A Solution for Automated Grading of QDA Homework

Andreas Kaufmann
Friedrich-Alexander-Universität
Erlangen-Nürnberg
andreas.kaufmann@fau.de

Dirk Riehle
Friedrich-Alexander-Universität
Erlangen-Nürnberg
dirk@riehle.org

Julia Krause
Friedrich-Alexander-Universität
Erlangen-Nürnberg
julia.krause@fau.de

Nikolay Harutyunyan
Friedrich-Alexander-Universität
Erlangen-Nürnberg
nikolay.harutyunyan@fau.de

Abstract

Teaching research methods is important in any curriculum that prepares students for an academic career. While theoretical frameworks for qualitative theory building can be adequately conveyed through lecturing, the practices of qualitative data analysis (QDA) cannot. However, using experiential learning techniques for teaching QDA methods to large numbers of students presents a challenge to the instructor due to the effort required for the grading of homework. Any homework involving the coding of qualitative data will result in a myriad of different interpretations of the same data with varying quality. Grading such assignments requires significant effort. We approached this problem by using methods of inter-rater agreement and a model solution as a proxy for the quality of the submission. The automated agreement data serves as the foundation for a semi-automated grading process. Within this paper, we demonstrate that this proxy has a high correlation with the manual grading of submissions.

1. Introduction

Qualitative data analysis (QDA) is a task that benefits tremendously from experience. Reading a textbook only goes so far in preparing a budding researcher for the first time to perform coding of rich qualitative data. It, therefore, is of benefit to integrate practical homework assignments into a university course teaching QDA methods (DeLyser et al., 2013).

At the core of QDA is the process of coding the data. This involves the identification of theoretical constructs in the data and assigning labels (codes) to those segments of data. However, coding usually does not have a single correct result that can be easily checked and efficiently graded for a large number of students. The effort required for reviewing, grading, and giving feedback on such homework is substantial. This provides a significant challenge for including experimental learning in such a course if student numbers are not limited to the low double digits.

While the iterative process of data gathering and analysis usually does not fit the scope of a course with a total work effort by participants of 120h to 150h, some basic experience with coding qualitative data can be gathered by coding existing data with a given code book. This mirrors the practice of enhancing the trustworthiness of a research project through *inter-rater agreement*. Inter-rater agreement measures "the extent to which different raters assign the same precise value for each item being rated" (Gisev et al., 2013).

This paper evaluates how to use inter-rater agreement metrics for evaluating student homework on QDA demonstrated at a worked example at our university. Its contributions are the following:

- Definition and demonstration of an approach to use inter-rater agreement for grading
- Empirical evaluation of the approach using data from several course generations

While grading against a model solution is not novel, and the agreement metrics used here are well established as a form of quality assurance for rigorous research, the area of application for such a metric is novel. Its utility as a proxy for homework quality in a teaching environment has never been empirically validated.

The remainder of this article is structured as follows. Section 2 presents related work. We then outline our research design in section 3 followed by the results in section 4 and a discussion in section 5. We acknowledge the limitations of our research in section 6 and finish with a conclusion in section 7.

URI: https://hdl.handle.net/10125/102635
978-0-9981331-6-4

2. Related Work

This paper focuses on grading qualitative research exercises. It draws on previous research on teaching qualitative research methods and grading students by using inter-rater agreement or peer assessment.

2.1. Teaching qualitative research methods

Teaching QDA and training students to use computer assisted qualitative data analysis software (CAQDAS) at the same time is a frequently identified challenge (Kalpokaite & Radivojevic, 2020; Roberts et al., 2013; Silver & Rivers, 2016). While methodological expertise is required to fully utilize CAQDAS (Silver & Woolf, 2015), the ability to proficiently use QDA software also increases the methodological awareness amongst postgraduate students (Silver & Rivers, 2016).

Roberts et al. (2013) studied teaching QDA to more than 60 undergraduate students using NVivo[1]. 67 students provided feedback on using NVivo in an online survey. Addressing the issue of combining methodology teaching and tool training was done by Silver and Rivers (2016). They presented a CAQDAS Postgraduate Learning Model to combine the teaching of methodology knowledge and technology training. Blank (2004) reported on teaching 15 students using Qualrus[2]. In contrast to our approach, each student had to create a code system, assign codes, and write a summary. Afterwards they got written feedback. However, preparing written feedback for more than 60 students in our case was not feasible.

The benefits of *experiential learning* (Kolb, 2014) are well documented (Gentry, 1990). Jiusto and DiBiasio (2006) have shown a link between experiential learning and readiness for self-directed learning and life-long learning. One common strategy to implement an experiential aspect in teaching QDA methods while also making use of peer-feedback rather than individual feedback from the instructor is following a lecture and classroom discussion "with small group work, in order to allow students to articulate and compare their interpretations" (Raddon et al., 2009).

The focus of these papers is on teaching students QDA supported by tooling. We could not find much insight on handling the reviewing and grading of results. Much effort and time is necessary to provide individual feedback or at least to grade the results. We want to present an approach to provide a fast and objective grading mechanism.

[1]NVivo, see https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[2]Qualrus, see http://www.qualrus.com/

2.2. Inter-rater agreement as grading method

Our solution to the problem of workload for teaching assistants (TAs) builds on the concept of inter-rater agreement. Deciding on the most suitable assessment strategy for inter-rater agreement and making the procedure as transparent as possible is a challenge that is frequently addressed (Campbell et al., 2013; Gisev et al., 2013; LeBreton & Senter, 2008; Spence & Lachlan, 2005; Tinsley & Weiss, 2000). The primary goal of inter-rater agreement is to evaluate the quality of a code system. This may be performed as a measure of quality assurance where a researcher wants to improve the trustworthiness of their analysis by adding investigator triangulation. Another common application is to synchronize a team of collaborative coders, where each team member codes a different part of the data within a large project (Popping, 2010). High inter-rater agreement can be interpreted as a sign that the whole team involved in the coding has a common understanding of each code and when to use it.

In our teaching case, we already have a high-quality solution that is used to evaluate the individual results of the inter-rater agreement created by students. Here the primary goal is to teach QDA supported by CAQDAS. In general, the difficulty of teaching QDA to a large number of students, and the effort required to evaluate and give feedback on an exercise that includes coding of qualitative data is a long-standing problem (Delyser, 2008; DeLyser et al., 2013; Lowe, 1992; Sidaway, 1992; Spence & Lachlan, 2005). However, while efforts of mitigating work effort due to better and more efficient execution of the exercises, for example through thorough handbooks with experiences from previous semesters (Pile, 1992) or group exercises (Hein, 2004; Madill et al., 2005; Raddon et al., 2009), the assessment of student performance, as discussed in our work, received less attention.

In a literature review of 113 papers on teaching qualitative research methods, Wagner et al. (2019) underscored the need for further research on this topic in their recommendations for future research.

2.3. Peer assessment as grading method

Another potential solution to the problem of workload for the instructor within the context of teaching QDA is to employ peer assessment, which is a type of peer-feedback. There exist three distinct types of peer assessment (Kane & Lawler, 1978). These types can be distinguished as the following:

- **Peer-nomination**, where students nominate high- or low-performers

- **Peer-rating**, where students rate each other's work from bad to good
- **Peer-ranking**, where students would rank-order each other's performance

Peer-rating provides the most similar type of output compared to our approach for evaluation, but it might not be the most reliable (Kane & Lawler, 1978). Dingel et al. (2013) found that in their course of 101 students, peer-rating did not strongly correlate with performance, potentially due to different, maybe less appropriate, rating criteria such as effort instead of the quality of output. In a meta-analysis, Li et al. (2020) also found that reported results on the application of peer assessment are mixed, but identified mediating factors that can improve the effect, such as rater training and computer assisted peer assessment.

Still, besides the rating itself, peer-evaluation provides a good opportunity for students to receive more feedback in order to improve future homework.

Peer-grading as a specific aspect of peer-assessment is especially challenging when it comes to teaching at scale. The situation was ameliorated with the introduction of tools like CrowdGrader[3] (Dasgupta & Ghosh, 2013; De Alfaro & Shavlovsky, 2014) and various massively open online courses (MOOCs) platforms (Gehringer, 2014). De Alfaro and Shavlovsky (2014) found that the grades computed by CrowdGrader were precise and suitable for student homework evaluation. They compared the grading accuracy of a peer-grading student with a fully random grader. In contrast, we evaluate the accuracy of our grading by comparing the automated grading with the manual grades of the TAs. Similarly, the results show that automated grading was not less accurate than human grading.

The application of peer-grading could be a solution to the described problem of instructor feedback being the bottleneck for scaling student numbers. However, the studies we found are unspecific to the peculiarities of teaching QDA, so we are lacking empirical evidence for the appropriateness for this type of exercise. Having had some experience with CrowdGrader in the context of a different course we decided to investigate a possible automated solution as described here first.

3. Research Design

3.1. Research Question

Our research question was triggered by the challenges of scaling the teaching of QDA, as laid out in section 2. A consequence we frequently encountered when talking to lecturers is that practical exercises are not offered, because it would turn grading into a nearly insurmountable task.

In an attempt to ease this task by partially automating homework grading for exercises to teach QDA to college students, we implemented an algorithm for inter-rater agreement in our own QDA software, QDAcity [4].

A possible alternative solution for scaling participation numbers, other than automation, is to distribute grading among multiple TAs. However, these human raters would naturally have a certain degree of variation between them, even when structured and pre-defined evaluation criteria are followed. To evaluate whether our automated evaluation is a fair substitution for manually evaluating each homework submission we were guided by our research question:

> ***RQ:*** *Is the disagreement between the automatically assigned categories based on our automated approach and manually assigned evaluation categories smaller or equal to the disagreements expected with multiple human raters?*

3.2. Data Source

Our evaluation data is gathered from a long-running research course that we have been teaching to students at both a Bachelor's and Master's level. The course has grown over time, drawing students from many departments outside of computer science and the engineering faculty. The growth in student numbers made teaching QDA cumbersome and gave rise to the software solution that forms the innovation described in this article. Typical class size in recent years was about 80 participating students at which point giving individual feedback was no longer feasible.

Our course teaches students how to perform research for their final thesis. Part of it is devoted to research methods, including a four-week long exercise on QDA. Each week, students deliver one homework assignment.

The homework requires students to code a set of six expert interviews on the topic of user experience design (UXD) in product line engineering (PLE). In the first three weeks, each participant is instructed to code two interviews with a given code system. After the first week of coding, a Q&A session is offered with the investigator who originally created the code book (MacQueen et al., 1998) to clarify any open questions in a classroom discussion.

In the final week, the task is to write a conclusion on

[3] https://www.crowdgrader.org/

[4] https://qdacity.com

a research question of choice, which the student feels is adequately addressed in the qualitative data coded in the previous weeks.

The learning objectives of this exercise are to empower the students to complete the following tasks:

- Identifying relevant concepts in the context of a large amount of qualitative data
- Working with a code book and distinguishing similar codes with slight distinction
- Synthesizing a written theory from the coded data

3.2.1. Evaluation of Homework Submissions The submission of the final hand-in of a two-page long written conclusion about a research question of choice was always graded manually by our teaching team. The preceding three weeks of coding exercises were evaluated automatically using an inter-coder agreement metric measuring the similarity of the student's coding to our model solution.

We measured agreement by using the *F-Measure* metric (see equation 1), with the unit of coding set to the paragraph level. The F-Measure is a common analysis method in information retrieval for binary classification. Other common agreement metrics which would work in our approach are Krippendorff's Alpha or Fleiss' Kappa. The calculation is automated in our tool, but other tools like MAXQDA, NVivo, or ATLAS.ti also support such metrics. The reason for our own tooling was mainly the handling of the number of submissions.

$$\text{FMeasure} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

To calculate recall and precision, we first need to identify the true positives (TP) and false negatives (FN). We also need to identify the true positives (TP) and false positives (FP). These elements are illustrated in Figure 1, where the correct coding the grader is looking for is shown in the model solution and the student output is shown below.

Recall is defined as the ratio between true positives and the sum of true positives and false negatives, as shown in equation 2.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Precision is defined as the ratio between the true positives, and the sum of true positives and false positives (see equation 3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

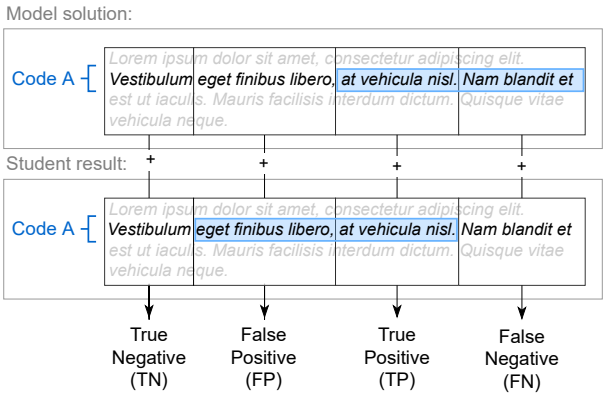


Figure 1. Information retrieval

Setting the unit of coding for evaluation to the paragraph level means, that it didn't matter how large a portion of any paragraph was coded by a student, or if one code was applied multiple times within one paragraph. Whenever a student applied a code at least once to a paragraph, and it was also applied at least once to that paragraph in our model solution, it was counted as a *true positive*. If it was only applied by the student it was a *false positive* and if it was applied only in the model solution it was counted as a *false negative*.

To map agreement scores to the nominal scale we use for grading, we randomly sampled submissions from every decile of agreement score (0-10%, 10%-20%, etc.). The sample was then manually evaluated according to the categories of our grading scale presented in table 1. We then determined thresholds for the agreement scores we consider to be sufficient to justify a particular item of the evaluation scale.

We sampled homework submissions from each evaluation category (exceptional, average, etc.) and manually evaluated the homework submissions to assess whether the defined thresholds and the resulting automatic classification, are fair. For instance, starting out we deemed submissions with over 30% agreement as excellent but decided to increase the threshold, thereby making it harder to earn full points in future semesters. The scale presented here (table 1) is based on our experience over six iterations of teaching the course using the same type of exercise with the same data to be analyzed. However, after just one correction after our initial definition of the boundaries for each category, the thresholds remained stable. We expect an iterative process of identifying appropriate thresholds to be necessary for any new type of data or change to the type of exercise.

The evaluation of the last assignment in this series, and thus the final result of the analysis, was never

Table 1. Grading Scale

Agreement (%)	Points	Description
< 10	0	No significant work
10-20	1	Minimal performance
20-40	2	Average performance
> 40	3	Exceptional performance

automated. It is important to note that exercises like these can, or should, not be fully automated, and we only argue for generating additional data points to influence grading and thereby lighten the load on the teaching team overall.

3.3. Data Sampling

Of the total number of 627 homework submissions graded by our inter-rater agreement metric, we sampled 11.16% (70 submissions) which were manually evaluated by the teaching team. We used the same grading scale of zero to three points based on table 1.

Before sampling, we stratified our population of data by ranges of the F-Measure score assigned through our grading system. Then we randomly sampled from each stratum resulting in a total of 70 automatically graded homework submissions which included 140 documents coded by course participants.

The representation of each stratum in our sample was deliberately chosen to be disproportionate to the occurrence of each stratum's characteristic in our overall population. Instead, we chose an equal representation of each stratum in our sample. The exception to this are submissions rated with zero points by our agreement algorithm, of which only ten submissions existed in our populations. Of each other category (1,2,3 points) we randomly sampled 20 submissions each.

We chose disproportionate selection from the strata because our goal was not to show that the population of our course was fairly evaluated, but whether our approach works for all of the four categories.

3.4. Evaluation Method

To evaluate the grading algorithm for our sample, we distributed the homework submissions randomly to three TAs. The TAs were tasked to manually evaluate each submission on the same scale of zero to three points as shown in table 1. Two TAs evaluated 23 submissions and one evaluated 24. The TAs were unaware of how many submissions of each stratum were within their sample. The human raters were presented with the results of the algorithmic evaluation only after their rating had concluded in order not to prime them with the expectation of an excellent or poor homework submission.

To ensure the objectivity of the manual evaluation performed by the TAs, we conducted several peer-debriefing sessions and followed a commonly agreed upon set of evaluation criteria including:

- Completeness (were all data coded?)
- Comprehension (were all codes applied?)
- Precision (were the codings appropriate?)
- Variability (were some codes over-/underused?)
- Relevance (were all codings relevant to the task?)
- Significance (were the key data coded?)

These evaluation criteria are a subset of defined terms previously used in the context of evaluative case study research (Harutyunyan, 2019). Given the similar nature of this study and the application of QDA methods, we adapted them to this research design.

The manual grading was an iterative process. We used investigator triangulation to increase the level of confirmability (and thus, trustworthiness) as defined by Guba, 1981. Each TA was assigned a random subset of student submissions and followed the above-mentioned criteria to grade the homework on the scale presented in table 1. We took detailed notes on the evaluation criteria.

The manual evaluation was restricted to double–checking a sample of the automatically evaluated submissions as quality assurance and to re-test the appropriateness of the chosen F-Measure thresholds. The results of the manual evaluation were then compared to the automatic evaluation algorithm based on inter-rater agreement.

In the second iteration of quality assurance, TAs swapped the homework submissions that diverged from the automated grades and reevaluated them. We used the following three questions.

Q1: *How would you rate your agreement with the following sentence? The automatically assigned category can be justified*

Q2: *How would you rate your agreement with the following sentence? The manually assigned category can be justified*

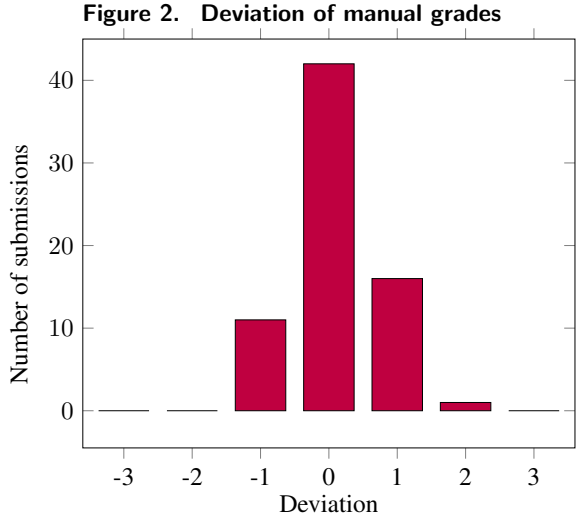
Q3: *With which category would you have rated the assignment?*

The answer to the first two questions were given on a five point Likert scale for agreement. The answer to the third question was given as a rating on our four-point scale for homework performance.

4. Results

We begin by presenting the analysis of the results of the first rater. Then we present the analysis of the

Figure 2. Deviation of manual grades



first rater, second rater, and the algorithm, which also provides insights into the variation in ratings between the different TAs involved in the course.

Figure 2 shows the number of deviations of the grades of the manually graded homework submissions from the automatically graded submissions. Zero means that the manually assigned grade is identical to the automatically assigned grade, -1 means that the human rater assigned one less point, and +1 means the human rater classified the assignment as one category higher.

Table 2 shows the number of homework submissions in our sample where our manual evaluation diverged from the algorithmic evaluation, as well as the average F-Measure agreement. The average F-Measure in this table is an average over the F-Measure score in only those specific types of divergences. Combinations that did not occur in our data are omitted from the table. The data shows, that divergence of the human rater from the algorithm is lower in the lower categories. In total, there were five instances of disagreements in distinguishing between the categories of 0 and 1 point and another seven instances of disagreement in distinguishing the categories of 1 and 2 points, however a total of 15 instances of disagreement between the categories of 2 and 3 points.

In distinguishing between the categories of 1 and 2 points the manual rater chose the lower category only once when the algorithm did not, but six times chose the higher category when the algorithm did not. This may suggest that the threshold for entering the F-measure range for the 2 points category could potentially be slightly lowered. This would also eliminate the one divergence of a distance between 2 categories. Such a change would have to be re-evaluated in a future semester, to avoid overfitting the categorization criteria to a specific subset of past data.

Table 2. Disagreements

Automated Eval.	Manual Eval.	# instances	Average F-Measure
0	1	2	5.93%
1	0	3	14.70%
1	2	6	18.04%
1	3	1	17.34%
2	1	1	24.54%
2	3	8	35.38%
3	2	7	45.58%

On average, a homework submission that was classified as one category better or worse by the algorithm had an F-Measure level, that was 2.7% above or below the threshold to the manual classification.

With many of the submissions which were evaluated differently from the human rater and the algorithm, the grading notes of the human rater already stated things like “strong 2 or weak 3”. So presumably either classification could have been justifiable, which is in line with many, albeit not all, of the deviating submissions being evaluated with F-Measure score close to the threshold for the range required for the manual evaluation category. One possible way of mitigating this issue could be more fine-grained grading categories.

Of the assignments, which the algorithm classified as zero points there was little disagreement. Only one deviation occurred, which was just 0.3% shy of the required 10% threshold. All of these submissions were incomplete, meaning long parts of the data were not coded by the student at all. Hence, the classification with the same amount of points that someone who did not submit anything was justified by the human rater.

More deviation occurred the other way around, where the algorithm gave one point, and the TA gave zero. This was a distinct category of error. If some significant portion of the data (for instance just one of the two documents) was coded with average precision and recall, but large parts remained not coded at all, then the algorithm averaged the whole submission to one point. The TA, however, argued that the assignment was not even finished, hence considered this a failed submission justifying a classification with zero points.

Figure 3 shows the answers to **Q1** and **Q2** of our questionnaire analyzing the disagreements of the human rater with the algorithmic classification by a second rater (another TA).

Both for the ratings of the first rater as well as the ratings from the algorithm, the second rater tended to agree with the statement that the rating would be justifiable rather than not, which manifests in the

Figure 3. Analysis of disagreement

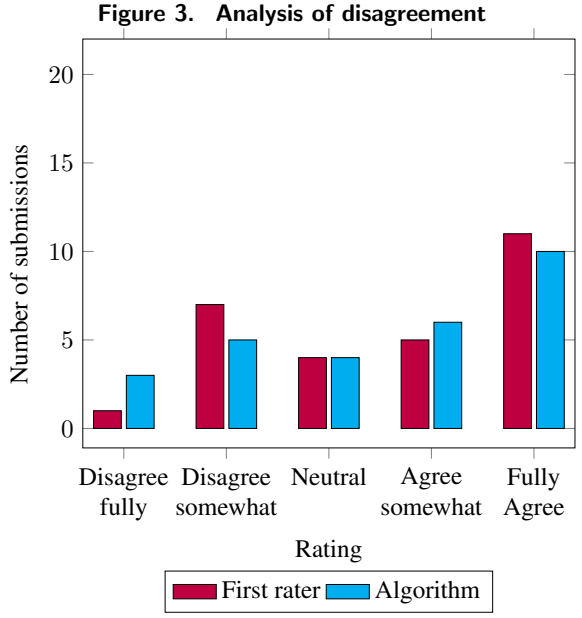
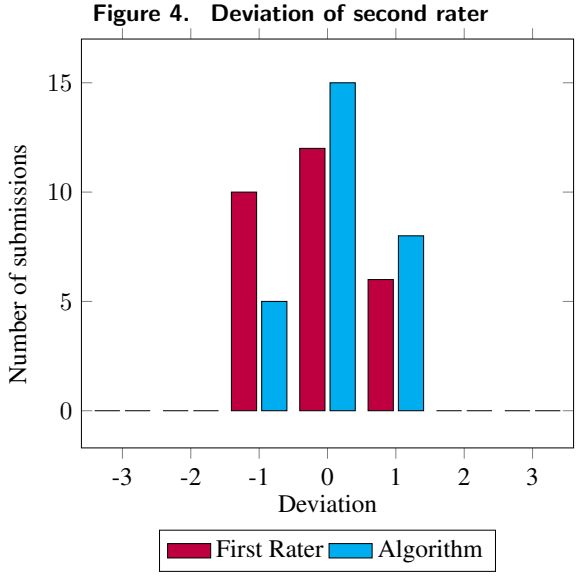


Figure 4. Deviation of second rater



most frequently used answer is "fully agree" with both ratings. In many cases this answer was chosen for both ratings of the same homework submission, indicating that either of the two ratings could be justified. This resonates with the frequent comments of our human raters about a submission being a borderline case between two ratings. In the one instance where there was a significant difference of two points between the algorithm and the first rater, the second rater disagreed with both ratings as either 1 point (algorithm) or 3 points (first rater) and instead argued for a rating of 2 points would be justifiable. Indeed, the first rater specified in his comments, that it was a borderline case between 2 and 3 points, and the algorithm categorized it as one of the stronger submissions of the 1-point-category.

When aggregating the number of disagreements, the second rater disagreed 8 times and agreed 16 times with both the algorithm and the second rater. However, there were slightly fewer strong disagreements with the first rater, as compared with the algorithm.

Figure 4 presents the deviation between the category assigned by the second rater and the automatically assigned category as well as the manually assigned category from the first rater. This addresses question **Q3**. In total, the second rater disagreed with the first rater 16 times while disagreeing with the automated rating 13 times.

This review stage corroborated a theme that had already been present in the evaluation by the first rater: Human raters, on average, have a tendency to rate submissions slightly higher than the algorithm.

Figures 2 and 3 support the claim that the disagreement between a human rater and the algorithm is not stronger than the disagreement between multiple human raters. Therefore, the algorithmic evaluation using inter-rater agreement metrics appears to be a valid substitute for evaluation through a human rater.

5. Discussion

We consider the results adequate for relying on our agreement metric as a proxy for the quality of the homework submission in the setting of our course. The thresholds for what can be considered a minimal average, and excellent performance used here are specific to our data and would need to be calibrated to any data that is used in a similar exercise. This is analogous to the question of what levels of inter-rater agreement metrics are sufficient to validate coding in a research project as being "good". While some classification of inter-coder ratings exists, any thresholds are in the end arbitrary and need to be contextualized.

While any threshold may be arbitrary there is always some discretion in manual evaluation that could be deemed arbitrary as well. And not every student receiving the same grade shows exactly the same performance. While a four-point scale is not nuanced if there are enough data points nuances can be expressed through the average on a larger data set of evaluations. For instance, our sample contained two submissions from the same student, which were both evaluated as a borderline case between average and excellent. Our manual rater ended up evaluating one submission with a weak three, the other with a strong two. While

both of these evaluations were in disagreement with the ratings from the algorithm, the averaged points for these students would have been the same, because the algorithm evaluated them the other way around.

Some level of disagreement between different raters is also common if two human raters evaluate different homework submissions. Our data show, that the level of disagreement between our human raters and the algorithm does not exceed the levels of disagreement which we could observe between multiple human raters. Out of the ratings in which the first rater and the algorithm disagreed, a large portion could be justifiably rated with either category, according to a second rater.

One frequently occurring category of disagreements between human raters and the algorithm was related to incomplete submissions. Many of these cases were evaluated with zero points from both the human rater and the algorithm. In some instances, however, the portion which was completed by the student was so similar to our model solution, that the achieved agreement was averaged so the assignment was categorized with one point by the algorithm. The TA rated such a case as a failed submission. This category of disagreements could be handled in multiple ways:

- Concede, that a partial good submission is as good as a complete, but less accurate submission.
- Require a minimum threshold of agreement for all documents (or even parts of documents), not just that the average meets a certain threshold.
- Give recall more weight over precision, leading to long swaths of uncoded data to drag down the score more heavily.

An advantage of the algorithmic evaluation is that it is free of biases in recognizing student names. If a student is known for their brilliant participation in the classroom this raises expectations when evaluating other parts of their work. Or if a student is recognized as below average in some different context, a spillover effect may put him or her at an unfair disadvantage, even if the teacher evaluating the submission actively tries to free themselves from such biases. The algorithm evaluates each submission without being primed with an impression of a student that is not directly linked to the homework's quality. This also guards against biases towards gender or ethnicity.

Our research agenda includes a replication study at a different university. Further, we are looking into an evaluation using a more fine-grained evaluation scale, which likely makes deviations more frequent, but each deviation less severe. Besides the scale, different agreement metrics are of interest for evaluation. Further, a combination of peer-assessment with our automated evaluation approach may be a promising avenue for future research.

6. Limitations

Through the high level of structure of the exercises presented here, some of the benefits of experiential learning that are linked to experiencing the, at times, messy complexity of a real world research project may be slightly stifled. However, according to the feedback we received from our students, even this rather structured exercise helped prepare them better than a lecture on its own could have done. Further, it should be regarded as just one part of a larger teaching concept, with other types of homework and lecture content filling the gaps that such a narrowly focused exercise can not deliver.

With polytomous data like an average code system, it should be acknowledged that disagreement between a pair of two distinct categories might be considered better or worse than disagreement between a different pair of categories (Maclure & Willett, 1987). For instance, within the data used in our exercise, the same basic concept was applied in different stages of product development, and for each phase, a code existed for a best practice on how to implement the concept within this stage. These codes were often confused by students. However, applying the code for the right concept in the wrong development phase could be considered better than applying a code that is completely nonsensical in the context of the data segment. For the agreement algorithm, this would not make a difference. While this specific problem could be observed with many of our participating students, we believe that if the number of such frequently confused codes is small, a sufficiently large sample of data to code can mitigate this effect.

We were not able to manually evaluate all homework submissions. Through stratification, we were able to ensure, that our sample was not biased towards containing more submissions from one category than the other. The exception to this is submissions in the category of 0 points because only 10 submissions of our population existed within this stratum. We believe, that under-representing this category is preferable to the alternative of a significantly smaller sample.

Anonymizing student names was not possible for technical reasons, so it is possible the human raters were biased towards particular names or the history of experiencing one particular student as a participant in this or in another course. This limitation was mitigated through a second round of evaluation through the second human rater in those cases where the first rater diverged

from the automated evaluation.

To mitigate researcher bias, the sample was randomized before being distributed to the TAs for evaluation. The human raters did not know the exact distribution of their assigned sample to the categories used for stratification.

Our evaluation scale was coarse-grained. The evaluation scale could be justified because each data point only constitutes one out of a minimum of ten submissions which constitutes one of three parts of the grade. At the end of class, students were free to demand an evaluation through an oral exam the grade of which would have superseded the grade calculated from the many data points gathered during the semester, but none of the students ever chose this option.

Our approach for scaling QDA exercises is specific to a setup with existing data and an existing code system. This replicates a scenario of investigator triangulation (Guion et al., 2002). Other types of exercises, which include data gathering and the creation of a new code book, could potentially be better scaled using an option for peer-feedback and peer-evaluation.

7. Conclusion

We present a way to scale teaching of QDA in the context of a course with practical exercises to a large number of participants. We use inter-rater agreement as a proxy for the quality of the homework submission.

We have shown, that an agreement metric strongly correlates with the manual evaluation of homework exercise submissions in a sample of 70 exercises and 140 coded documents. In our sample the human rater aligned with the rating of the algorithm in 42 cases (60%), in 11 cases (15.7%) the human rater chose one category lower, and in 16 cases (22.86%) one category higher. Only in one instance (1.43%), the divergence was two categories higher in the manual evaluation, in which case the second rater strongly believed the rating between the two extremes was appropriate.

We found that our teaching of QDA can be partially automated by evaluating one particular type of coding exercise with an inter-rater agreement algorithm instead of manually assessing each homework submission.

We compared the disagreement between a second rater and the algorithm as well as between the same second rater and the first rater in those cases where the first rater disagreed with the algorithm. Our data shows, that disagreement between human rater and the algorithm was on the same level as disagreement between multiple human raters.

While there is no full agreement between our manual evaluation and the algorithm, almost all of the deviations concerned borderline cases where either grade could be justifiable. In fact, a survey of the second raters showed many instances where the rater agreed fully, or partially with the statement that the grade from the first rater and the algorithm could both be justified.

This answers our ***RQ*** by finding in our data that disagreement between human raters and the algorithm is not larger than the disagreements among human raters.

References

Blank, G. (2004). Teaching Qualitative Data Analysis to Graduate Students. *Social Science Computer Review*, *22*(2), 187–196.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research*, *42*(3), 294–320.

Dasgupta, A., & Ghosh, A. (2013). Crowdsourced judgement elicitation with endogenous proficiency. *Proceedings of the 22nd international conference on World Wide Web*, 319–330.

De Alfaro, L., & Shavlovsky, M. (2014). Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. *Proceedings of the 45th ACM technical symposium on Computer science education*, 415–420.

Delyser, D. (2008). Teaching qualitative research. *Journal of geography in higher education*, *32*(2), 233–244.

DeLyser, D., Potter, A. E., Chaney, J., Crider, S., Debnam, I., Hanks, G., Hotard, C. D., Modlin, E. A., Pfeiffer, M., & Seemann, J. (2013). Teaching qualitative research: Experiential learning in group-based interviews and coding assignments. *Journal of Geography*, *112*(1), 18–28.

Dingel, M. J., Wei, W., & Huq, A. (2013). Cooperative learning and peer evaluation: The effect of free riders on team performance and the relationship between course performance and peer evaluation. *Journal of the Scholarship of Teaching and Learning*, *13*(1), 45–56.

Gehringer, E. F. (2014). A survey of methods for improving review quality. *International Conference on Web-Based Learning*, 92–97.

Gentry, J. W. (1990). What is experiential learning. *Guide to business gaming and experiential learning*, *9*, 20.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Ectj*, *29*(2), 75–91.

Guion, L. A., Diehl, D., & McDonald, D. (2002). *Triangulation: Establishing the validity of qualitative studies* (tech. rep.). University of Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, EDIS.

Harutyunyan, N. (2019). *Corporate open source governance of software supply chains* (Doctoral dissertation). Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

Hein, S. F. (2004). " i don't like ambiguity": An exploration of students' experiences during a qualitative methods course. *Alberta Journal of Educational Research*, *50*(1).

Jiusto, S., & DiBiasio, D. (2006). Experiential learning environments: Do they prepare our students to be self-directed, life-long learners? *Journal of Engineering Education*, *95*(3), 195–204.

Kalpokaite, N., & Radivojevic, I. (2020). Teaching qualitative data analysis software online: A comparison of face-to-face and e-learning ATLAS.ti courses. *International Journal of Research & Method in Education*, *43*(3), 296–310.

Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. *Psychological bulletin*, *85*(3), 555.

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods*, *11*(4), 815–852.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? a meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193–211.

Lowe, M. S. (1992). Safety in numbers? how to teach qualitative geography? *Journal of Geography in Higher Education*, *16*(2), 171–175.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American journal of epidemiology*, *126*(2), 161–169.

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. *CAM Journal*, *10*(2), 31–36.

Madill, A., Gough, B., Lawton, R., & Stratton, P. (2005). How should we supervise qualitative projects? *The Psychologist*, *18*(10), 616–618.

Pile, S. (1992). Oral history and teaching qualitative methods. *Journal of Geography in Higher Education*, *16*(2), 135–143.

Popping, R. (2010). Some views on agreement to be used in content analysis studies. *Quality & Quantity*, *44*(6), 1067–1078.

Raddon, M.-B., Raby, R., & Sharpe, E. (2009). The Challenges of Teaching Qualitative Coding: Can a Learning Object Help? *International Journal of Teaching and Learning in Higher Education*, *21*(3), 336–347.

Roberts, L. D., Breen, L. J., & Symes, M. (2013). Teaching computer-assisted qualitative data analysis to a large cohort of undergraduate students. *International Journal of Research & Method in Education*, *36*(3), 279–294.

Sidaway, J. D. (1992). Qualitative change? innovation and evaluation in the course at reading.

Silver, C., & Rivers, C. (2016). The CAQDAS Postgraduate Learning Model: An interplay between methodological awareness, analytic adeptness and technological proficiency. *International Journal of Social Research Methodology*, *19*(5), 593–609.

Silver, C., & Woolf, N. H. (2015). From guided-instruction to facilitation of learning: The development of five-level qda as a caqdas pedagogy that explicates the practices of expert users. *International Journal of Social Research Methodology*, *18*(5), 527–543.

Spence, P., & Lachlan, K. (2005). Teaching Intercoder Reliability: A Gentle Introduction to Content Analytic Methods for Graduate Students. *Texas Speech Communication Journal*, *30*, 71–76.

Tinsley, H. E. A., & Weiss, D. J. (2000). 4 - Interrater Reliability and Agreement. *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 95–124). Academic Press. https://doi.org/10.1016/B978-012691360-6/50005-7

Wagner, C., Kawulich, B., & Garner, M. (2019). A mixed research synthesis of literature on teaching qualitative research methods. *SAGE Open*, *9*(3), 2158244019861488.