Manual for the data collection and analysis of self- and peer assessment via think aloud protocols

Version April 2023

Ernesto Panadero

Javier Fernández-Ruiz

Leire Pinedo

(ERLA Group)

For more information contact: ernesto.research@gmail.com

Acknowledgements: Nathan Rickey for English editing (version April 2023).

Manual for the data collection and analysis of self- and peer assessment via think aloud protocols

1. Aim

This manual describes how the ERLA group has been coding the think aloud data obtained in self-assessment studies (Panadero et al., 2020, 2022, in review A, in review B) and how we plan to do it in future peer assessment studies. Beyond this manual, the higher aim is to open the black box of self and peer assessment. For decades, research has conceptualized self and peer assessment as strategies on their own without exploring learners' specific actions that comprise these processes (Panadero et al., 2016). Only through investigating that black box can we really understand what a student is doing when asked to, for example, self-assess.

Acquiring such knowledge would provide us with accurate and specific information on what are the specific aspects of peer and self-assessment and, even further, what is the role of individual differences in shaping these processes. Our team has presented these ideas and data in a number of studies (Panadero et al., 2020, 2022, in review A, in review B) and keeps developing new studies to further investigate these processes. This line of work is also being pursued by other researchers (e.g., Yan & Brown, 2017), but something currently unique to our work is the type of data we collect. We believe that, to understand self and peer assessment, we have to capture the actions of learners while they are engaged in these processes, which implies data collection methods such as eye-tracking, video-observation, physiological data, or thinking aloud. In this manual we focus on the latter.

2. Glossary

2.1. Think-aloud protocol: A scientific method that involves the concurrent verbalization of thoughts while performing a task (Ericsson & Simon, 1993). It is used with the aim of obtaining data on the cognitive, emotional and motivational processes underlying task performance. Another version of think aloud is called stimulated recall in which the participants reflect out loud about actions performed at an earlier occasion. This is sometimes called retrospective think aloud.

2.2. Process: General, content-independent actions performed by the students during their self- or peer assessment (e.g., Rate). In our studies, we have identified and categorised six general processes that students perform across different tasks when self-assessing: Read, Recall, Compare, Rate, Assess, and Redo. Information about each process is available at section 4.

2.3. Level: Some processes (i.e., Assess and Redo) can be performed at two levels of expertise. These are distinguished by adding (level 1, simple) or (level 2, complex) after the name of the process.

2.4. Strategies: Specific, content-dependent actions performed by the students during their self- or peer assessment (e.g., compare question-response). They differentiate from the processes as they are more specific and content dependent.

2.5. Assessment criteria: Standards used by the students to assess their performance. They can be grouped by level of complexity.

2.6. Profile: Classification of different ways of performing self- or peer-assessment, based on the differences among the type and level of the processes performed. In our self-assessment studies, we have identified and categorised four different profiles:

No Self-Assessment (NSA)

Superficial Self-Assessment (SSA)

Intermediate Self-Assessment (ISA)

Advanced Self-Assessment (ASA).

Information about each profile is available at section 5. Data for peer assessment still needs to be collected and analyzed to see whether the classification is valid for peer assessment too.

2.7. Flowchart: Visual illustration of the self- or peer assessment processes performed by the students. It can be used to represent the processes of one student, or to integrate several students' processes in the same figure if they belong to the same profile. More information is available in section 6.

3. Data collection process

3.1. General description of the process: Students are asked to think aloud while conducting self or peer assessment of an academic task (e.g., written essay, mathematical exercises). As students might be unfamiliar with the think aloud process, it is recommended to ask them to first perform it with a task they are familiar with. A typical example we use is asking the students to count the windows in their homes explaining out loud the mental steps they are performing.

Importantly, in some of our studies we explored the effects of feedback or tools such as rubrics. Thus, we asked our participants to first perform an unguided self-assessment and then a second guided self-assessment with the aid of written feedback and/or a rubric.

We have not implemented time restrictions to perform the self-assessment. The same procedure can be easily applied to peer assessment, substituting students' own tasks with peers' tasks, which is something ERLA will do shortly.

3.2. Instructions to participants

Verbal instructions: Students are instructed to express their thoughts, feelings and motivation, and reminded that if they are silent, they will be prompted to think out loud. After the aid was provided, students are once again asked to talk about their thoughts and feelings concerning the aid and to repeat the think aloud process of self-assessing their task.

Written instructions: Students receive a page containing the instructions of the task they have to perform. This way, if the student asks the researcher for instructions, the researcher can direct the student to the written instructions in order to reduce irrelevant variance caused by differences in instruction delivery.

3.3. Frequent challenges during data collection

The student remains silent during the think-aloud protocol: It is common for students to stop thinking out loud at some point of the experimental phase. In our case, as we are working physically next to the student and following what they are doing, we are able to identify if the silence is due to a distraction (e.g., a noise in the corridor) or if the student is concentrated in the task. If it is the latter and the student remains silent for 30 seconds, we prompt them to explain what it is that they are doing at that particular moment as we are interested in that data.

The student asks what to do: First, the student is reminded of the written instructions that were giving at the beginning. That is usually enough for the vast majority of the students. If they ask specific follow-up questions, we try to deliver an answer in the most succinct way possible as it can be the case that the student is trying to get the work done by the researcher.

The student asks how to use the materials (e.g., rubric): A description of the materials is provided, as well as an indication of their use (e.g., people look at each row and see if the description matches their work). Importantly, we state that not using the material is also a valid option.

4. Data structure

4.1. General description of the database

To integrate all the data collected, each researcher can use their preferred software tool (e.g., Microsoft Excel, SPSS, R). The most important sources of data are those related with processes, strategies, and criteria. Using these variables, second-level variables (such as profiles) will be calculated. Next, a list with some of the main groups of variables will be provided.

General variables: These are usually situated at the beginning of the database such as student ID, educational level (primary, secondary or higher ed.), experimental condition (nominal), academic year (ordinal) and gender (nominal).

Strategies and criteria data

> *Self-assessment strategies:* A list of variables regarding self-assessment strategies are coded next. These strategies are different depending on the educational level, subject, and occasion (with or without aids). These strategies are coded dichotomously, using 0 if the student did not use such strategy, and 1 if they did. Importantly, a student can use several strategies. The complete list of strategies coded in our studies can be found in Appendix A.
>
> *Assessment criteria:* Similar to the strategies, assessment criteria may differ depending on the educational level, subject, and occasion (with or without aids). These criteria are also coded dichotomously, using 0 if the student did not use the criterion, and 1 if they did. The same student can use several criteria. The complete list of criteria coded in our studies can be found in Appendix A.

Profile data: this section of the data is used to identify the self or peer assessor profile.

> *Processes:* Processes are coded differently from the strategies and criteria. Instead of creating dummy variables for each process, we create four variables representing the first, second, third and fourth process carried out. It has to be done this way because we want to record the order in which the processes are carried out, not just their occurrence. After analysing close to 200 videos, we identified that students usually enact up to four different processes. Hence, we created four variables, but other approaches can be considered, as discussed in section 4.3. Next, we explain the six processes that have been identified.
>
> **Read**. The student reads the question, the answer, or any other part of the task.
>
> **Recall**. The student describes the question and their answer, without directly reading what is written and without making a judgement about the quality of the task.

Compare. The student compares different sources of information, usually between the question and their answer.

Rate. The student makes an estimation (e.g., good, bad, average) about the quality of specific parts of the task, or about the whole task.

Assess. The student assesses the quality of their work using different criteria. Usually, it follows the pattern of "I think the task is _____ because _____". Assessing has two levels of complexity. In the first level, the student uses assessment criteria which are not based in rules and instructions (e.g., "I think I did a poor job because I was nervous"). In the second level, the student uses criteria based on the task, workshop, or teacher's instructions (e.g., "I think I did a poor job because some of my paragraphs are not really coherent").

Redo. The student performs again the exercise or part of it. Redoing has two levels of complexity. In the first level, the student would think on different responses to the exercise, without actively modifying their response. In the second level, the student actively modifies their response.

Weighted processes: Four new variables are created next, to match the four processes but this time substituting the name of the process by its weight. More information on how processes are weighted in section 5.2.1.

Profiles

Qualitative profiles. Profiles were coded as ordinal variables: (1) No Self-Assessment (NSA), (2) Superficial Self-Assessment (SSA), (3) Intermediate Self-Assessment (ISA) and (4) Advance Self-Assessment (ASA). In our studies, we calculated as many profiles as occasions of self-assessment (e.g., self-assessment of the Spanish task without aid, self-assessment with aid, self-assessment of the mathematics task without aid, self-assessment with aid). Additionally, we calculated a general profile as a self-assessor considering the more specific profiles. More information about the process in section 5.1.1.

Quantitative profiles. These variables are related to the qualitative profiles and are also coded from 1 (NSA) to 4 (ASA). However, these quantitative profiles are calculated by scoring the different processes as coded qualitative profile. Meaning that a particular process (e.g., assess) is given a score. Therefore, the main difference from the qualitative profiles is that here a quantitative score is reached. Importantly, these categories are in correspondence with their qualitative homologous, meaning that for each qualitative profile we also calculated a quantitative profile (i.e., general, self-assessment of the Spanish task without aid, with aid, of the mathematics task without aid and with aid). More information about the calculation of quantitative profiles in section 5.2.

Other: Questionnaire or interview data can be also included in the database, coded following the usual procedures.

4.2. General recommendations

Naming the variables: We divide between profile variables and processes variables. First, regarding profile variables (see "qualitative profiles" in the previous section), the general profile can just be named "GENERAL_PROFILE". For other profile variables, we recommend the sequence "SUBJECT_OCCASION" (e.g., SPANISH_WITHOUT_FB). For processes variables, the following naming was used "OCCASION/SUBJECT/ORDER". For occasion, one indicator has to be created to differentiate self-assessment with and without aid (e.g., N for self-assessment without aid and Y with aid). For subject, another indicator has to be created to differentiate between subjects if necessary (e.g., S for Spanish and M for Maths). For order, one last indicator has to be created to identify the order in which the process was performed (e.g., 1, 2, 3, or 4). For weighted processes, one last indicator can be added at the end of the variable columns. Therefore, the variable corresponding to the third process performed in the self-assessment of the mathematics task with aids would be named "YM3". For strategies and criteria related variables, the pattern "OCASSION/SUBJECT/STRATEGY OR CRITERIA NAME" can be used. An example can be found in the Excel appendix.

Ordering the variables: We recommend the following order of variables in the database: General variables (e.g., gender, age) -> Qualitative profile data -> Quantitative profile data -> Qualitative processes -> Quantitative processes -> Strategies -> Criteria -> Other data. Different orders can be considered by other researchers.

4.3. Decisions to make

Number of processes coded: As explained previously (section 4.1.5), we observed that students usually carried out up to four processes before repeating the cycle. Therefore, we created four variables representing the first, second, third and fourth processes carried out. It is possible, however, that in other contexts (e.g., more experienced students) the cycles become more complex and more variables are needed. Another option, if the order of the processes is not considered a relevant variable, is to create six dummy variables (one per process), as was done in our studies with the strategies and criteria.

Repeated processes: A student can enact the same process multiple times during a self-assessment. How to manage these occasions depends on the focus of the study. If the aim is to investigate the temporal sequence of processes, then all multiple sequential uses should be registered. On the other hand, if the aim is to investigate just what are the processes used regardless of the order of use or repetition, then coding it just one time is an appropriate method.

For example, for the first aim, a sequence could be Read -> Rate -> Assess -> Rate. For the second aim, just registering that the student perform the processes of Read, Assess and Rate would be enough.

Levels overlapping: A student can perform the same process in both levels of complexity (e.g., a participant may enact Assess at level 1 first and then subsequently enact Assess at level 2). As in the previous case, our recommendation depends on the focus of the study. If the aim is to analyse the level of complexity of the self-assessment process, then the same process performed with both levels of complexity should be coded independently. If the aim is to investigate just which processes are used regardless of their level of complexity, then they can be coded once.

5. Data analysis

5.1. Qualitative profile calculation

We use a standardized procedure to assign students to a profile. Next, a description of the procedure is presented.

Key processes: The assignment of students to the profiles is based in the analysis of three specific processes (i.e., Rate, Assess and Redo) as these have been shown to be the key processes that are distinguishable among profiles.

> First, if the student does not perform any of these three processes, she is assigned to No Self-Assessment.
>
> Second, if the student estimates the quality of the task, but does not properly assess or redo it, then she is assigned to Superficial Self-Assessment. In general, the students in this category judge the quality of the performance (e.g., good, bad, average) but do not explain the criteria to reach such judgement.
>
> Third, in the Intermediate Self-Assessment category, students estimate the quality of the task and assess it based on some assessment criteria. Here, while the student is able to perform a complete self-assessment employing assessment criteria (generally based on teacher's instructions), they still do not perform any process to improve the work's quality or correct the previous performance.
>
> Lastly, a student is categorized as Advanced Self-Assessment if she redoes parts or all the task. This can be done only through thinking aloud or by actually performing changes in the document itself. These students usually perform all the other key processes (e.g., Assess), but some might not perform all and because they redo the task they are also categorized here. Ultimately, the most salient feature of these students is that they are able to perform a complete self-assessment of their previous performance and improving it by fixing the mistakes detected.

Special cases: We have identified that the academic year and level of complexity are variables that need to be taken into special consideration when assigning participants to a profile. In some conflicting cases, the processes need to be reviewed by two or more researchers to reach an agreement.

> **Academic year**. As our sample was comprised of secondary and higher education students, this division was used to determine profile assignment when raters disagreed. When in doubt, secondary education students were assigned to a higher

profile. On the contrary, higher education students were assigned to a lower profile when in doubt.

Level of complexity. The level of complexity of the processes of Assess and Redo needs to also be considered. When in doubt, students who perform these processes with a level 2 of complexity are assigned to a higher profile than those with a level 1.

5.2. Quantitative processes calculation

Weights: To calculate a quantitative profile based in the quantitative coding of the processes—see above section 5.1—a numerical value has to be assigned to each of the six processes. After analyzing the videos and exploring several numerical value options, we propose the values shown in Table 1 as an appropriate solution. Obviously, other values can be considered (e.g., different scorings for Read and Recall, or different scorings for Assess Level 2 and Redo Level 1). However, based on our experience analyzing these types of data, the final decision was to use the scorings shown at the table above.

Table 1.

Processes' scores

Process	Score
Read	**0,5**
Recall	**0,5**
Compare	**0,75**
Rate	**1**
Assess (Level 1)	**1**
Asses (Level 2)	**1,5**
Redo (Level 1)	**1,5**
Redo (Level 2)	**2**

Total and weighted calculation: Students using a larger number of processes, even of lower quality, could be benefited when calculating their quantitative profile. For example, imagine a student that reads, recalls, compares, rates, and assesses at level 1 (total 3,75), vs. a student that only assesses at level 2 and redoes at level 1 (total 3,50). Thus, to eliminate the bias towards those students carrying out more processes, it is recommended to calculate a weighted score. This is done by dividing the total score by the number of processes performed.

5.3. Quantitative profile calculation

Thresholds: The quantitative profiles are calculated using the weighted scores. Based on our previous analyses, the following cut-off points have been calculated and are the ones we recommend to use:

From **0 to 0,5** – Mostly No Self-Assessment
From **0,51 to 0,92** – Mostly Superficial Self-Assessment
From **0,93 to 1,35** – Mostly Intermediate Self-Assessment
From **1,36 to 3** – Mostly Advanced Self-Assessment

5.4. Quantitative validation of profiles

There are several quantitative analyses that can be performed to test the validity of the qualitative identification of the profiles. It is helpful to perform these analyses to evaluate the qualitative coding performed in previous steps[1]. So far, we have used three data analysis procedures, namely cross-tables, discriminant analysis, and ANOVAs. Next, these are presented briefly and we exemplify the results we obtained in each of them with data from Panadero et al. (In review).

Cross-tables: It is a type of table in a matrix format that displays the frequency distribution of the variables/categories. They are also known by other names such as cross tabulation, crosstabs, or contingency table. They show how two variables are related and can help identify any interactions between them.

> *In our data we used all cases, divided by subject, educational level, and occasion. Five of the six crossed tables obtained a Kappa scoring above 0.6, which indicates an acceptable relationship between qualitative and quantitative profiles*.

Discriminant analysis: This method is used to find a linear combination of features that characterizes or separates two or more classes of objects or events.

> *In our data, we calculated four discriminant analyses. First, analysis of Spanish without feedback revealed three discriminant functions, the first of which alone explained 73.7% of total variance, canonical R^2 = .79 and significantly differentiated among profiles, Λ = .08, x^2 (24) = 426,05, p = .000. Second, Spanish with feedback revealed three discriminant functions, the first of which alone explained 67.3% of total variance, canonical R^2 = .75 and significantly differentiated among profiles, Λ = .09, x^2 (24) = 399.90, p = .000. Third, Mathematics without feedback revealed three discriminant functions, the first of which alone explained 71.3% of total variance, canonical R^2 = .60 and significantly differentiated among profiles, Λ = .24, x^2 (24) = 80.98, p = .000. And,*

[1] Please note that there would be a lot to consider when carrying out this step. See for example Creswell, J. W., & Plano-Clark, V. L. (2018). Designing and conducting mixed methods research (3rd. ed.). Sage. As it is not our intention to discuss these considerations in detail here, we just want to state that our thinking is more complex that what is written.

fourth, Mathematics with feedback revealed three discriminant functions, the first alone explained 57.2% of total variance, canonical R^2 = .78 and significantly differentiated among profiles, Λ = .04, x^2 (21) = 174.07, p = .000. All these results reinforce the previous results in that the four qualitative profiles we identified are also found statistically via discriminant analysis.

Analysis of variance (ANOVA) over dependent variables: It is a set of statistical models and their related methods (such as the "variation" within and between groups) that are used to compare the differences among averages.

In our data, we ran a one-way ANOVA using performance (secondary Spanish, university Spanish, and secondary mathematics) as the dependent variable and general qualitative profile (four categories) as the independent variable. There was a significant effect of the general profile F(4, 226) = 2.80, p = .041, η^2 = .036. Pairwise comparisons post hoc revealed that students in Superficial SA profile (M = 6.22 SE = .23) showed a significantly (p = .027) lower academic performance than students in Intermediate SA profile (M = 6.86 SE = .17), as well as significantly lower performance (p = .01) than students in Advanced SA profile (M = 7.22 SE = .32). Students in No SA profile (M = 6.31, SE = .50) did not differ significantly from the other groups. We interpret these results as another type of discriminant validation for the qualitative profiles.

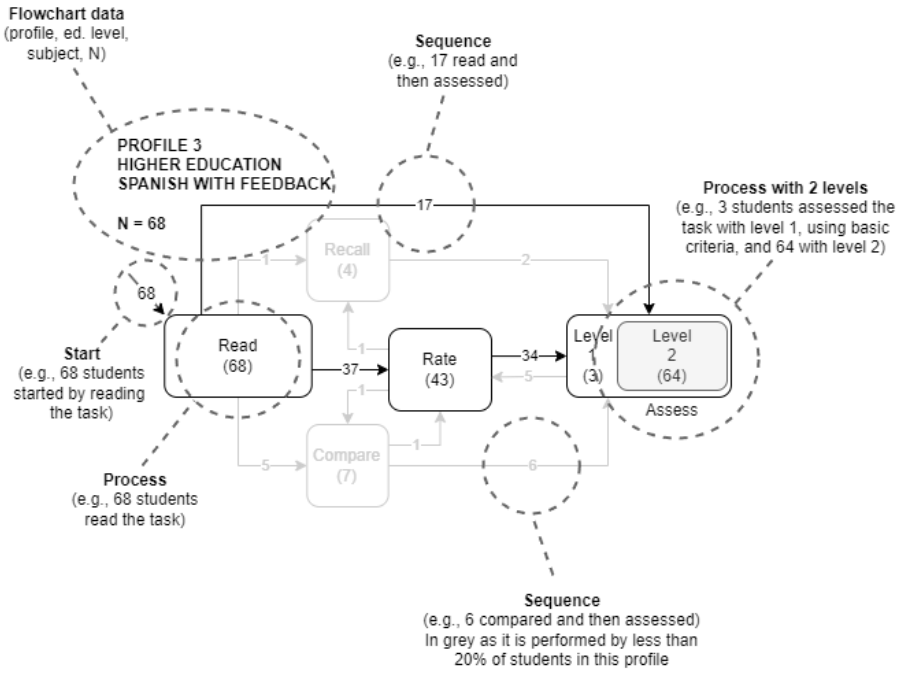
6. Data reporting and visualization

6.1. Flowchart figures

The flowcharts are central for interpreting the qualitative data to understand the differences among profiles. Thus, we encourage future users of this manual to include them. The flowchart should visualize the self or peer assessment processes and the temporal relationship among them (e.g., order or execution). Figure 1 is an example of a flowchart we have used, containing very detailed information that is crucial for the reader to understand what happens. We recommend using two colours in the figure: black and pale grey, the latter indicating that those processes are performed by less than 20% of the students in that particular profile. For the design of the figures, we recommend specific figure creation software, such as Draw.io or Tableau. Next, we describe the elements of the flowchart.

Figure 1

Example of the flowchart figure



Arrowed lines: The arrows represent the sequence of processes (i.e., order of the processes). In each arrow, there is a number which represents the number of students that followed such sequence. Some arrows do not have an origin (e.g., in figure 1 the arrow with 68 which is the start). Arrows with no origin indicate the number of students that started the self-assessment in such particular process.

Boxes: The boxes contain the name of the process, and the frequency represents the number of students that performed such process.

Levels: For the processes with more than one level (e.g., Assess), the number of students in each level are represented with Level 1 (in white) and Level 2 (in grey).

7. Conclusions

As mentioned in the opening section, this manual describes how the ERLA group has been coding the think aloud data obtained in self-assessment studies. Our aim is to also apply this method to future peer assessment studies in the coming months.

The manual will probably be further developed, so please keep in mind the version you are using and whether there is an updated one. Further developments will probably include more examples, application to peer assessment, and the integration with behavioural data such as eye tracking and electrodermal activity. However, we believe the current version, focusing on thinking aloud, represents a significant step for self-assessment research.

References

Hevey, D. (2010). Think-aloud methods. In N. J. Salkind (Ed.), *Encyclopedia of research design (pp. 1505-1506)*. SAGE Publications, Inc., https://dx.doi.org/10.4135/9781412961288.n460

Panadero, E., Brown, G. T. L., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review, 28*(4), 803-830. https://doi.org/10.1007/s10648-015-9350-2

Panadero, E., Fernández-Ruiz, J., & Sánchez-Iglesias, I. (2020). Secondary education students' self-assessment: The effects of feedback, subject matter, year level, and gender. *Assessment in Education: Principles, Policy & Practice, 27*(6), 607-634. https://doi.org/10.1080/0969594X.2020.1835823

Panadero, E., García-Pérez, D., Fernández-Ruiz, J., Fraile, J., Sánchez-Iglesias, I., & Brown, G. T. L. (2022). University students' strategies and criteria during self-assessment: Instructor's feedback, rubrics, and year level effects. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-022-00639-4

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation In Higher Education, 42*(8), 1247-1262. https://doi.org/10.1080/02602938.2016.1260091

APPENDIX A.

Strategies		Processes	Criteria	
Secondary Ed.	Higher Ed.		Secondary Ed.	Higher Ed.
Read the question		**Read**		
Read the response	Read the essay			
Read the text				
Read/process the feedback received	Read the feedback/ rubric received			
	Remember the instructions	**Recall**		
	Remember the seminar			
Compare text-response		**Compare**		
Compare question-response	Compare instructions-essay			
Access their memory to compare				
Compare own evaluation to feedback	Compare essay to feedback or rubric			
		Rate		
Review the signs (operators)		**Assess (Level 1)** *Subjective (guts-feeling) indicators*	Without clear criteria	Without clear criteria
			Intuition	Negative intuition
				Positive intuition
			Hindsight	Negative hindsight
				Positive hindsight
			Experience/self-efficacy	
Replace the X		**Assess (Level 2)** *Compliance with external indicators*	Rules	Instructions
			Spelling	Spelling
			Mistakes identified	
			Given by the teacher	Feedback received
			Comparative	
				Writing process
				Paragraph structure
				Sentences and punctuation marks
Evaluate the procedure followed			Adequacy to the question	
			Coherence of the result	
			Steps followed	
			Formula application	

Think of different responses	Think of different responses	**Redo (Level 1)**
Perform the exercise again	Perform the exercise again	**Redo (Level 2)**