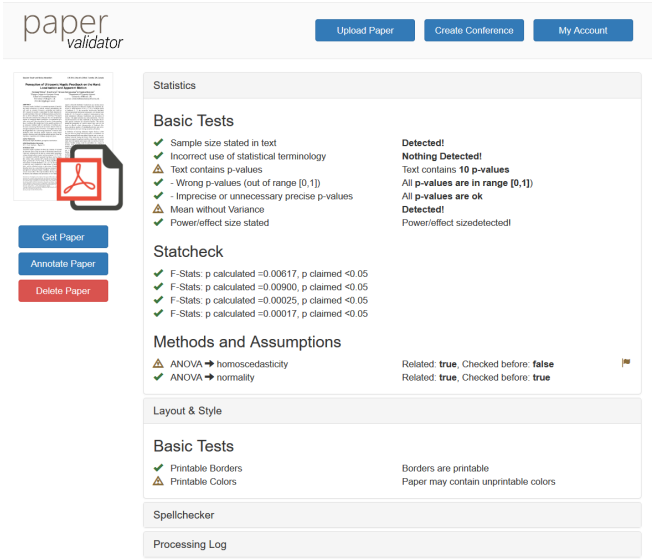


**Figure 5. The interface of a paper overview page in *PaperValidator* containing the analysis results, logging information and annotation options.**

- Are there any colors used in the paper, which are difficult to read when printed in gray scale?

Notice that the analysis in part (4) is not directly related to statistics but indirectly; e.g. diagrams presented in unreadable colors makes it challenging for a reader to follow the reported explanations. Besides, part (4) is also a proof of concept, that the *PaperValidator* can be easily extended so that not only the contents but also the layout can be checked.

Having finished the paper analysis parts (1)-(4), the author, who has uploaded the PDF, will be notified by an email containing a hyperlink to the paper analysis result overview page as shown in Figure 5. For each of the four analysis parts, the results are listed and depending on the result, a warning or an error is generated.

Furthermore, the analysis results overview page also includes a spell checker, which can be used besides spell checking, to verify the conversion process from PDF to text. If there are exceptional spelling mistakes listed, which are not present in the initial PDF file, there was an error in the conversion process and the analysis results are therefore not reliable.

Another source of information when an error happens during the PDF processing is the processing log, which also can be found on the result overview page. This log shows all the important events and reports all errors thrown by the tool. There is also a summary of all method-assumption snippets and their corresponding Mturk answers.

The result overview page also allows the download of the analyzed PDF in two versions; one is the blank version, which is equal to the one which was uploaded to the system, and the other is an annotated version in which all the findings are highlighted. The most dominant highlighting, thereby, is applied to methods with missing assumption.

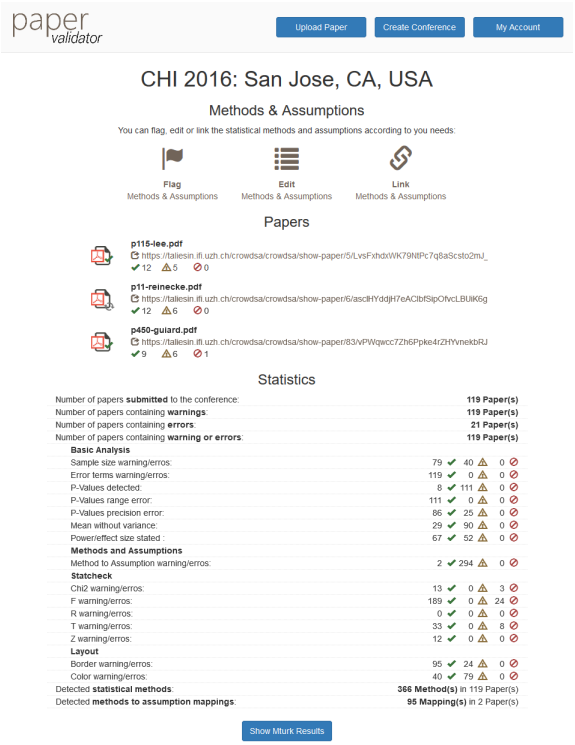


**Figure 6. The interface of the conference overview page in *PaperValidator* showing a conference where three papers have been uploaded so far.**

### *Functionality for a Conference Chair*

The main functionality for a conference chair is related to the creation and administration of a conference. *PaperValidator* provides for this purpose several interfaces such as a conference creation form, conference settings pages as well as a conference overview page. In the following paragraphs, each of these interfaces is explained in more detail.

First, the conference creation form, allows the chair to create a conference by choosing a name for the conference and selecting a method-assumption template, which later builds the base for the method-assumption validation process. It is worth mentioning that this template is only the base and it is freely adaptable later. Having created the conference, the creator gets an email with a hyperlink to the conference overview page.

This conference overview page, as shown in Figure 6, consists of three parts. On the top, there are three buttons relating to different conference settings concerning the method-assumption validation. The next part, in the middle, lists all the papers, which have been uploaded to the conference so far. Besides the processing status of the uploaded papers, the list also shows how many warnings and errors have been found for each paper. Moreover, by clicking on a paper, a conference chair can get to the paper results overview page and use all its functionality, as has already been presented in the Functionality for Authors Section. The bottom of the conference overview page provides some statistics about all the uploaded papers and the findings of its validation process.

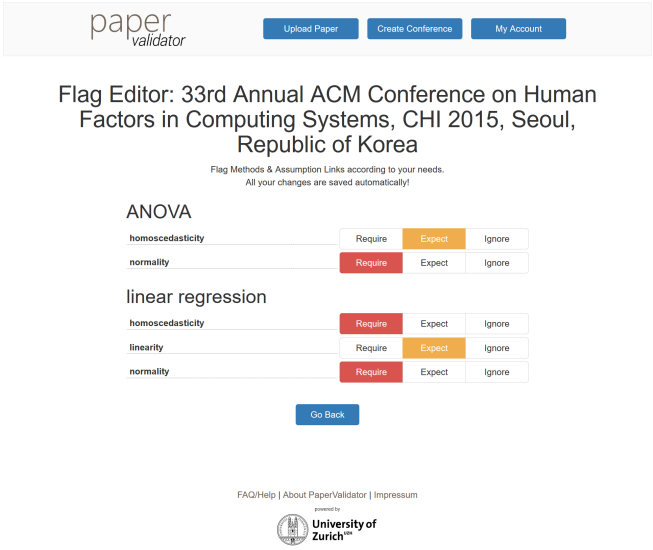


**Figure 7. *PaperValidator*'s conference setting interface where a conference chair can flag the assumption checking for certain methods according to their needs.**

The three method-assumption validation settings pages, which are at the top of the conference overview page, have the following meanings. The first relates to the interface for inserting and editing methods, assumptions and their synonyms; the second is for linking methods with their associated assumptions, and the last is for flagging the linked method and assumptions as shown in Figure 7. With this flagging option, a conference chair can assign an importance to each of the method-assumption linking. So for example, if a chair flags ANOVA and its assumption of a normal distribution as required, every paper will show an error when ANOVA is used without checking for a normal distribution first. The flagging also directly influences the highlighting color on the paper overview page.

### *Feedback from Statisticians*

To ensure that the functionality of our tool is accessible to its target users and has their acceptance, we performed interviews with six statistics professors from the University of Zurich. The semi-structured interviews brought about some valuable input for enhancing the interface and functionality, especially in the analysis part, as well as some ideas for future work.

The most interesting insight from these interviews was the fact that even though the professors have a similar broad knowledge of statistics, they do not have a consensus on how statistics should be reported, and therefore, they also estimated the usefulness of *PaperValidator* differently. While some of them think the *PaperValidator* is exactly what the research community needs, others believe that the tool is too rigid and authors will never report the statistics in as much detail as demanded. One interviewee brought up a possible reason why there is such a difference in opinions. He sees a problem in different research cultures and in different fields. In medical science, for example, the statistical results are reported in more detail

| | Normality | Homoscedacity | Linearity | Multicollinearity | Variance-covariance |
|---|---|---|---|---|---|
| t-test | x | x | | | |
| ANOVA | x | x | | | |
| MANOVA | x | x | x | x | x |
| ANCOVA | x | x | | | |
| Linear regression | x | x | x | | |

**Table 1. The method-assumption mapping as used in the method-assumption analysis of *PaperValidator*.**

because the validity of the results is much more important, considering that they can affect human life.

### Method and Assumption Mapping

The analysis of statistical methods and assumption is the central part of *PaperValidator*. Because this part relies on an initial set of statistical methods, assumption and their correct mapping, we wanted to ensure that we had an elaborated set, which is widely accepted by statisticians.

In order to create such a set, we asked a statistician from the University of Zurich to create an initial set of all possible methods (and synonyms) and their associated assumptions (and synonyms). This set was then reviewed by another statistician, and the resulting set was presented in an interview session to six statistics professors of the University of Zurich. Here again, there was no consensus, but we tried to create a mapping which optimally satisfies all the concerns of the different scholars. The interview sessions also brought up some ideas for assumption synonyms, which were also added to the mapping.

The resulting set was, in the next step, compared to the occurrences of statistical methods in CHI conferences from 1989 to 2016. For that we downloaded all the conference papers and searched for methods using regular expressions. The methods used as search terms come from a method glossary provided by Leeper [11]. The results from this analysis show that our set covers three of the top five most occurring methods over the years and the other two from the top five do not have any testable assumption to our best knowledge. The final set resulting from this analysis was again double-checked using [7] and [14] before it was finally implemented as a template in *PaperValidator*. The final mapping (without synonyms) is shown in Table 1.

## EVALUTATION

With *PaperValidator*, we have built a system which is able to extract automatically certain aspects of statistics and analyze them. To prove the validity of this analysis process, we performed a test run using CHI papers. The focus of this validation was on the method-assumption validation part of the system, which is crowd-sourcing based.

The CHI papers had been chosen for the evaluation because CHI is multidisciplinary and this may lead to a variety of statistics reporting styles. Such a variety is ideal for a reliable evaluation, which is meaningful for different research fields. The CHI papers have been extracted from the *dblp computer science bibliography*[12] using *iMacros*[13]. There were, in total, 5132 papers extracted from the conferences from 1989 to 2016.

### Experimental Setup

For the evaluation, a stratified sample of 100 papers out of the 5132 CHI papers was processed by the *PaperValidator*, using the method-assumption mapping as shown in Table 1 and the result of the analysis was compared with the ground truth, which we generated manually by annotating the PDFs and extracting the methods and method-assumption pairs. In the following, sampling, ground truth generation and the measurements are explained in more detail.

### Sampling

The required sample size was calculated using a standard formula for population proportion estimation [19] aiming at the target statistical power of 90 %. The outcome of this calculation revealed that a sample size of 100 papers is sufficient for our requirements.

Having determined the required sample size, we had to select 100 papers from the total 5132 CHI papers using a stratified sampling technique. We decided to use a stratified sample instead of a random sample to ensure that the papers in the sample contained statistics. This was important considering that around 30% of all CHI papers do not use statistics according to our glossary method, which is explained later. We decided to use only papers with at least one occurrence of a p-value in our sample because all statistical methods (see Table 1) implemented in *PaperValidator* report p-values.

Furthermore, we introduced two sampling categories: papers which contain many statistical terms and papers using few statistical terms, similar to what was done in the work of Dugan et al. [5]. Such a categorization was required because of reporting differences between papers. Many authors use in their papers only a few statistical terms to describe their statistics, while a few use them excessively which leads to a long-tail distribution of method-assumption pairs per paper. A random sampling would lead to a bias, since most of the paper would be selected from the long tail; therefore, in order to account for that, we introduced these two categories.

For the assignment of a particular paper to one of these two categories we used the glossary method, what means that a glossary with statistical terms was created and used to count the number of statistical terms contained in each paper. We created the glossary combining the statistics glossary from five

[12] http://dblp.uni-trier.de/db/conf/chi/

[13] http://imacros.net/

different sources[14,15,16,17,18] and removed common words from the English language[19]. This glossary method was applied to previously selected papers containing at least one p-value, and the papers were ranked according to the total number of statistical terms from our glossary that were found in the paper. The higher ranks thereby belong to the first category and the lower ranks to the second and therefore our final selection was the first and the last 50 papers in the ranking so that we had sampled, equally from each of the two categories, enough papers to reach our target sample size of 100 papers.

### Generation of the Ground Truth

For the evaluation of the *PaperValidator's* statistical analysis, a ground truth was required representing the analysis results as performed by experts. This ground truth was necessary for a comparison between the *PaperValidator* results and expert results. For the generation of this ground truth, we had two approaches. The first was the usage of a freelance expert portal, but it delivered insufficient results. Therefore, a second approach was used, based on manual reviewing, verified by an inter-rater agreement analysis. More about the two approaches can be found later in this section.

The creation of the ground truth was realized in both approaches using two steps. First, all statistical methods and assumption occurring in a particular paper were highlighted directly in a modified version of the original PDF. This modification includes the underlining of all statistical terms (as they occurred in the glossary from the sampling) and the assignment and annotation of a unique ID, as used by the *PaperValidator* system, to each of the underlined terms. The second step was the extraction of all methods and method-assumption pairs to an external text file, extracting in each case the method name plus method ID and the assumption name and assumption ID. These created text files could then be used directly for comparison with the *PaperValidator* results.

#### *Crowd Expert Approach*

For our first approach we wanted to hire experts on the freelance portal Upwork[20] and therefore we created a movie, which introduced our project and explained the task, including the PDF highlighting and method-assumption extraction as explained above. We hired three freelance experts in statistics with a high job success rate, top rating and a comprehensive job history, paying 2.5$ per paper page. The results from these three experts were not sufficient and differed strongly from the expected results generated by us. For that reason, we decided to use a manual reviewing approach.

#### *Manual Reviewing Approach*

In the manual reviewing approach, we performed the task of extracting statistics from papers ourselves. To ensure the validity and unbiasedness of this approach, we processed some of the papers up to three times by different people from the *PaperValidator* project so that we could conduct an inter-rater agreement analysis. In this analysis, we compared the different paper review outcomes to each other with the conclusion that the agreement between the reviewers is sufficiently high (Cohen's Kappa > 0.8). This means that our ground truth is accepted by multiple researchers in the field of statistics and computer science and is therefore reliable.

[14] http://www.stat.berkeley.edu/ stark/SticiGui/Text/gloss.htm
[15] https://www.st-andrews.ac.uk/psychology/current/statisticsglossary/
[16] http://www.stats.gla.ac.uk/steps/glossary/alphabet.html
[17] http://isi.cbs.nl/glossary/bloken00.htm
[18] https://en.wikipedia.org/wiki/Glossary_of_probability_and_statistics
[19] http://www.wordfrequency.info/free.asp
[20] https://www.upwork.com/

| Paper ID | Method | M-ID | Assumption | A-ID | GT | PV |
|---|---|---|---|---|---|---|
| 7 | ANOVA | 4:4236 | Normality | 4:4278 | 1 | 1 |
| 7 | ANOVA | 6:6742 | Normality | 5:7623 | 0 | 1 |
| 8 | t-test | 3:4823 | Homoscedasticity | 3:3457 | 1 | 0 |
| 9 | ANOVA | 2:8642 | Normality | 5:4263 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... |

**Table 2. Example of the data format as used for the comparison between ground truth and *PaperValidator* results.**

### Measurements

The processing of 100 sampled papers through *PaperValidator* using Mturk happened in different sessions conducted during business hours of working days, Eastern Time (EST). For the experiment, only US workers with more than 4000 solved and approved tasks and less than 4% rejected tasks were hired. The tasks were randomized and could only be answered once per crowd worker. The reward for each task was 40 cents, which resulted in a total cost of 214$ for the analysis of 100 papers. The task, as posted to the crowd workers on Mturk, is shown in Figure 4.

The data produced by the *PaperValidator* had the same format as the text files of the ground truth containing method name and ID, as well as assumption name and ID of each detected pair. The IDs used for methods, as well as for assumptions were unique using a combination of the page number and its relative text position on the page, representing the position of a particular word. In Table 2, there are examples of such IDs; in the row where ANOVA has the ID 4:4236, which means that it can be found on page 4 at text position 4236.

The consistent use of these IDs in the *PaperValidator* data, as well as in the ground truth data, made it possible to automatically combine the data sets using the unique IDs of methods and assumptions, which results in the data set as shown in 2. The column GT represents the ground truth data and column PV, the *PaperValidator* data using 1 for a valid method-assumption pair (the author has checked the assumption before applying the method) and 0 for an invalid method-assumption pair (the assumption does not belong to the method or the author did not check the assumption before applying the method). Besides this table, another one was generated, similar to the Table 2, containing only the detected methods without assumptions.

## RESULTS

The 100 CHI papers of the stratified sample lead to 197 data entries of possible method-assumption pairs in the format as shown in Table 2. Besides, we were also able to generate a table with 303 entries of extracted methods. These two tables

| | | *PaperValidator* | |
|---|---|---|---|
| | | positively classified | negatively classified |
| ground truth | positively classified | 257 | 4 |
| | negatively classified | 42 | 0 |

**Table 3. Confusion matrix showing the *PaperValidator's* method extraction results.**

| | | *PaperValidator* | |
|---|---|---|---|
| | | positively classified | negatively classified |
| ground truth | positively classified | 73 | 17 |
| | negatively classified | 13 | 93 |

**Table 4. Confusion matrix showing the *PaperValidator's* results on identifying correct method-assumption pairs using crowd workers.**

were used to answer two important questions concerning the method-assumption analysis of the *PaperValidator* tool:

1. How does the *PaperValidator* perform in extracting statistical methods from a text?
2. How is the *PaperValidator's* performance on extracting statistical assumptions, mapping them to the extracted methods and identifying correct method-assumption pairs using crowd workers on Mturk?

The tool was assessed comparing the ground truth data with the *PaperValidator* data (In Table 2 column GT and PV) using four performance measures for binary classifiers:

1. Precision: The fraction of correctly positively classified elements out of all positively classified elements.
2. Recall: The fraction of correctly positively classified elements out of all elements that should be positively classified.
3. Accuracy: The fraction of correctly classified elements out of all elements.
4. F1 score: Measure of a test's accuracy calculating the harmonic mean of precision and recall.

The results of the method extraction are shown as confusion matrix in Table 3, leading to a precision of 85.9%. A recall of 98.4%, as well as an accuracy of 84.8%, with a relatively high F1 score of 0.91. The high recall value shows that the system is able to extract correctly almost all the occurring statistical methods in the text, which was expected since statistical methods are usually reported using a fixed terminology and their extraction using a text search and regular expression is straightforward.

Table 4 shows the confusion matrix of the method-assumption classification. With a precision of 84.9%, a recall of 81.1% and an accuracy of 84.6% under the agreeable F1 score of 0.83, demonstrates a high reliability of the *PaperValidator* system. 84.9% of the method-assumption pairs, which were classified by the crowd worker as valid pair (the author has checked the assumption before applying the method), were actually valid and 81.1% of the total 90 correct method-assumption pairs (as revealed by the ground-truth) could be correctly detected.

Besides the comparison of our system to the ground truth, it was also compared to two baseline algorithms called PV-Auto1 and PV-Auto2. The first one does not use any heuristics at all, considering every extracted method-assumption pair from the text as valid. The second algorithm uses the heuristic that the assumption checking is usually reported closely to where the method is reported and therefore, all pairs occurring on the same page are considered as valid.

The analysis of PV-Auto1 resulted in a recall of 100%, which means that all the valid method-assumption pair could be detected by the system. However, with a precision of 46.4% PV-Auto1 was not appropriate for a comparison of a crowd based system to a fully algorithmic system. For such a comparison, we used PV-Auto2 using a basic heuristic.

With PV-Auto2 we reached a precision of 66.7%, a recall of 73.4% and an accuracy of 70.9% under an F1 score of 0.7. It can be seen that all the measures are considerably lower than the measures when using crowd workers. Precision falls by 18.2%, recall by 7.8% and accuracy by 13.8%, which clearly reveals the need and usefulness of crowd workers to our system.

To conclude the results, we would like to add two critical remarks concerning the data and data analysis. First, we had to exclude two papers from the analysis with an exceptionally high error rate. These papers are a special case because both do not apply the stated statistical method but discuss them on a meta-level. Our system does not account for that but could be extended in future work to cover this case by including a corresponding question in the crowd worker task.

Second, it was also noted that, despite stratified sampling, we faced a long-tail distribution of method-assumption pairs with only 12 of 100 papers reporting any assumptions. This circumstance indicates that the assumption checking in CHI is uncommon. In the Discussion section, this claim is further examined and assessed.

## DISCUSSION

The results of the 100 processed CHI papers revealed that the validation process of *PaperValidator* works reliably despite the non-expert human component. Moreover, the analysis indicated that the assumption checking is rather sparse considering most of these 100 papers do not have assumption checks for the methods used. This raises the question about the validity of the statistical results in several papers because it can only be assumed the author has performed the necessary assumption checks. To gain a further indication into this problem, we analyzed the use of methods and assumption in CHI from 1989 to 2016.

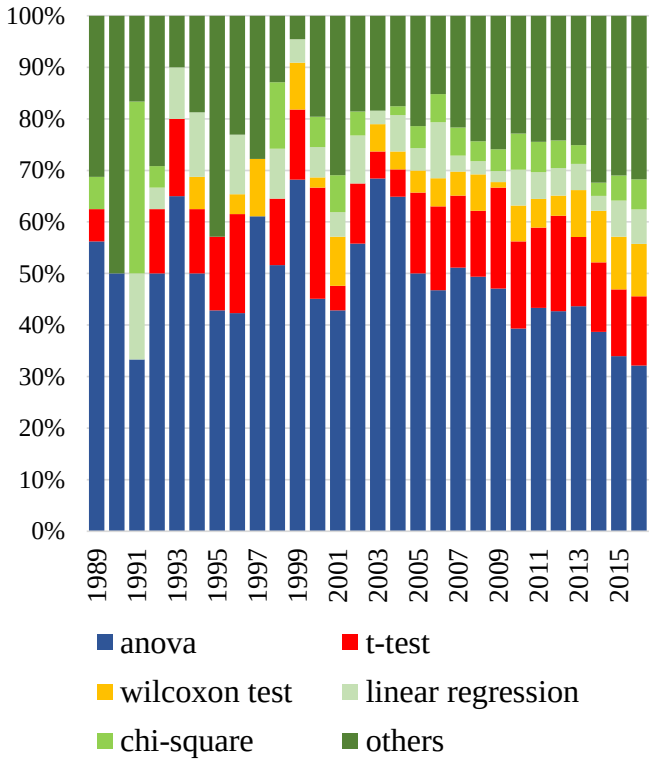


**Figure 8. Development of the usage of statistical methods in CHI from 1989 to 2016.**

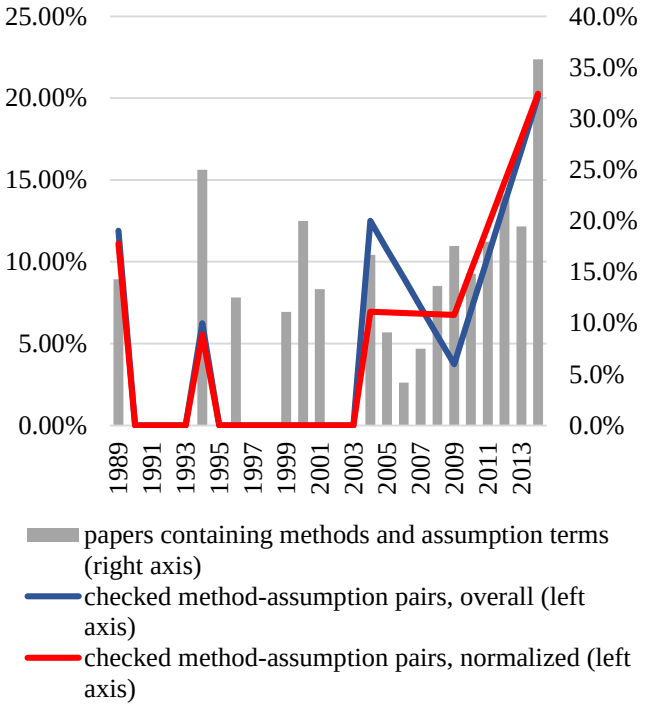


**Figure 9. The development of the share of checked method-assumptions pairs over the years in CHI.**

### Methods and Assumptions in CHI over the years

In a first step, we analyzed the occurrence of statistical methods in all the downloaded CHI papers using the method glossary of Leeper [11]. The result demonstrates that the portion of papers containing statistical methods has roughly tripled over the last two decades. However, method term occurrence alone is a weak indication that statistical reporting has improved over the years and therefore we performed a deeper analysis using the *PaperValidator* tool.

For this analysis, we considered only two methods; ANOVA and t-test since, as we can clearly see in Figure 8 which shows the shares of the top five statistical methods of CHI, they are dominant in almost every year and, therefore, most suitable to sample from. So in a first step we randomly sampled papers (containing either ANOVA or t-test) from different years starting with 1989, and using a five-year interval with a sample size of 30 papers for each year. This sampling and interval was necessary to save costs.

Next, the sampled papers were processed by the *PaperValidator*, similar as we did with the 100 CHI papers from the previous analysis, so that we had two data sets at the end containing all extracted methods and all the processed method-assumption pairs. These two data sets could then be used to examine the development of method-assumption reporting over the years. The result of this analysis is shown in Figure 9. It should be noted that for the years 1989,1994 and 1999, the number of papers sampled is less than 30 because there were not enough papers available. This leads to a reduced power of the analysis during those years. Furthermore, we added data points at 0% for years without any method-assumption pairs.

It can be seen in Figure 9 that there appears to be a slight improvement over time starting from 1999, which is the first year where we reached a sample size of 30 papers. Both, the share of checked method-assumptions pairs over all possible pairs, as well as the normalized share of method-assumption pairs (the same influence for every paper by normalizing the number of method-assumption pair occurrences) have an upwards tendency. However, this tendency is far from steady, which is probably due to our reduced sample size and the limited number of papers per conference every year. To indicate the possible steadiness in a bigger sample, the share of papers, which contain both statistical methods and assumption terms (necessary to build a valid method-assumption pair), was added as grey bars to the chart in Figure 9. It should also be noted that this analysis was performed using all 5132 papers but the results are not reliable because the bars indicate only the possibility of a valid pair. Even though this approach is not reliable, it indicates a relatively stable growth in the last decade and, therefore, for future work, this phenomenon will be analyzed on a bigger scale.

Despite this growth, there are still many unchecked assumptions and a big potential for improvement of statistical reporting in CHI. And not only the unchecked assumptions are a problem; the *Statcheck* part of the *PaperValidator* revealed that 8.7% of the 5132 CHI papers contain errors and the other

analysis parts brought up more flaws, like the lack of effect size respectively statistical power reporting in 15.3% of the papers, or the absence or vagueness of sample size reporting in 3.2% of the papers.

## LIMITATIONS AND FUTURE WORK

During the interview session we were confronted with many different opinions about which assumption can and should be checked given a statistical method. As the core functionality of *PaperValidator* and its accuracy builds on a solid method-assumption mapping a reassessment and extension of the existing mapping is important in future work. Besides, since we faced different opinions about the reporting style in different disciplines during the interviews, we need to evaluate *PaperValidator* in other research fields in order to confirm its accuracy.

Furthermore, the cost per paper can be minimized by implementing other heuristics to the crowd process so that more pairs are pruned without losing too much precision. We could, for example, create a library of phrases which express unambiguously that an author has checked the assumptions for a method and use them to confirm valid method-assumption pairs directly without using Mturk.

In addition, *PaperValidator* was built as an all-purpose research paper validation tool, which is open for many extensions. Further statistical analysis tests could be implemented alongside other tests concerning layout, grammar or content. This would lead to a tool as wished by one of our interviewees: *"It would be great to have a tool, which automatically checks all the pre-requisites for a certain conference on a paper and perform a basic review"*.

## CONCLUSION

With *PaperValidator*, we have created a system which can extract and validate certain parts of statistics reported in a publication. Pivotal in this validation are statistical methods and if their assumption had been checked before the method was applied. We proved the system to be accurately working using a stratified sample of 100 CHI papers and furthermore, we used the system to examine the development of method-assumption reporting from 1989 to 2016 with the outcome that it slightly increases but there is still lot of room for improvement.

We believe that our results and the proposed system could help improve the statistical reporting in CHI papers and increase, through enhanced assumption checking, the quality and validity of the reported results.


## ACKNOWLEDGEMENTS

A special thanks goes to Patrick de Boer for his great advice and patient support, as well as the numerous lines of code he added to the project. A big thank you also goes to Michael Feldman for his valuable input, and Mattia Amato for his preliminary work and analysis.

Further, we would like to show our gratitude to all the professors who participated in our interview sessions, as well as to the journal staff who took the time to answer our questions about their publication processes.


## REFERENCES


1. Douglas G Altman. 1998. Statistical reviewing for medical journals. *Statistics in medicine* 17, 23 (1998), 2661–2674. http://www.medicine.mcgill.ca/epidemiology/moodie/AGLM-HW/Altman1998.pdf
2. Patrick M De Boer and Abraham Bernstein. 2016. PPLib: Toward the Automated Generation of Crowd Computing Programs Using Process Recombination and Auto-Experimentation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (2016), 49. http://dl.acm.org/citation.cfm?id=2897367
3. Peter T Choi. 2005. Statistics for the reader: what to ask before believing the results. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie* 52 (2005), R46–R46. http://link.springer.com/article/10.1007%2FBF03023086
4. Douglas Curran-Everett and Dale J Benos. 2004. Guidelines for reporting statistics in journals published by the American Physiological Society. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 287, 2 (2004), R247–R249. http://ajpregu.physiology.org/content/287/2/R247.short
5. Casey Dugan, Werner Geyer, and David R Millen. 2010. Lessons learned from blog muse: audience-based inspiration for bloggers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1965–1974. http://dl.acm.org/citation.cfm?id=1753623
6. Phillipa J Easterbrook, Ramana Gopalan, JA Berlin, and David R Matthews. 1991. Publication bias in clinical research. *The Lancet* 337, 8746 (1991), 867–872. http://www.sciencedirect.com/science/article/pii/014067369190201Y
7. Andy Field. 2013. *Discovering statistics using IBM SPSS statistics*. Sage.
8. Rink Hoekstra, Henk Kiers, and Addie Johnson. 2012. Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in psychology* 3 (2012), 137. http://journal.frontiersin.org/article/10.3389/fpsyg.2012.00137/full
9. Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1105–1114. http://dl.acm.org/citation.cfm?id=2208557
10. Thomas A Lang and Douglas G Altman. 2014. Statistical Analyses and Methods in the Published Literature: The SAMPL GuidelinesâĂİ. *Guidelines for Reporting Health Research: A User's Manual (eds D Moher, DG Altman, K. F Schulz, I Simera and E Wager), John Wiley & Sons, Ltd, Oxford, UK. doi* 10 (2014), 9781118715598. http://www.wame.org/PDFs/SAMPL.pdf
11. James D. Leeper. 2016. What statistical analysis should I use? (Aug. 2016). http://www.ats.ucla.edu/stat/spss/whatstat/

12. Pascal Lessel, Maximilian Altmeyer, and Antonio Krüger. 2015. Analysis of recycling capabilities of individuals and crowds to encourage and educate people to separate their garbage playfully. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1095–1104. dl.acm.org/citation.cfm?id=2702309
13. Robert J MacCoun. 1998. Biases in the interpretation and use of research results. *Annual review of psychology* 49, 1 (1998), 259–287. http://www.annualreviews.org/doi/full/10.1146/annurev.psych.49.1.259
14. D.S. Moore, G.P. McCabe, and B.A. Craig. 2012. *Introduction to the Practice of Statistics*. W.H. Freeman.
15. Michèle B Nuijten, Chris HJ Hartgerink, Marcel ALM Assen, Sacha Epskamp, and Jelte M Wicherts. 2015. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods* (2015), 1–22. https://link.springer.com/article/10.3758/s13428-015-0664-2/
16. Alexander M Strasak, Qamruz Zaman, Karl P Pfeiffer, G Gobel, and Hanno Ulmer. 2007. Statistical errors in medical research-a review of common pitfalls. *Swiss medical weekly* 137, 3/4 (2007), 44. http://www.isdbweb.org/app/webroot/documents/file/855_11.pdf
17. Coosje LS Veldkamp, Michèle B Nuijten, Linda Dominguez-Alvarez, Marcel ALM van Assen, and Jelte M Wicherts. 2014. Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS one* 9, 12 (2014), e114876. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114876
18. Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding novices in statistical analysis. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2693–2702. http://dl.acm.org/citation.cfm?id=2702347
19. Christopher J Wild. 2000. Chance encounters: A first course in data analysis and inference. (2000). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.175.808