Reflection - quantifying a rare good

Thomas Daniel Ullmann*, Fridolin Wild, and Peter Scott

Knowledge Media Institute, The Open University
Walton Hall, MK7 6AA Milton Keynes, United Kingdom
{t.ullmann,fridolin.wild,peter.scott}@open.ac.uk
http://kmi.open.ac.uk

Abstract. Based on a literature review, reflections in written text are rare. The reported proportions of reflection are based on different baselines, making comparisons difficult. In contrast, this research reports on the proportion of occurrences of elements of reflection based on sentence level. This metric allows to compare proportions of elements of reflection. Previous studies are based on courses tailored to foster reflection. The reported proportions represent more the success of a specific instruction than informing about proportions of reflections occurring in student writings in general. This study is based on a large sample of course forum posts of a virtual learning environment. In total 1000 sentences were randomly selected and manually classified according to six elements of reflection. Five raters rated each sentence. Agreement was calculated based on a majority vote. The proportions of elements of reflection are reported and its potential application for course analytics demonstrated. The results indicate that reflections in text are indeed rare, and that there are differences within elements of reflection.

Keywords: quantification, reflection, reflective thinking, reflective writing, reflection detection, reflection analytics

1 Introduction

The phenomenon "reflection", a pivotal thinking skill, has a rich theoretical tradition. Several methods have been developed to measure reflection, especially in the area of reflective learning. Analytical models of reflection in writings can be distinguished by three types, covering the depth, the breadth, or the process of reflection (e.g. [2–4, 6, 7, 9, 10, 13–15]). These models decompose reflection into several elements characterising reflection.

Little is known about the quantities of these reflective elements in texts. Quantification is the mapping of phenomena into a set of numbers. It is core to scientific research, as it allows to investigate the properties of the phenomenon in their context, its relations, probabilities, and patterns, to test theoretical assumptions, in order to develop rules, laws, towards a general theory.

Regarding reflection, it is still a largely unmapped territory, when it comes to the quantification of properties of reflection in texts. This research tries to

* Corresponding author

find answers to the question of what can be expected regarding the frequency of occurrences of reflection in texts. Intuitively reflection occurs rarely in writings. But, how rarely does reflection actually occur?

While there are attempts to quantify the proportions of elements of reflection in reflective text the reported results are hard to generalise. Guidelines to compare reflection studies do not exist yet. This research proposes a method based on a comparable unit of analysis, which allows to estimate frequencies of reflective elements in texts. It describes the method used to quantify reflection, and exemplifies the process on six elements of reflection. The study was conducted in the context of course forum posts of a virtual learning environment.

2 Proportions of reflection in texts

The following literature review outlines findings on three levels of reflection research, starting with a meta-analytical view, followed by research on the depth of reflection, and proportions of elements of reflection.

Dyment and O'Connell [1, p. 90-91] undertook a meta-review of the quality of reflection in student writings. They included 11 studies in their review. Amongst other criteria, they looked at the distribution regarding the depth of reflection. They classified the outcome of each study either as low, moderate, or high. A study classified as low for example had a high percentage of texts, which were mostly descriptive and less reflective. Five studies were categorised as low, four studies had a moderate level of reflection, and two had a high level of reflection. According to their review a relatively high proportion of studies report low to medium levels of reflection, while only two studies achieved high levels of reflection.

Although the categorisation into levels of reflection is informative, they warn that the categorisation is to a degree subjective as the used methods in each study vary making the comparison difficult.

The following section outlines results of individual studies describing proportions of elements of reflection, and their context. Some of the studies are already covered by Dyment and O'Connell [1]. Additional information was added to situate these studies in their context. In one case further reported results are presented. Studies not included in their review are marked with an asterisk.

*Wald et al. [13, p. 43] report on the distribution of reflective levels (depth of reflection) within a corpus of 93 reflective writings. The reflective writings stem from second-year students self-selected best reflective writing field notes. The reflective writings were selected from archived material and were not in connection with a instructional setup of the researchers. The levels started at level 1 with "nonreflective: habitual action" (0% of students), "nonreflective: thoughtful action" (18%), "reflective" (41%), "critically reflective" (30%), to level 5 "transformative learning" (11%).

Wong et al. [15, p. 53-54] conducted a content analysis (and interviews) of reflective writings in the context of an instructional design specifically targeting reflective writing. The writings were however not graded. The percentages of 45

students regarding the level of reflection were as follows: 13% non-reflectors, 76% reflectors, and 11% were critical reflectors.

Plack et al. [7, p. 204] analysed reflective writings of 43 journals of students participating in a course for clinical practice, which had an emphasis on reflective practice (instructions were given about reflection, journals were not used for grading). They report frequencies of elements and depth of reflection. Regarding the depth, 15% of the journals had no evidence of reflection, 43% showed evidence of reflection, and 42% evidence of critical reflection. The percentages for the elements of reflection were (percentage of journals containing an element): Reflection in action (23%), reflection on action (38%), reflection for action (28%), content (35%), process (38%), premise (18%), returns to experience (39%), attends to feelings (38%), and re-evaluation (32%). The most frequent element was return to experience, which is the category label for a description of an important experience. The premise element - critique of own assumptions - is the less frequent one (see Plack et al. [7, p. 206-7] for descriptions of the elements).

Hatton and Smith [3, p. 41] assessed the written work of 60 students in the context of a professional program, which used reflection-fostering instructions. They reported overall percentages of coded units. 60-70% were descriptive reflective, and more than 30% of dialogic reflection was found in essays after a special instruction. On average 19 reflective units were found in a writing of 8-12 pages.

*The analysis of Ross [8, p. 24-25] took into account 134 papers from 25 students (average 5.4 papers per student). The students took a course with a special focus on reflection. The article reported the following percentages of papers: 22% were highly reflective, 34% moderately reflective, and 44% low reflective.

Williams et al. [14, p. 7 and 12] report the highest achieved level of reflection of 56 students, who had to write a reflective journal during a course (the journal made up 10% of the grade, at least one journal had to be written per week). The percentages from the lowest level to the highest are: 0% describe learning, 2% analyse learning, 23% verify learning, 36% gain new understanding, and 39% indicate future behaviour.

The proportions of students or texts regarding level or elements of reflection have to be used with care. The mapping from the evidences of reflection to elements or levels depends largely on the interpretation of the researchers, and thus the percentages might be different if the mapping process would have been done differently. In addition, the reported percentages are on either a level of a person or journal. Thus, not much can be said about the distribution of reflective elements or levels in texts. Tentatively, the presented research results might indicate that most texts/students are written/write with a low or medium level of reflection. Highly reflective texts/students are rather rare. All presented studies describe that the course was especially designed to foster reflection, which compared to normal courses might result in higher proportions of reflection. The articles reporting on elements of reflection indicate that some elements occur more frequently than others. The work of Hatton and Smith [3] is insofar from special interest for this research as they provide percentages of units of reflection, which might give indications about the frequencies of reflective utterances

in text. However, they do not specify exactly their understanding of a unit, which makes estimates speculative. For example, if a unit is a sentence, then 19 reflective units in 8-12 pages would indicate that reflections are rare instances (about 1% of the text assuming 200 sentences for 10 pages). If a unit equals a paragraph and if we assume that a page consists of five paragraphs, then 38% of the units of a text would be reflective considering 10 pages.

3 Method

The chosen approach to quantify reflections follows seven points:

- Choose a text corpus and describe its domain and characteristics.
- Unitise the text corpus (in here the unit of analysis are sentences).
- Draw a random sample of units.
- Based on theory of reflection derive elements of reflections, which characterise it.
- Operationalise each element of reflection.
- Device a strategy to gain annotations for the units. In this paper the units are manually annotated.
- Calculate and report the proportions of the annotated elements of reflection (quantification).

It is worth to consider controlling the text length (amount of units of each text), especially for a smaller corpus. For example, if a corpus with a very long text and several very short texts is used, the randomly selected units will come mostly from the long text. If the long text is about describing a problem and the short texts are mostly acknowledging or short introductions of members of a forum, the uncontrolled corpus will be biased.

The theories on reflection vary regarding the elements which together model reflection. Furthermore, the elements of reflection are often high level descriptions, which may be too abstract to measure. For each element several subelements can be designed, with the aim to arrive at measurable elements. This list of subelements may be too large to be administered in a single experiment. A pre-test can help to find the optimal amount of items. A pre-test is also advisable to check the measurability of each item. Inhere six items were selected. For each element of reflection one or two items were selected. The rational is that the items are better distinguishable by human raters, instead of using several items from one element, which may be too similar.

4 Text corpus

The text corpus used for this research is based on forum data of the virtual learning environment[1] of the Open University, UK. It consists of two courses on eLearning, two courses about social work, and one course on science. The data was de-identified by the researcher.

[1] `https://github.com/moodleou/`

course	description	posts	unique sentences with personal pronoun	unique sentences without personal pronoun
eLearning 1	postgraduate course (30 credits)	410	3454	4639
eLearning 2	postgraduate course (30 credits); different semester than eLearning 1 course	274	2115	2480
social work	level 2 course (60 credits)	475	3787	3689
social work	level 3 course (60 credits)	103	903	872
science	postgraduate course (30 credits)	355	2170	3820

Table 1: Description of text corpus

The forums serve several purposes, for example to support the students, to have a platform for discussion, exchanging ideas, and to socialise. The forum posts therefore contain a wide spectrum of writings. The eLearning and social work courses explicitly stated that reflection is one of the learning goals. However, special assignments regarding reflection were in general[2] not conducted within the forums, but with forum external means, and thus may be only indirectly present in the forum corpus. The science course did not explicitly aim at reflection.

From all posts, posts of a character length between 1500 and 3500 characters were selected. On average the 1677 forum posts were 2121 characters long (sd=512).

Only course forums, which were core to the course and explicitly embedded in the activity of the course were kept. Forums for technical support, student talk, and general course-wide forums were excluded.

From all posts the ones made by the role "student" are kept, while the other roles like "tutor", "moderator", or "production staff" were filtered.

All texts were split into single sentences. These were then divided into sentences that contain a personal pronoun and sentences that did not contain a personal pronoun (using the reflection detection architecture of Ullmann [11] and its extension [12, p. 106f.]). 500 sentences from each set were randomly selected. They form the input of the corpus of sentences used for the annotation process.

[2] some courses had forums to discuss reflection assignments

5 Questionnaire

The questionnaire contained four sentences on each page to rate. A maximum of 30 sentences could be coded by a coder per batch. The instruction of the questionnaire explained the task. Each category was described with an example sentence and explanation. The workflow for the raters was: read a sentence, categorise it, and write a short explanation justifying your choice.

The raters could choose from one of the following seven categories. These categories follow the reflection model outlined in Ullmann et al. [12, p. 103f.]. Something could/should have been done differently, drawing a conclusion based on a premise (reasoning), taking another perspective (point of view), intention to do something, something was successfully learned (achieved), something is interpreted in a new way (new understanding), and none of these. The raters were prompted to choose the best category for each sentence. In addition, the instruction stated that the sentence had to speak for itself. The categories were presented for each sentence in random order. Pre-experiments showed a preferred answer bias for the first category. Shuffling the categories aimed at minimising this bias based on the sequence of the categories. After the rater selected their answer, they had to justify their choice with a short free text answer.

The categories can be seen as one of several variations to capture an element. This also means, that the questions do not exhaustively represent an element, but they cover to a certain extent the essence of it. The following table shows the mapping between the elements of reflection [12, p. 103f.] and the categories.

Elements of reflection	Categories
Description of an experience	Something could/should have been done differently
Critical analysis	Drawing a conclusion based on a premise (reasoning)
Taking other perspectives into account	Taking another perspective (point of view). Something is interpreted in a new way (new understanding)
Outcome	Intention to do something. Something was successfully learned (achieved). Something is interpreted in a new way (new understanding)

Table 2: Mapping of elements of reflection to the categories of the questionnaire.

The question "something is interpreted in a new way (new understanding)" can be seen as an outcome dimension of reflection, but also as another take/perspective on something.

6 Survey

The questionnaire was distributed via a crowdsourcing platform[3]. The participants had to fill out a minimum of four gold questions before they saw the first item of the survey. Gold questions are items with known answers used to stop those participants that fail to correctly answer a certain amount of gold questions from filling out the survey. Additionally, a set of own text validators was used to discourage participants filling out the questionnaire randomly.

10 batches of 100 sentences were administered. Participants from previous batches were allowed to rate a new batch. Each sentence was rated by at least five raters.

The 411 raters came from 17 countries. Most of them were from the USA (n = 202), GBR (n = 94), and IND (n = 45). The remaining 70 raters came from 14 other nations.

7 Results

The inter-rater agreement between the annotators was measured using Krippendorff's α [5] for nominal data. The α for the gold data is 0.43, for annotation and gold data combined it is 0.32, and for the annotation data only is 0.22.

The annotated sentences were then filtered. Only sentences where three or more raters agreed on (majority vote) remained in the data set. From the original 1000 sentences 623 remain. Krippendorff's α for this set is 0.36. The α values of the participants are relatively low, which means that by running the same experiment on the crowdsourcing platform again, some of the sentences will be classified differently. It has however the benefit that it does not rely on expert ratings, which also might be difficult to replicate by other researchers.

The following table shows sentences from the experiment and their classification.

Category	Example
Something could/should have been done differently	Victor and Morgan you are right that I should have applied better my own learning instead of using the Uni ones. [names were de-identified before the survey]
Drawing a conclusion based on a premise (reasoning)	I imagine this is probably in order to have a focus and provide enough detail rather than skim over the whole area.
Taking another perspective (point of view)	When I am doing FRT work, I often think about how the parents view me when they know I haven't got children!

[3] `http://crowdflower.com/`

Something is interpreted in a new way (new understanding)	After I saw how this lifted her mood and eased her anxiety, I will remember that what we can view sometimes to be small can actually make a significant difference.
Intention to do something	I would like to be involved in helping with the site too -although I'm a novice!
Something was successfully learned (achieved)	This has helped me reflect on my own life and experiences whilst allowing me to empathise with others in their own circumstances, I feel proud of what I have achieved so far as the work/life/study balance is always difficult to navigate but I'm lucky that I have a supportive family to help.
None of these	Bye the way, Audacity is also run under the CC Attribution licence.

Table 3: Example sentences for each category.

As we are interested in the proportions of reflective elements and not so much in the replication of every sentence's judgement, an analysis of the stability of the proportion over different points in time was conducted. The assumption is that similar proportions will arise, if the number of sentences is big enough and the sentences are randomly selected.

To test this assumption the 10 runs were split into 3 equally sized batches of 208 items. The following table shows the results.

	batch 1	batch 2	batch 3
none of them	109	102	121
intention to do something	32	30	18
reasoning	26	32	22
successfully learned	21	23	23
could have been done differently	12	7	10
interpreted in a new way	4	5	7
taking another perspective	4	9	6

Table 4: Stability over batches

The proportions of the categories do not vary much in each batch, which may indicate that although the inter-rater reliability is small since each time the experiment is replicated the sentences will receive different annotations, the proportion of elements will stay relatively steady. One exception (in batch 3 of the element intention) may indicate that there are some fluctuations, which should be considered in an experiment with a bigger sample size.

The next table reports on the proportions of reflective elements in the dataset. Based on the sentences, which received the majority vote, it shows for all ele-

ments its frequencies and its percentages. In addition, it shows the frequencies for personal and non personal sentences.

	all	%	personal	non personal
could have been done differently	29	4.65	18	11
reasoning	80	12.84	42	38
taking another perspective	19	3.05	13	6
interpreted in a new way	16	2.57	13	3
intention to do something	80	12.84	53	27
successfully learned	67	10.75	43	24
none of them	332	53.29	113	219

Table 5: Number of sentences in each category

Focusing on the six elements of reflection, the categories "reasoning", "intention to do something", and "successfully learned" are present in more than 10% of the sentences. Besides these relatively frequent, but still rare elements, "interpreting something in a new way", "taking another perspective", and the recognition that "something could have been done differently" occur in two to five percentage of the sentences. The category "none of them" has the highest percentage. "None" does not necessarily mean that this category represents a "not reflective" category. It might include sentences, which are reflective, but not captured in the six items representing reflective elements. The chosen six elements of reflection are not an exhaustive model of reflection capturing all possible ways of expressing reflection in texts.

In order to test the influence of personal pronouns in reflective sentences, the original sample contained the same amount of sentences with personal pronoun as those without. The table shows that personal sentences have higher category counts for all reflective elements than non personal sentences. This might indicate that sentences containing a personal pronoun are more likely to be annotated as one of the reflective elements.

Assuming that there are relatively constant proportions of reflective elements, one area of application might be the analysis of courses regarding the baseline proportions of reflective elements. Compared to this experiment however, more data from several domains should be taken into account in addition to a more fine grained model of elements of reflection.

The next table serves as an example for this possibility. It inspects the proportions of reflective elements over the five courses of the data set. The amounts of sampled sentences from each course vary. The number of sentences for the first eLearning course was 196, from the second eLearning course was 111, from the social work level 2 course was 146, level 3 course was 39, and from the science course 131. As the social work level 3 course has a much lower sentence count than the other courses, it was excluded from the results. The courses were balanced according to their number of sentences. The following table shows the balanced percentages of elements of reflection for each course.

	%el1	%el2	%swl2	%sci	%all
could have been done differently	3.1	5.4	4.8	6.9	4.7
reasoning	13.3	9.9	17.8	9.9	12.8
taking another perspective	2.6	2.7	4.8	2.3	3.0
interpreted in a new way	3.6	1.8	2.1	2.3	2.6
intention to do something	12.2	10.8	14.4	10.7	12.8
successfully learned	11.2	12.6	8.2	10.7	10.8
none of them	54.1	56.8	47.9	57.3	53.3

Table 6: Percentage of balanced sentences for elements of reflection per course. el1 and el2: elarning courses; swl2 and swl3 is the social work course level 2 and level 3; sci: science course

The last column contains the previously reported percentages of the whole corpus. The courses with more sentences will have more influence on the overall percentages. Notable is the 9.4% difference between the social work level 2 course and the science course on the element "none". This means that the former course has nearly for each 10 sentences an additional reflective sentence.

8 Discussion

Based on the intuition that reflection is a rare good (valuable but does not occur frequently), this research provides evidence about the proportions of reflective elements in texts. Based on the literature review, research indicates that the proportions vary on the study level, level of reflection, and elements of reflection. This research presents the proportions of reflection, based on the unit of sentences, from six elements of reflection. It concludes that these elements of reflection are indeed rare, and that some of the elements, for example the element "change of perspective", or "something was interpreted in a new way", are especially rare. In addition, sentences that contain a personal pronoun, are rated as having higher frequencies of reflection. This may indicate that sentences, in which the writer expresses a personal view, are more likely to be rated as reflective.

In addition, the number of sentences of reflective elements varies between courses. While these results are interesting for the comparison of courses, the results have to be taken cautiously, as a bigger sample size would be necessary to carry out a thorough analysis of the elements of reflection. This would be needed to balance the cells containing a small number of sentences.

Compared to the research outlined in the theory part, this research shows proportion of reflective elements from course forum posts, that did not focus on developing reflective writing skills. This may help as a reference to other courses that are especially designed to enhance reflective writing.

This research used sentences as the unit of analysis. While this decision helps to calculate percentages of elements of reflection based on the total sentence count of a text, it bears its own problem. Certain elements may not be able to

be captured in one sentence, and thus the raters might have annotated it as none, although from the wider context it would belong to an element of reflection.

As stated, the studied elements of reflection were not exhaustive to describe all forms of reflection. Further research with other elements, and other operationalisations of the elements would help to extend this research. Furthermore, it may be fruitful to study text corpora with different contexts. This could help to determine, which context factors stimulate reflection or hinder it.

The used approach however allows to quantify the proportions of reflective elements, and it indicates that reflection is a rare good.

References

[1] Dyment, J.E., O'Connell, T.S.: Assessing the quality of reflection in student journals: a review of the research. Teaching in Higher Education 16, 81–97 (Feb 2011), http://www.tandfonline.com/doi/abs/10.1080/13562517.2010.507308

[2] Gulwadi, G.B.: Using reflective journals in a sustainable design studio. International Journal of Sustainability in Higher Education 10(2), 96–106 (Oct 2009), http://www.emeraldinsight.com/journals.htm?articleid=1776291&show=abstract

[3] Hatton, N., Smith, D.: Reflection in teacher education: Towards definition and implementation. Teaching and Teacher Education 11(1), 33–49 (Jan 1995), http://www.sciencedirect.com/science/article/pii/0742051X9400012U

[4] Korthagen, F., Vasalos, A.: Levels in reflection: core reflection as a means to enhance professional growth. Teachers and Teaching: Theory and Practice 11, 47–71 (Feb 2005), http://www.tandfonline.com/doi/abs/10.1080/1354060042000337093

[5] Krippendorff, K.H.: Content Analysis: An Introduction to Its Methodology. Sage Publications, Inc, third edition edn. (Apr 2012)

[6] Moon, J.A.: A handbook of reflective and experiential learning. Routledge (Jun 2004)

[7] Plack, M., Driscoll, M., Blissett, S., McKenna, R., Plack, T.: A method for assessing reflective journal writing. Journal of allied health 34(4), 199–208 (2005)

[8] Ross, D.D.: First steps in developing a reflective approach. Journal of Teacher Education 40(2), 22–30 (Mar 1989), http://jte.sagepub.com/content/40/2/22

[9] Scanlan, J.M., Chernomas, W.M.: Developing the reflective teacher. Journal of Advanced Nursing 25(6), 1138–1143 (Jun 1997), http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2648.1997.19970251138.x/abstract

[10] Thorpe, K.: Reflective learning journals: From concept to practice. Reflective Practice 5(3), 327–343 (2004), http://www.tandfonline.com/doi/abs/10.1080/1462394042000270655

[11] Ullmann, T.D.: An architecture for the automated detection of textual indicators of reflection. In: Reinhardt, W., Ullmann, T.D., Scott, P., Pammer, V., Conlan, O., Berlanga, A. (eds.) Proceedings of the 1st European Workshop on Awareness and Reflection in Learning Networks. pp. 138–151. Palermo, Italy (2011), http://ceur-ws.org/Vol-790/

[12] Ullmann, T.D., Wild, F., Scott, P.: Comparing automatically detected reflective texts with human judgements. In: 2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning. Saarbruecken, Germany (Sep 2012), http://ceur-ws.org/Vol-931/paper8.pdf

[13] Wald, H.S., Borkan, J.M., Taylor, J.S., Anthony, D., Reis, S.P.: Fostering and evaluating reflective capacity in medical education: Developing the REFLECT rubric for assessing reflective writing. Academic Medicine 87(1), 41–50 (Jan 2012), http://journals.lww.com/academicmedicine/Abstract/2012/01000/Fostering_and_Evaluating_Reflective_Capacity_in.15.aspx

[14] Williams, R.M., Wessel, J., Gemus, M., Foster-Seargeant, E.: Journal writing to promote reflection by physical therapy students during clinical placements. Physiotherapy Theory & Practice 18(1), 5–15 (Mar 2002)

[15] Wong, F.K., Kember, D., Chung, L.Y.F., Yan, L.: Assessing the level of student reflection from reflective journals. Journal of Advanced Nursing 22(1), 48–57 (Jul 1995), http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2648.1995.22010048.x/abstract