BEYOND PERFORMANCE RESULTS: ANALYZING THE INFORMATIONAL AND DEVELOPMENTAL POTENTIALS OF STANDARDIZED MATHEMATICS TESTS

PIETRO DI MARTINO, ANNA BACCAGLINI-FRANK

Over the past years the importance and use of standardized assessments in mathematics has been steadily increasing: besides the well-known PISA and TIMSS, many countries have been developing their own standardized assessments, often sharing, in part or completely, the frameworks of the famous international assessments. Frequently the results of these assessments are object of harsh debates within national educational communities, and they have repercussions on the educational decisions made by politicians (Breakspear, 2012; Kanes, Morgan & Tsatsaroni, 2014), which can have important consequences trickling down to the classroom level. This is why the use of standardized assessments is a phenomenon that cannot be ignored by the research community.

National and international standardized assessments of mathematical competencies can offer a huge quantity of data and the possibility of comparing results of students from different places and schools; moreover, in many cases the items challenge students to solve stimulating non-routine problems. However, the use of standardized tests also raises various other issues that have been criticized (Kanes, Morgan & Tsatsaroni, 2014; Carnoy, 2015). These have to do with aspects on three different levels:

- Structure: tests are essentially made up of closed questions. The results give information only on the *product* (the student's answer) rather than on the *process*.
- Administration: there is a tight time limitation and students cannot ask for clarifications of any sort on the text of the problems.
- The way the results are discussed.

We see the last level as the most critical: on the one hand, the popular media discourse focuses primarily on *competitive* issues, *i.e.* how each country/school/classroom is positioned in the rankings. On the other hand, in many countries, the students' results on standardized mathematics tests have paved the way for significant educational reforms, even though there are relatively few additional primary or secondary analyses of the data by researchers, evaluators and experts. Particularly worrisome is that often in the popular media discourse the quantitative results of these large-scale assessments become "an unavoidable (and 'obvious') provider of information 'based on proof'" (Carvalho, 2012). In this respect Pons (2012) highlights the lack of institutional and cognitive spaces of enrollment and interest-building in which results of these standardized assessments could be analyzed, translated and reinvested by people. Finally, as Doig (2006) underlines, the information provided by the standard summative reporting methods have little effect on mathematics teachers' development and everyday practice. Embracing this perspective we ask what might we gain educationally from standardized mathematical assessments.

Standardized assessments are interested in *describing* quantitative macro-phenomena rather than in *understanding* them, also at a micro level. To gain such understanding, using Crespo's words (2000), it is necessary to acquire an interpretative orientation, collecting students' ideas with the purpose of *accessing* rather than *assessing* students' mathematical understanding. Through these means, we believe it is possible to achieve two goals:

1. To foster the development of a critical approach to standardized tests, interpreting, analyzing and limiting the potential of the information offered by the results;
2. To work in synergy to transform the tests into an educational opportunity for teachers and researchers.

In this article, we will use results from a recent teacher education project in Italy to describe practical ways of achieving these two goals. In particular, we will analyze ways to exploit two potentials for teachers and researchers we see in the critical approach to standardized tests: the *Informational Potential,* seen as the information that can actually be obtained by interpreting and analyzing students' performance results on standardized tests (this includes understanding the limitations of such potential); the *Developmental Potential,* seen as the educational opportunities offered by a critical approach to standardized tests and by a re-elaboration of the informational potential.

The project: context, goals and approach to the items

The teacher education project—focused on the Italian national standardized mathematics test, called INVALSI [1]—lasted one year and involved a Team of 31 members: 26 in-service teachers, 14 from primary school (grades 1 to 5) and 12 from high school (grades 6 to 8), and five

Mathematics Education researchers. The project was inspired by the idea of co-learning community introduced by Jaworski and Goodchild (2006), and it was based on two tightly interconnected and well established goals. The first one was to develop new knowledge and competence through fusion of the different specific competencies of the researchers and the teachers during *a priori* and *a posteriori* analyses of the students' processes. The second one was to provide professional development for teachers, guiding their transformation into teacher-researchers and, in particular, broadening their *interpretative knowledge* (Ribeiro, Mellone & Jacobsen, 2016). In this perspective, teachers and researchers both played a crucial, though asymmetric, role.

The work was structured in three phases, carried out cyclically for each of the four content categories defined in the INVALSI framework [2]. The only actors present in all three phases were the teachers.

In Phase 1 the Team analyzed the INVALSI database (unlike PISA, all the INVALSI items are public and each item is explicitly related to a learning goal) and the teachers were asked to select items that they considered particularly significant. The Team then carried out an *a priori* analysis of each of the selected items, aimed both at highlighting the main difficulties that the item might cause for the students, and at predicting possible pathways leading to a student's answer. In order to emphasize, and not condition, teachers' voices, the researchers would present their observations only after a first analysis developed by the teachers.

Phase 2 was the experimental one. Selected items were assigned individually to classes of the appropriate grade level, adding the request "Explain how you reasoned". There was not an explicit time limitation and students could ask the teacher for explanations about the wording of the text of the item. The teachers took notes of all the students' requests of clarification and collected the students' written answers. During a later class period, a mathematical discussion was orchestrated by the teacher (notes were taken during each of these periods). The students' difficulties, their different solution strategies and argumentations were analyzed and compared.

During Phase 3 the Team analyzed students' excerpts and teachers' notes with the aim of recognizing the variety of processes involved in reaching each answer to the item.

Through the implementation cycles of the three phases, based on what emerged during each meeting, the researchers attempted to shift the focus of the discussions from students' products towards their processes. One means used for accomplishing this was a comparison between the *a priori* analysis developed in Phase 1 with the results obtained in Phase 2.

In the following sections, we present two examples of this 3-phase structure in order to illustrate how the Informational Potential and Developmental Potential were explored and exploited throughout these (and similar) episodes.

Example 1: Count the stars (grade 2)

This item (see Figure 1) was one of the few non-multiple choice questions on the INVALSI 2013 test. The teachers selected this item mostly because of their surprise at the low performance on a task they considered relatively easy (only about 55% of the national sample of students answered correctly). Primary school teachers also stressed the crucial importance of the ability involved (counting) for their school level.

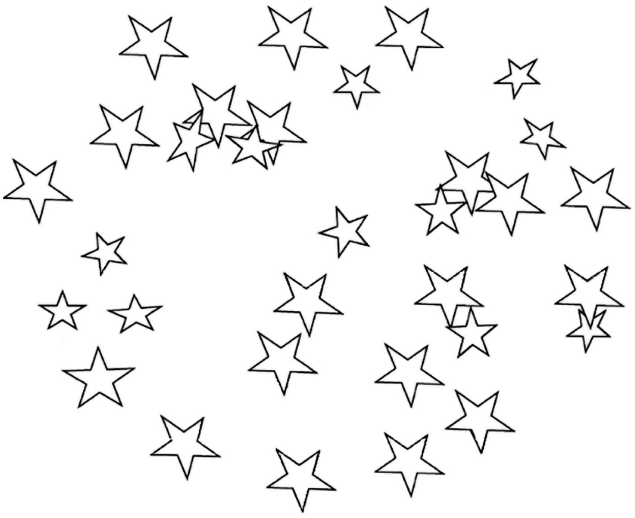


Figure 1. Count the stars. How many stars are there in the picture?[3]

The researchers, as well, considered the item interesting, but for different reasons. On the one hand, the item appears to be not so easy due to the chaotic arrangement of the objects to be counted, and their *nature* (the stars are not movable); on the other hand, it appears to be a manageable mathematical problem. Indeed, according to Ball (1988), these two conditions are essential for fostering, and thus observing, different solution processes. Moreover, the educational goal declared ("verify the mastery of efficient counting strategies") was clearly not verifiable through a standardized test, and it could be assessed only through an analysis and discussion of the counting processes enacted by the children.

In Phase 2 the item was assigned to 45 students from two classes, and it generated many different answers. Indeed, in answering the question "How can we know who counted correctly?" students explained how they had counted and willingly took into consideration strategies other than their own.

In their reports, the teachers were fascinated but, above all, surprised by the various approaches enacted and described by the students. They noticed four different types of counting strategies, which they described as follows:

1. *Free* counting: pointing with a finger or trying to follow a visual order between counted and uncounted stars;
2. *Marked* counting: marking the counted stars with a pen;
3. Counting through one-to-one correspondence: marking for each star on the page an X on another area of the page, and then counting the Xs drawn one under the other;
4. Counting by partitioning: partitioning the stars on the page into subsets in various ways, counting the

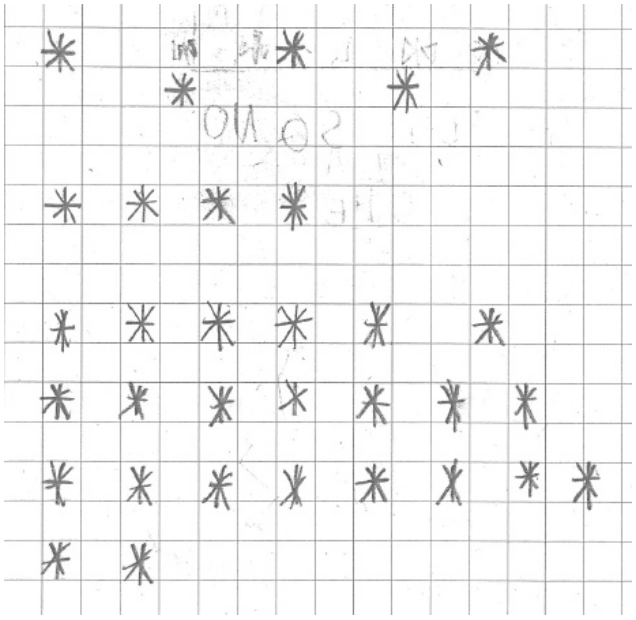


Figure 2. Cristian's drawing.

numbers of stars in each subset and then summing the numbers. This process was carried out in many different ways: some students used horizontal lines, some used vertical lines, some used sets with an equal number of stars (usually 3, 4 or 5), some used sets based on the proximity of stars.

To the researchers, what was more interesting than the account of students' strategies was to hear how the students explained how they knew that they had counted correctly, and to get insight into students' awareness of the two properties of the specific set of objects that made the counting difficult. In both classes students agreed that strategy 3 is the best for solving the problem of the chaotic arrangement of the stars, but in one class a student, Matteo, seemed to realize that it does not address the fact that the stars are not moveable:

Matteo: But then also the crosses you have to count them on your notebook. I think it would be better, instead of putting a mark on the stars, to put a little piece of paper on top of each star, and then count the little pieces of paper, so you can throw away the ones already counted.

The class happily took on Matteo's suggestion, but the students soon realized that the strategy was too hard to carry out empirically because the pieces of paper used to cover the stars had to be too small and they would easily move during the process. Therefore they decided to apply strategy 3, with specific control "to be sure that they counted well", suggested by Stefania:

Stefania: To be sure we have counted well, let's count 10 times.

Stefania's way of checking the solution may not be efficient (the children could keep on making a same mistake), but it is also true that in addition to the repetition of the task her strategy foresees the possibility that the counting would be carried out by different people. Moreover, as researchers, we noticed how the will to check the solution was in itself an important goal to achieve in the context of problem solving.

In the discussions with the teachers we also drew attention to how the possibility of enacting the strategy proposed for checking the solution depends heavily on the variable of *time*: during the official test children do not usually have this time, and in any case they tend to perceive the time as being insufficient for such strategies. This discussion led to more general considerations about what the standardized test actually assesses: the ability to solve problems or the ability to solve problems rapidly?

Finally, a conspicuous amount of time was devoted to discussing the case of Cristian, a child who had shown many difficulties during the phase of individual thinking about the problem, but who seemed to have identified the main sources of difficulty. Indeed when the teacher asked the children if they enjoyed it and why, he answered:

Cristian: No, I did not like it because the stars were hard to count, because they were on top of one another and all over the place.

Then the teacher asked the students to work individually and, "Try to draw the stars so that they are easier to count". All the children drew the stars arranging them in arrays, with groupings of 3, 4, 5 or 10, but Cristian did not (see Figure 2).

Both teachers and researchers were interested in the apparent discrepancy between the child's metacognition and his problem solving strategies. Discussing this discrepancy later, the teachers commented on the role played by the seemingly (to them) unusual request to make the task easier.

Example 2: comparing decimal numbers

Which of the following numbers <u>is closest</u> to 100? [4]

A) 100.010 B) 100.001 C) 99.909 D) 99.990

Example 2 (intended for grade 5) was of interest to all the teachers on the Team primarily because the answer most often chosen by the students in the national sample is not the correct one (44.6% answered D, 43.9% answered B); moreover, it struck their interest because of the mathematical concepts involved, considered significant at the transition between primary and middle school (similar items are typically used also in the test for grade 6). To the researchers the item was of interest not only for the reasons described by the teachers, but also because the explicit educational goal this item was designed to assess is "to compare decimal numbers," but in order to answer correctly the students needed to do more than simply compare decimal numbers. Indeed, another crucial and notoriously difficult concept is involved, that of distance (proximity) between numbers. We expected that there might also be difficulties arising from the polysemy of the word 'close' which carries everyday language meanings as well as a specific mathematical meaning.

During the phase of *a priori* analysis many teachers stated that the students' difficulties in responding to this item would be related mostly (if not univocally) to their understanding of decimal numbers.

During Phase 2 the item was proposed to 63 students from three different classes. The results we obtained were similar to those of the national sample: the most chosen option was still D, practically with the same percentages of the national sample. What was extremely interesting with respect to our Project, however, emerges from the analyses of the students' argumentations and from the discussions carried out in the three classes. Analyzing these argumentations, the Team noticed that actually very few choose an incorrect answer because of mistakes in carrying out calculations with decimal numbers. The teachers appeared to be shocked.

As they reported on the classroom activities, teachers started mentioning factors other than "dealing with decimal numbers" as possible sources of difficulties. Among these there were the following: a difficulty related to carrying out the algorithm for calculating the distance between numbers, when the concept of absolute value has not been introduced; a linguistic difficulty related to the term 'closest'.

During the Team discussions the researchers drew attention to the algorithm for calculating the distance between numbers. In analyzing how it might be enacted by their students, the teachers recognized its *asymmetry*: if X is greater than 100, the distance between X and 100 is X minus 100; if X is lesser than 100, the distance is 100 minus X. They also recognized that mistakes can be made when students try to *make the algorithm symmetric:* they calculate X minus 100 for each value of X, and—after that—they choose the least positive value (D). The second source of difficulties discussed was related to the linguistic difficulty related to the term 'closest'; the researchers leaped on their insight into this aspect, and instantly the awareness seemed to act as a catalyst in initiating numerous discussions, which led to the recognition of 'closest', in the analyses of classroom discussions, as "the most common source of mistakes" for this item. The teachers commented on their being surprised at seeing this source of difficulties, especially because they saw it (correctly) as being completely unrelated to the educational goal the item was designed to assess.

Indeed, from the transcripts and teacher notes that we analyzed together, it was clear that for many students, the expression 'closest to 100' means 'the number does not exceed 100'.

> *Carla:* I have not considered the numbers greater than 100 because 'the closest to 100' means that it does not reach 100.

In other words, for many students—practically for all who choose answer D—the sentence 'the number X is close to 100' implies that X *precedes* 100. This is explicit in Luca's words:

> We have to exclude 100.010 and 100.001 [the answer options A and B] because they are over 100, therefore they exceed it and they move away from it!

During the followup discussions the researchers shared international research results with the teachers, such this finding of Boero, Douek and Ferrari: "Some difficulties generally arise from the differences in meanings and functions between the word component (*i.e.,* the words and structures taken from ordinary language) of mathematical registers and the same words and structures as are used in everyday life" (2008, p. 265). The teachers responded enthusiastically and seemed proud to understand and "be understood by" international research findings, reporting on numerous other representative episodes from their classroom discussions.

The teachers seemed to be highly motivated and proposed investigating further how linguistic aspects related to the concept of 'close to' influenced students' behavior in mathematics. So the Team decided that the teachers would ask students in different classes to think of a new formulation of the text of the item, designed to overcome the linguistic difficulty related to the meaning of 'close to', but without omitting the words 'close to'. After an animated discussion, one of the classes proposed the following text: "Which of the following numbers is closest to 100, going back and forth on the number line?".

This formulation was used in three other classrooms. Although this text does not apparently simplify the mathematical task, the result of its implementation in the other classrooms was that very few students chose an incorrect answer. This was extremely interesting both to the researchers and to the teachers because it confirmed the conjecture that being able "to compare decimal numbers"—the declared educational goal the item was designed to assess—actually was not a source of difficulty; although the quantitative results were interpreted as evidence that most students had not achieved this goal!

Exploiting Informational Potential and Developmental Potential

Two criteria were dominant in the teachers' processes of choosing items in the first implementation of Phase 1:

1. Surprise in seeing an unexpected (usually low) performance result by the students on a specific item of the INVALSI test;

2. Relevance of the item with respect to the specific mathematical content involved.

A result regarding the Developmental Potential was the appearance of a third criterion in the later implementations of Phase 1: *The relevance and the variety of the hypothesized thinking processes activated by the student to answer the item*. This criterion was not present in the first implementation. Indeed, as was the case with the two examples above, the teachers took a very normative approach, being quite certain about why a specific answer was given by the students. They tended to identify as causes (few) cognitive aspects related to the specific mathematical content of the item, and see no other possible interpretations of the students' responses. This could be described in terms of teachers' initially limited interpretative knowledge (Ribeiro, Mellone & Jacobsen, 2016).

Another index of lack of such interpretative knowledge can be seen in the mismatch between the teachers' expectations and the actual results obtained in the experimentations, which indeed led to *surprise* both in the case of Cristian (Example 1) and in the case of the answers provided by the students to "the number closest to 100" (Example 2).

In the teachers' initial responses there were few possible interpretations of a student's incorrect answer, typically

related to a low knowledge about the mathematical topic involved in the item. This leads to a first crucial consideration: a lack of interpretative knowledge and a narrow attention to the students' scores inhibit the Informational Potential of standardized tests. Instead, to exploit the Informational Potential it seems necessary to critically discuss the 'automatism' according to which a given distractor is associated with a precise (incorrect) solution process. In order to do that, the request for argumentation is crucial: the awareness of different ways in which a distractor may be selected can emerge only from the analysis of the argumentations or from a mathematical discussion.

By the end of the project, teachers' interpretative knowledge had certainly grown. This result was also confirmed by an *affective clue:* the instances of surprise practically disappeared in teachers as the project progressed. However, we found quite interesting and satisfactory that the teachers were still eager to foster in-class discussions with their students on test items to explore their own conjectures on difficulties that they expected might emerge. In other words, we witnessed how growth of interpretative knowledge can go hand in hand with growth of awareness of the Informational Potential of standardized mathematics tests, leading to the development of teacher-researchers.

Another interesting result is how the teamwork aimed at developing a critical approach to standardized tests can lead to developing interpretative knowledge, which is a goal *per se* in professional development (Ribeiro, Mellone & Jacobsen, 2016), but it is also a necessary condition for exploiting the Informational Potential of the items considered. Looking further into these results, we see how the researchers could focus on *unexpected processes* to explore and exploit the Developmental Potential. This is something we learned during the project; indeed, for us it was important to gain insight into how elements of the Informational Potential can be used to exploit the Developmental Potential. A related finding has to do with the emergence of a key role played by the *cognitive evaluation of the unexpected.* While the unexpected was of interest to the researchers from the very beginning of the project, for the teachers this was not the case; what teachers seemed to attend to was an evaluation of the didactical significance of the unexpected, of its being connected to a significant goal, and of its being related to something done in class (and therefore being responsible for it). We see this as an important difference between teachers and researchers, related both to the Informational Potential and the Developmental Potential.

The critical approach methodology we adopted allowed us to notice matches and mismatches in what was considered to be interesting for the teachers and the researchers. The discussions initiated about them guided the construction of a shared perception of the Informational Potential. We learned that reaching such a shared perception can include shifts in focuses of teachers' attention (*e.g.,* from products to processes) but also shifts in the researchers' view, and this is an important aspect of the Developmental Potential. Zooming into specific classroom episodes and cases, as in the case of Cristian, allows us to bring teachers' attention to aspects such as control on solutions, metacognitive aspects of learning, the issue of "time" in regular class activities and on standardized tests, the influence of language and choice of words on standardized tests. Focusing on these aspects in students' responses and using these as entry points for talking about international studies in math education, and providing teachers with new lenses seems to broaden their interpretative knowledge (in Example 2, the teachers seemed to suddenly recognize a phenomenon). This strategy seems particularly powerful in exploiting the Developmental Potential.

Moreover, reaching an awareness of how an aspect like time can impact students' performance also leads to deeper understanding of the actual Informational Potential of standardized tests. Indeed, an important outcome of the project was the ignition of teachers' curiosity about questions they saw as significant from a didactical point of view, and how by the end of the project they had also come to see as significant other issues than the ones they initially attended to. For example, if an item had a very low rate of correct answers, then, exploiting the Informational Potential, an issue can be proposed as a likely cause of students' errors (and this might not have to do with the expected cognitive difficulties related to the topic). This issue can become a new source of interest for the teachers, leading them to develop further analysis on this issue (assuming the attitude of a teacher-researcher). In this way the Developmental Potential is also exploited.

Standardized mathematics tests as a resource for teachers and researchers

We conclude with a comment about the significance of the choice of analyzing items from a standardized mathematics test.

First and foremost, there is a motivational reason. Standardized tests are relevant to the educational system almost worldwide. Teachers are usually highly motivated in everything that revolves around standardized tests because of the social pressure that they feel around their students' performance. Moreover, the tests can foster teachers' motivation to collaborate because tests represent a common experience for them (the same tests are administered to all students). In the project, we tried to increase motivation by assigning teachers the role of main actors in choosing the items to work on.

A second reason is that the items on the standardized tests are usually considered to be actual *problems* for teachers and students. That is, they are usually tasks "for which the solution method is not known in advance" (NCTM, 2000) and not exercises drilling the students on the mere application of a studied algorithm. In other words, the items are designed to assess productive rather than re-productive processes: this is key for having the opportunity to encounter a wide range of solution processes. This is surely true in the Italian context for the INVALSI items: they are generally perceived by students and teachers as difficult problems. The perceived difficulty of the items thus can act as a trigger for teachers' curiosity in interpreting where the difficulties reside; and the variety of students' answers given in class allows teachers to shift students' focus onto processes while teaching, capitalizing on the need to choose and agree on which processes are correct and why.

Finally, we commented earlier how the items on the standardized test are designed to contain distractors, possible choices assumed to be the outcome of frequent incorrect solution processes. In other words, the choices are designed

according to a very normative interpretation. Learning to analyze this perspective critically, through growth of one's interpretative knowledge, opens windows onto students' actual solution processes and leads to deeper appreciation of the Informational Potential of standardized tests. Indeed we recognize that the work accomplished during the project on the items analyzed by the Team, especially the analyses of single student cases, led to deeper appreciation of the Informational Potential not only on behalf of the teachers but also of the researchers.

In conclusion, a different approach to the standardized assessment is possible and desirable in order to exploit the Informational Potential and the Developmental Potential of standardized mathematics tests. This project indeed shed light on how exploitation of the Informational Potential, and especially the Developmental Potential, yields the possibility of developing teacher-researchers, and researchers with a sharper eye on many educational issues through synergetic teamwork on behalf of all the participants.

Notes

[1] INVALSI is largely inspired by the PISA theoretical framework. The assessment is developed annually for grades 2, 5, 8, 10.
[2] INVALSI considers four content categories: Quantity, Uncertainty and data, Change and relationships, and Space and shape.
[3] All quotations from our data in this article are translations.
[4] Underscore in the original text.

References

Ball, D.L. (1988) Unlearning to teach mathematics. *For the Learning of Mathematics* **8**(1), 40–48.

Boero, P., Douek, N. & Ferrari, P.L. (2008) Developing mastery of natural language. In English, L. (Ed.) *International Handbook of Research in Mathematics Education,* pp. 262–295. New York: Routledge.

Breakspear, S. (2012) The policy impact of PISA: an exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers, No. 71.* Paris: OECD Publishing. http://dx.doi.org/10.1787/5k9fdfqffr28-en

Carnoy, M. (2015) *International Test Score Comparisons and Educational Policy. A Review of the Critiques.* Boulder, CO: National Education Policy Center. http://nepc.colorado.edu/files/pb_carnoy_international_test_scores_0.pdf

Carvalho, L.M. (2012) The fabrications and travels of a knowledge-policy instrument. *European Educational Research Journal* **11**(2), 172–188.

Crespo, S. (2000) Seeing more than right and wrong answers: prospective teachers' interpretations of students' mathematical work. *Journal of Mathematics Teacher Education* **3**, 155–181.

Doig, B. (2006) Large-scale mathematics assessment: looking globally to act locally. *Assessment in Education: Principles, Policy & Practice* **13**(3), 265–288.

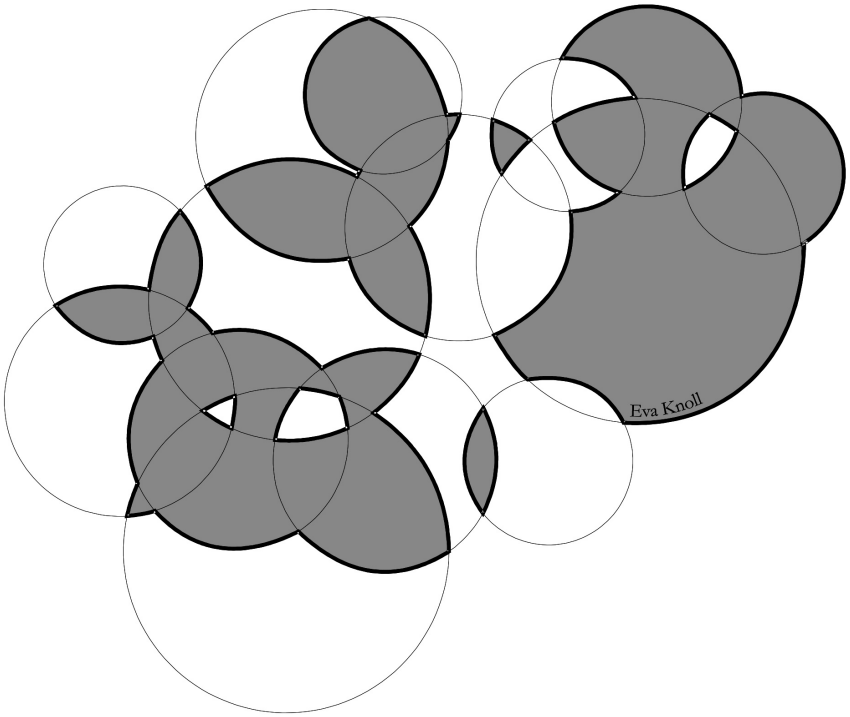
Jaworski, B. & Goodchild, S. (2006) Inquiry community in an activity theory frame. In Novotná, J., Moraova, H., Kratka, M. & Stehlikova, N. (Eds.) *Proceedings of the 30th Conference of the International Group for the Psychology of Mathematics Education,* Vol. 3, pp. 353–360. Prague: Charles University.

Kanes, C., Morgan, C. & Tsatsaroni, A. (2014) The PISA mathematics regime: knowledge structures and practices of the self. *Educational Studies in Mathematics* **87**, 145–165.

National Council of Teachers of Mathematics [NCTM] (2000) *Principles and Standards for School Mathematics.* Reston, VA: Author.

Pons, X. (2012) Going beyond the 'PISA shock' discourse: an analysis of the cognitive reception of PISA in six European countries, 2001–2008. *European Educational Research Journal* **11**(2), 206–226.

Ribeiro, M., Mellone, M. & Jakobsen, A. (2016) Interpreting students' nonstandard reasoning: insights for mathematics teacher education. *For the Learning of Mathematics* **36**(2), 8–13.



What kinds of configurations can you obtain if you systematically mark every second segment on each of a randomly arranged collection of overlapping circles?