

Combining a theoretical framework and a statistical measure to assess the reliability of didactic criteria in the analysis of large corpora

Ismail Mili¹, Mickaël Da Ronch² and Marie-Line Gardes³

¹University of Fribourg, Faculty of Educational Sciences, Fribourg, Switzerland;
ismail.mili@edufr.ch

² University of Teacher Education Valais (HEP-VS), Saint-Maurice, Switzerland

³ University of Teacher Education Vaud (HEP-VD), Lausanne, Switzerland

The aim of this article is to present the articulation of a theoretical framework and a statistical measure using Fleiss' Kappa to assess the reliability of didactic criteria. These didactic criteria have been developed within the framework of the analysis of a corpus of consistent resources and aim to evaluate the potential of problems enabling students to develop a mathematical activity of research. The results show that the criteria constructed are reliable, with significant inter-coder agreement, and provide interesting perspectives for the analysis of resources in the context of teacher training.

Keywords: Mathematical activity, problem, data analysis, reliability, Fleiss' kappa.

Background and research question

Internationally, recommendations from research and the field suggest that the teaching and learning of mathematics should take place through problem solving, preferably linked to the activity of the mathematician (see in particular OECD, 2014; Ministère de l'Éducation Nationale, 2019; National Research Council, 2001). But what is the reality? Do the resources available to teachers encourage such activity? And how can we identify that a teaching resource has the potential to encourage mathematical research activity? Our contribution aims to propose a theoretical framework and a methodological tool for analysing the potential of resources to encourage mathematical activity.

This contribution follows on from an article published at CERME 13 (Da Ronch et al., 2023), in which initial criteria for assessing the potential of school level problems were defined. Since that prospective study, these criteria have been revisited, broadened and refined, and the corpus studied extended. The research question now focuses on methodological aspects: can these criteria be considered “reliable” (in the sense of concordance between ratings made by several coders, see Sim & Wright, 2005)?¹

To do this, we begin by defining mathematical research activity and the notion of problem. In the second part, we detail the development of criteria for identifying, within a teaching resource, whether a mathematical problem is likely to encourage mathematical research activity. In the third part, we put these criteria to the test by applying them to the analysis of a corpus from French-speaking Switzerland, in order to test their reliability. Finally, in the fourth part, we open up perspectives on

¹ In our view, this is an essential step before testing the reliability of the criterion in terms of predicting student activity.

measuring the validity and reproducibility of these criteria, as well as their reinvestment in teacher training.

Theoretical Background

Da Ronch's (2022) epistemological study of the notion of mathematical activity as a human activity shows that for most mathematicians and mathematics didacticians, this activity is strongly linked to problem-solving (see Brousseau, 1997; Chevallard, 1998; Da Ronch, 2022; Gardes 2013; Halmos, 1980; Perrin, 2007; Thurston, 1994). During this activity, there is no question of limiting oneself to the use of definitions and theorems for the sole purpose of applying them to specific situations (Brousseau, 1997). Neither is the production of a mathematical result the ultimate goal of this activity. It is above all the processes involved in solving the problem that are indicative of genuine mathematical activity. The finished product, *i.e.*, the result of the problem, is in itself only a secondary objective of this activity (Chevallard, 1998). Consequently, mathematical activity involves different processes leading to the elaboration of a finished product, as the result on the problem, which are linked to the dialectics, which Brousseau qualifies, of action, formulation and validation. This means proposing problems that allow us to articulate, during these phases, “the reasoning of plausibility, purely inductive, and that of the necessities attached to deductive reasoning” (Da Ronch, 2022, p. 11) in the search for mathematical truth and its explanation in order to answer the problem. Doing mathematics therefore requires solving problems and accepting a certain form of scientific responsibility in the face of the challenge of truth and the need for proof (Da Ronch, 2022; Da Ronch et al., 2023). These different dialectics emerge if, and only if, specific knowledge is mobilized in the resolution of a problem (e.g., Da Ronch, 2022; Da Ronch et al., 2023). For example,

[...] the mobilization of specific knowledge linked to the practice of the mathematical activity. For example, entering into the resolution of a problem by studying particular cases is a knowledge linked to the experimentation process. The formulation of conjectures resulting from the study of these cases, the change of the register of representation of the mathematical objects at stake, sometimes leading to modelling premises, are examples of knowledge linked to the formulation process. Finally, validation is based on hypotheses (e.g., law of excluded middle, law of noncontradiction) and rules of logic (e.g., *modus ponens*, *modus tollens*) which allow, thanks to different mathematical reasoning (e.g., direct implication, contradiction, induction, counterexample...) to enter into the argumentation and proof process and thus to rule on the truth or falsity of a mathematical statement. This knowledge is therefore knowledge related, this time, to the validation process [...] (Da Ronch et al., 2023, p. 97).

Furthermore, the processes of action, formulation and validation and their associated knowledge are indicative of a real articulation between inductive and deductive reasoning, testifying to a genuine experimental approach in mathematical activity (see, Da Ronch, 2022; Gardes & Durand Guerrier, 2016; Giroud, 2011; Perrin, 2007).

These processes are visible only through the encounter with a problem and the proven need to solve it for those who intend to investigate it (Da Ronch, 2022). The etymological and epistemological study carried out in Da Ronch's work (2022, chap. 2) has enabled us to characterize a problem according to a double aspect, syntactic and semantic. The syntactic aspect concerns the structure of

the problem presented in the form of a general question and a set of instances, in reference to Garey and Johnson (1979).

For our purposes, a problem will be a general question to be answered, usually possessing several parameters, or free variables, whose values are left unspecified. A problem is described by giving: a general description of all its parameters, and a statement of what properties the answer, or solution, is required to satisfy. An instance of a problem is obtained by specifying particular values for all the problem parameters. (Garey et Johnson, 1979, p. 4)

The general question refers to the existence of a search variable (Grenier & Godot, 2004; Da Ronch, 2022), *i.e.*, a problem parameter whose value is not fixed and which is therefore the responsibility of the problem solver.

The semantic aspect of a problem refers to its epistemological quantity, determined by a detailed mathematical analysis of the problem (Da Ronch, 2022). This quantity can be identified by proximity to other related problems and by changes in the scope of the question and instances of the problem studied. The proofs used are also indicators of proximity to other problems that mobilize similar invariants in their resolution process (Da Ronch, 2022; Giroud, 2012; Da Ronch et al., 2023). All this highlights the non-isolated nature of the problem and its richness. To carry out such an analysis in order to determine the epistemological quantity of a problem, we pointed out a deficiency in the method of conduct to be adopted. Indeed, during the preliminary analysis (Artigue, 2015), no method of conducting a mathematical analysis with a view to building solid didactic engineering, with real consistent problems, leading to real mathematical activity was made explicit. This has led to the development of a method for carrying out such analyses (see Da Ronch, 2022; Da Ronch & Gravier, 2024). However, this method is relevant for the construction of didactic engineering, but remains too complex and not very operationalizable to evaluate the mathematical potential of problems on large corpora of data. Indeed, it is not possible to carry out such fine analyses for a wide variety of problems. This calls for the development of a new pragmatic evaluation tool that can effectively analyse a large corpus of problems for their potential in the implementation of mathematical activity.

Development of didactic criteria

In order to determine the potential of a problem to foster mathematical activity as we have defined it, we use the concepts of research variable (Grenier & Godot, 2004; Da Ronch, 2022) and didactic variable (Brousseau, 1997).

As mentioned in the previous section, a research variable is a problem parameter that is not initially fixed in the statement. It is up to the problem-solver to set different values for this variable. For example, if we wish to determine whether a rectangle of size $p \times q$ can be tiled with 1×2 dominoes, the search variable here is the size of the grid, which is not fixed in the statement. It is therefore up to the subject to appropriate it, *i.e.*, to set its own values in the process of solving the problem. The existence of a search variable thus guarantees that the problem is syntactically correct and semantically rich. When solving the problem, the problem-solver's mathematical activity is focused on the study of particular cases (as sub-problems) and then on generalization.

A didactic variable is a parameter of a learning situation that the teacher can vary to influence the way students interact with the proposed problem. The choice of a didactic variable value by the teacher can thus enable him or her to encourage mathematical activity in the pupils, in particular to take them gradually from the study of particular cases to the study of the generalization of a problem (see Figure 2).

The analysis of a problem's potential to encourage mathematical activity is based on a four-level criteria, relating to the nature of the variables present in the problem statement. We have formulated and coded them as follows:

- *Code 3: existence of a search variable in the problem statement*

For example, in the exercise Ribambelle¹ (see Figure 1a), the research variable is the number of “men” (where “man” is the basic figure created with the matches and then reproduced iteratively). It's up to the student to make this number of men vary, in order to study a few special cases that will enable him to determine a general formula.

- *Code 2: presence of a didactic variable with several values set by the problem statement*

For example, in the exercise Les étapes² (see Figure 1b), the number of steps is a didactic variable that the statement makes vary (20, 50, 100) to encourage students to study particular cases in order to determine the general formula.

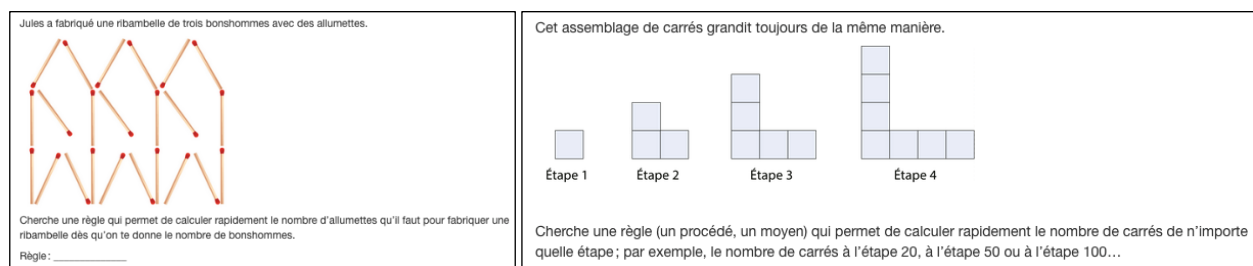


Figure 1: Statement of “Ribambelle”² (1a) and “Les étapes”³ (1b) – fifth grade textbook in french-speaking Switzerland

- *Code 1: presence of a didactic variable with a value fixed by the problem statement, but which can be modified by the teacher.*

For example, in the exercise The string (see Figure 2a), a didactic variable is the length of the string. The teacher could vary this to study different cases and guide the students' activity towards a more general question, e.g. what are the triplets of three numbers such that the perimeter of the triangle is equal to the perimeter of the square?

² Jules made a string of three men out of matches. Find a rule to quickly calculate the number of matches needed to make a string once you've been given the number of men.

³ This set of squares always grows in the same way. Look for a rule (a process, a way) that allows you to quickly calculate the number of squares in any step; for example, the number of squares in step 20, step 50 or step 100...

- *Code 0: no didactic variable*

For example, the exercise Achats (see Figure 2b) is a verbally stated problem that can be solved with simple arithmetic operations. There is no didactic variable that would enable us to develop a mathematical research activity in our meaning of the term.

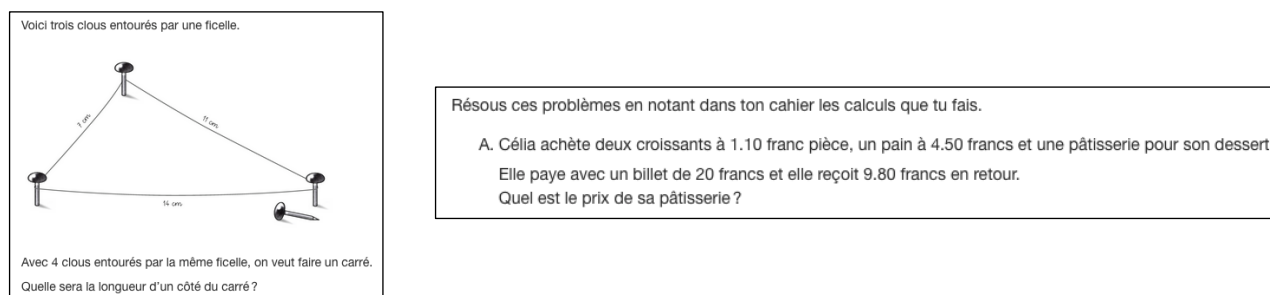


Figure 2: Statement of “La ficelle”⁴ (2a) – fourth grade – and “Achats”⁵ (2b) – sixth grade

Testing criteria on a corpus, choice and method

In addition to the fact that the authors work in Swiss institutions, the French-speaking Swiss corpus was chosen for the following three reasons. Firstly, because of the country's administrative structure, only one “official” teaching resource (textbook and book) is available to teachers: the Moyens d'Enseignement Romands (hereinafter MER). Although teachers have pedagogical freedom to supplement these MERs according to their needs, they are compulsory at ministry level, as they have been designed, developed and written to cover the subject content of the official curricula, grouped together in the Plan d'Études Romand (hereafter PER). This uniqueness of resources makes it possible to quickly get an overview, without having to analyse several different manuals. Secondly, some of the PER's injunctions are similar to our definition of mathematical activity. Indeed, in the “General comments” of the “Mathematics” domain presentation text, we find that

“[Mathematics] promotes [...] an attitude of research through trial-and-error, generalization, conjecture and validation. In this way, the practice of mathematics develops the ability to imagine strategies, and to organize and structure knowledge” (free translation, General comments on the MSN field, CIIP, 2024).

This is in line with the components of mathematical activity presented above. Other recommendations are along the same vein, such as the organization of scientific debates in mathematics, the formulation of questions, the development of reasoning, the use of trial and error to reconstruct a thought process and understand missteps, and the analysis of the relevance and limitations of a chosen model, among others (see “Contribution au développement des Capacités Transversales”, CIIP, 2024)

⁴ Here are three nails surrounded by a string. With 4 nails surrounded by the same string, we want to make a square. How long will one side of the square be?

⁵ Solve these problems by writing down the calculations you make in your notebook. A. Célia buys two croissants at 1.10 francs each, a loaf of bread at 4.50 francs and a pastry for dessert. She pays with a 20-franc bill and receives 9.80 francs in return. How much does her pastry cost?

Thirdly, a section of the MER entitled “Aide à la Résolution de Problèmes” (Problems solving support, hereinafter ARP), distinct and circumscribed and constituting around 1/5 of all the statements, is explicitly intended to “work on problem solving”. The editors justify this section by pointing out that “it is not enough to let students find problems and then correct them on the basis of those who have succeeded, for everyone to learn to solve problems”, and that it is “necessary to put in place real teaching in problem solving” (free translation - ARP comments, CIIP, 2024).

Using our criteria, we analysed the entire Cycle II ARP corpus (students aged 8 to 12). The analysis was carried out in parallel by the three authors on a corpus of 312 items, with an initial sampling of 10 items per level (over 4 levels) to calibrate the coding. Inter-coder similarity, an indicator of criterion reliability, was measured using the appropriate Fleiss' Kappa coefficient (see Fleiss, 1971). It measures the agreement between several coders when they classify items into discrete categories. The Fleiss' Kappa coefficient ranges from 0 to 1, where 0 indicates a match equivalent to that expected by chance, and 1 corresponds to a perfect match between coders. A medium match is considered to exist when $\kappa > 0.41$, and a large match when $\kappa > 0.61$. The fact that four categories are available increases confidence in the interpretation of the coefficient obtained. Moreover, testing it on a wide range of problems (312) reinforces this confidence (Fleiss, 1971; Landis & Koch, 1977; Sim & Wright, 2005).

Results and discussion of the reliability of the criteria used and research perspectives

The results of our study show the reliability of the didactic criteria obtained by calculating Fleiss' Kappa coefficient. The latter indicates a significant concordance value ($\kappa \approx 0.7$) according to Landis and Koch's (1977) table. Indeed, of the 312 problems coded in triple blind according to our four criteria, 44 obtained a different coding with each time exactly 2 coders in perfect agreement out of the 3 with only a deviation of 1 in the coding. Thus, 268 problems obtained perfect agreement in coding. Consequently, the results show that the criteria constructed are reliable, with significant inter-coder agreement. This paper has thus presented a reliable method for processing and analysing resources in order to identify the potential of problems enabling students to develop a mathematical research activity. This method of conducting research is based on epistemological and didactic criteria for categorizing problems in the study of a corpus involving a significant quantity of resources. Finally, our criteria-based processing and analysis method seems effective for analysing the potential of problems to generate mathematical activity in students, since it enables us to process a large corpus of data, without having to carry out an exhaustive *a priori* analysis of the problems to measure their epistemological quantity.

A first perspective provided by these results is to pursue the articulation of our theoretical framework with the use of statistical measures to question the validity of the didactic criteria developed, as well as the reproducibility of their use. Validity could be assessed through an analysis of student activity during problem solving, showing that the nature of the variables involved does indeed have an impact on students' mathematical research activity. With regard to reproducibility, initial results suggest that these criteria are sufficiently precise and operational for other coders to take them up and use them on different data corpora. However, this remains to be proven. A second perspective is to deepen the

analysis of the corpus studied. We can already state that this corpus includes few problems that allow students to develop a mathematical activity of research. Indeed, there is only one problem with a search variable (Ribambelle, see Figure 1a) and five problems with a statement proposing several values for a didactic variable. We can thus see that the MERs do not really enable teachers to implement the PER injunctions concerning problem-solving. Indeed, with the exception of these six problems, the statements do not propose any work on the formulation and proof of conjectures, generalization is little worked on, the issue of truth is absent, as are discussions on the existence of solutions, their multiple or universal nature. A third perspective is to consider the use of these didactic criteria in teacher training, for analysing resources or designing situations. Indeed, since the institutional resource only partially addresses the requirements of curricula regarding mathematical research activity, teachers must be able to identify problems that allow students to engage in mathematical activity. The developed didactic criteria could thus be a relevant tool for refining their didactic analyses.

References

- Artigue, M. (2015). Perspectives on design research: the case of didactical engineering. In A. Bikner-Ahsbals, C. Knipping, N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education: examples of methodology and methods* (pp. 467–496). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9181-6_17
- Brousseau, G. (1997). *Theory of the Didactical Situations in Mathematics*. Kluwer Academic Publishers. <https://doi.org/10.1007/0-306-47211-2>
- Chevallard, Y. (1998). Analyse des pratiques enseignantes et didactique des mathématiques : l'approche anthropologique [Analysis of teaching practices and mathematics didactics: the anthropological approach]. *Actes de l'UE de la Rochelle*, pp. 91–118.
- CIIP (2024). *L'aide à la résolution de problème (ARP) en 7e et 8e* [Problem-solving assistance in grades 7 and 8]. Plateforme des Moyens d'Enseignement Romands. <https://www.ciip-esper.ch/#/discipline/5/7/objectif/1002>
- Da Ronch, M. (2022). *Pratique de l'activité mathématique en médiation: modèles didactiques et conception d'ingénieries* [Practice of mathematical activity in mediation: didactic models and engineering design] (Doctoral dissertation, Université Grenoble Alpes [2020-....]). <https://cnrs.hal.science/tel-04089443/>.
- Da Ronch, M., Gardes, M-L., & Mili, I. (2023). Study of the potential of problems to practice a research activity in mathematics at elementary school in French-speaking Switzerland. In Drijvers, P., Csapodi, C., Palmér, H., Gosztönyi, K., & Kónya, E. (Eds.). *Proceedings of the Thirteenth Congress of the ERME (CERME13)* (pp. 96–103). Alfréd Rényi Institute of Mathematics and ERME. <https://hal.science/hal-04408292v1>
- Da Ronch, M., & Gravier, S. (2024). Didactical engineering: an approach for carrying out an epistemological analysis from research problems in mathematics. In *Pre-proceedings of the Fourth Conference of the International Network for Didactic Research in University Mathematics*

(INDRUM 2024, 7-14 June 2024) (pp. 629–630). Barcelona, University of Barcelona and INDRUM. <https://hal.science/hal-04608843>

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>

Gardes, M.-L. (2013). *Étude de processus de recherche de chercheurs, élèves et étudiants, engagés dans la recherche d'un problème non résolu en théorie des nombres* [Study of a research process for researchers, pupils and students involved in the research of an unsolved problem in number theory] (Doctoral dissertation, Université Claude Bernard-Lyon I). <https://theses.hal.science/tel-00948332/>

Gardes, M.-L., & Durand-Guerrier, V. (2016). Designation at the core of the dialectic between experimentation and proving: A study in number theory. In E. Nardi, C. Winsl w & T. Hausberger (Eds.), *Proceedings of the First Conference of the International Network for Didactic Research in University Mathematics (INDRUM 2016, 31 March-2 April 2016)* (pp. 286–295). Montpellier, France: University of Montpellier and INDRUM. <https://hal.science/hal-01337922>

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability*. 174. Freeman San Francisco.

Giroud, N. (2011). *Etude de la démarche expérimentale dans les situations de recherche pour la classe* [Study on the experimental approach in Research Situation for the Classroom] (Doctoral dissertation, Université de Grenoble). <https://theses.hal.science/tel-00649159/>

Grenier, D., & Godot, K. (2004). Research situations for teaching: a modelling proposal and example. *ICME-10, Denmark*.

Halmos, P. (1980). The heart of mathematics. *American Mathematical Monthly*, 87, 519–524. <https://doi.org/10.2307/2321415>

Landis, J., & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>

Minist re de l' ducation nationale. (2019). *Programmes d'enseignement de math matiques pour les cycles 2, 3 et 4* [Mathematics curricula for cycles 2, 3 and 4]. Minist re de l' ducation Nationale.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. National Academy Press.

OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (5). OCDE Publishing. <https://doi.org/10.1787/9789264208070-en>

Perrin, D. (2007). L'exp rimentation en math matiques [Experimentation in mathematics]. *Petit x*, 73(6), 6–34.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3), 257–268. <https://doi.org/10.1093/ptj/85.3.257>

Thurston, W. P. (1994). On proof and progress in mathematics. *Bulletin of the American mathematical Society*, 30(2), 161–177. https://doi.org/10.1007/0-387-29831-2_3