Primary school students reasoning about and with the median when comparing distributions
Susanne Schnell, Daniel Frischeimer

To cite this version:
Susanne Schnell, Daniel Frischeimer. Primary school students reasoning about and with the median when comparing distributions. Eleventh Congress of the European Society for Research in Mathematics Education (CERME11), Utrecht University, Feb 2019, Utrecht, Netherlands. hal-02412841

HAL Id: hal-02412841
https://hal.archives-ouvertes.fr/hal-02412841
Submitted on 15 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Primary school students reasoning about and with the median when comparing distributions

Susanne Schnell¹ and Daniel Frischemeier²

¹ University of Paderborn, Germany, Susanne.Schnell@math.upb.de
² University of Paderborn, Germany, dafr@math.upb.de

Statistical reasoning can be enhanced at an early stage by facilitating informal statistical reasoning and developing conventional measures based on individual conceptions. This paper aims at identifying young students’ conceptions of the median and their relations to its definition. For this, we present a design research study with grade 3 students (age 9) focusing on the comparison of two groups. The qualitative analysis of complementary interviews shows that young students already develop an initial understanding of the median but struggle to find the conventional reference for the middle. Consequences will be drawn for further design cycles.

Keywords: statistics education, early statistics, primary school, comparing groups, median

Introduction: Early statistics in German curriculum

Statistical literacy is a fundamental component for active citizenship (Engel, 2017). Some facets of statistical literacy can be approached as early as primary school – an overview is presented in Leavy, Meletiou-Mavrotheris, and Paparistodemou (2018). In Germany, nation-wide standards for primary education in mathematics require students to gain competences in dealing with “data, frequency, and chance” (Hasemann & Mirwald, 2012). The core idea is to develop conceptual fundamentals, e.g. by posing statistical questions, collecting data, and creating and interpreting representations of data (Hasemann & Mirwald, 2012). German textbooks and teaching materials for primary school, however, often emphasize singular aspects such as reading a diagram rather than introducing context-rich and complex data analysis projects as sustainable teaching-learning arrangement (Bakker, 2004). Such an arrangement aims at developing students’ competences in drawing and dealing with informal statistical inferences (Makar, Bakker, & Ben-Zvi, 2011) and building on existing informal conceptions in order to develop conventional ideas (Bakker, 2004). Hence, researching initiated learning processes is an important task (Makar, 2014). This paper presents a design-based research project (Prediger et al., 2012) which was conducted over four cycles so far in order to develop a statistics teaching-learning arrangement for classrooms in German primary schools. After summarizing theoretical considerations, the main design principles and framework of the teaching-learning setting are discussed. Lastly, we will present empirical insights of students’ statistical reasoning about and with the median from interviews with students grade 3 (age 9) conducted after the whole-classroom activities.

Statistical projects in primary school

Going beyond calculation and algorithms in statistics education means enabling students to deal with statistical inquiries and applying the phases problem, plan, data, analysis and conclusions of the PPDAC-cycle (Wild & Pfannkuch, 1999). Especially at the starting point of a statistical inquiry – e.g., when posing statistical questions – the desire arises to not only consider the distribution of one variable, but to look for a relationship between two or more variables. So, for example, the question whether third graders tend to have more games on their smartphones than fourth graders
leads to the comparison of two groups. While easily motivated, group comparisons are also sophisticated statistical activities taking into account many of the fundamental statistical ideas like data, variation, representation, etc. (Burrill & Biehler, 2011). Frischeimer (2017) identified several important and sustainable group comparison elements like center, spread, skewness, shift, p-based and q-based comparisons. While these concepts can be quite complex, the fundamental ideas behind them can be addressed already in primary school. Informal comparisons are a fruitful task to provide in-depth insights into students’ conceptual development and informal statistical reasoning (e.g. Schnell & Büscher, 2015, Makar, 2014). In this paper, we focus on using (informal) conceptions of the median as well as visual features of distributions for comparing groups in early statistics.

To describe the center of a distribution is a key idea of statistics (e.g. Bakker, 2004; Makar, 2014). The concept of the median can be introduced on an enactive level via “animated statistics” in the sense of embodied cognition by Lakoff & Núnez (2000). In this case the students are statistical units themselves, line up in an ascending order in regard to a certain attribute (e. g. their height) and identify the middle. Even the more formal procedure of determining the median (sorting values and finding the middle value) is easily and non-formally presentable, but the interpretation and a deeper conceptual understanding can be challenging (Mokros & Russel, 1995) as it is interwoven with ideas like representativeness and typicality. However, these ideas can be pre-formally existent from even young students’ everyday experiences (Mokros & Russel, 1995). Mokros and Russel (1995) identified approaches to the average, which include among others the average as an algorithm without predominant focus on contextual interpretation or the average as midpoint which is chosen as representative of the data and draws on the ‘middle’ which alternately is the median, the middle of the X-axis or the middle of the range. To be able to interpret the median correctly and in a meaningful way, it has to be understood as a value in reference to the rest of the distribution (e.g. cutting the data set in two halves). Overall, even students who are familiar with formal ways to determine the median sometimes do not apply it when comparing distributions (e.g. Watson & Moritz, 1999; Konold et al., 1997). This might be due to a lack of understanding that measures such as the median are representative of a distribution (Konold et al., 1997). However, Makar (2014) shows how even young students of age 8 can be supported in developing a fundamental understanding of the median and using it for informal inferential reasoning when presented with an adequate teaching-learning arrangement. The key is to give students room to develop informal ideas which serve as important base for building foundational knowledge of statistical concepts (Makar, 2014; Bakker, 2004; Smith, diSessa, & Rochelle, 1993).

Especially in terms of visual aspects, young students draw on a variety of different strategies to describe and compare distributions (e. g. Watson and Moritz, 1999). For instance, they focus on ‘hills’ to describe the shape of the data (Bakker, 2004). To facilitate a proto-concept of center, Konold et al. (2002) introduced so-called modal clumps as “a range of data in the heart of a distribution of values” (p. 1) in order to find ways to identify the center in stacked dot plots and “allow students to express simultaneously what is average and how variable the data are” (p.1). Bakker and Gravemeijer (2004) build on these concepts to lead students from a local perspective (focusing on singular dates such as the minimum) to a global perspective on distributions (with a
certain shape, spread etc.) and see modal clumps as proto-concepts not only for center but also for spread and therefore provide students with the opportunity to compare two distributions.

Overall, our aim is to design and realize a teaching-learning arrangement to enhance the statistical reasoning of primary school students and to enable them to compare groups with proto-concepts like modal clumps and concepts like medians when working on statistical projects. This paper focuses on exploring which concepts of the median were developed by the children during the teaching-learning unit and applied in the interview. Therefore, the research question is: Which conceptions of the median do students verbalize in the interviews and what is their (visual) point of reference for the median?

Research Design

Methodological framework: Design Research

This study is situated in the methodological framework of topic-specific didactical design research (Prediger et al., 2012). Thus, it aims at designing a specific teaching-learning arrangement for data competence in primary school on the one hand, and the development of local theories of teaching and learning conceptual basics at an early age on the other hand. These intentions are addressed in iterative research cycles of (re-)designing, implementing, and analyzing the materials (Prediger et al., 2012). So far, four cycles were conducted, exploring different settings with whole classes and special interest groups in grade 3 and 4. After the classroom-activities, additional interviews for in-depth analysis of students’ reasoning were conducted.

Design principles and teaching-learning arrangement

For the design of the teaching-learning unit we implemented several recommendations for the design of Statistical Reasoning Learning Environments (Garfield & Ben-Zvi, 2008) such as using real data sets and using digital tools like TinkerPlots (Konold & Miller, 2011), which can help young students to manipulate and modify the data in regard to a specific question. Furthermore, we facilitated collaborative work on statistical projects (first with small datasets, later with larger datasets), introduced the students to data analysis concepts on different representation levels (enactive, iconic, symbolic) and promoted the use of informal concepts such as modal clumps (Konold et al., 2002).

In the following we refer mainly to the version of the teaching-learning arrangement by Breker (2018) from the last design research cycle in which 22 third grade students participated. All of the students have no specific statistical pre-knowledge – they have only been introduced to the reading and interpreting of pie graphs and tallies. The teaching-learning arrangement consisted of seven modules spread over 13 lessons (45 minute each):

<table>
<thead>
<tr>
<th>Module</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Getting to know each other</td>
</tr>
<tr>
<td>1</td>
<td>First basics in data analysis and getting to know the questionnaire</td>
</tr>
<tr>
<td>2</td>
<td>Statistical representations on different representational levels (enactive, iconic, symbolic)</td>
</tr>
<tr>
<td>3</td>
<td>Introduction to data analysis with TinkerPlots and creating stacked dot plots in TinkerPlots</td>
</tr>
</tbody>
</table>
Getting to know modal clumps, medians and hatplots to compare groups

Group comparison projects: Preparing the posters

Presentation of the group comparison posters

Conclusion of the project

The median was introduced in module 4. As mentioned above, one fundamental idea was to present data analysis concepts on different representation levels. Hence the median was introduced on an enactive level with animated statistics: The students were asked to line up in a row ascending by their height and determined the child in the middle. After completing a fill-in-the-blank text with the characteristics of the median, the teacher demonstrated how to use TinkerPlots to identify the median in a distribution. The median was then used by the students to compare different distributions.

Our aim was to investigate the ways in which the students use statistical concepts such as the median to compare groups in data sets in TinkerPlots after completing the teaching-learning arrangement in class. Hence we conducted subsequent interviews with a convenience sample of participants. Here, we provided them with a real data set including variables about leisure time activities and media use of about 680 primary school students (grade 3 and 4). The participants’ task was to investigate whether the third or the fourth graders have more games on their smartphones. The interview consisted of two main activities with regard to the task: First the students were asked to conjecture two dot plots (grade 3 vs. grade 4) based on their expectation. Second the students were asked to use TinkerPlots to compare the groups of third and fourth graders in the real data set.

Data collection and analysis

Ten students participated in the interview study of cycle 4 after they completed the teaching-learning arrangement with their class. In the interviews, the students worked in pairs of two. All interviews, including the activities in TinkerPlots, were recorded and fully transcribed. The analysis was conducted under the interpretative paradigm, scrutinizing the transcripts for sequences in which the students use the median and then reconstructing their cognitive activities and conceptions.

Findings

The middle of what? – The case of Paula and Linda

Paula’s and Linda’s initial hypothesis is that fourth graders have more games on their smartphone. After creating two conjectured dot plots, the researcher asks the students to explain the meaning of the median.

80 R: Could you maybe explain it again: what is the median, Paula?

81 Paula: The center point of the/ well it’s the center of the dot plot.

Paula calls it the ‘center point’ and seems to be aware that she has to specify the point of reference of the center. Her offered reference ‘the center of the dot plot’ (line 81) is ambiguous as it could refer to the X-axis, the range of data dots or the distribution of data dots (similar to average as midpoint, Mokros & Russel, 1995). For determining the median, she counts only the columns of dots (with the same value) rather than every single dot, which equals determining the center of the
range. Both girls interpret this median correctly in terms of the higher median indicating more games for fourth graders.

When introduced to the real data set, the girls first determine and interpret the hills. Asked to summarize the findings in regard to the overall question (line 218), Paula argues with the median which they have not determined for this distribution, yet.

![Figure 1: Paula and Linda’s TinkerPlots screen with drawn-in hills and later added medians](image)

**Figure 1: Paula and Linda’s TinkerPlots screen with drawn-in hills and later added medians**  
(median of 3rd graders: 8, median of 4th graders: 4)

218 R: Okay. Good. Now, what did you find out for your question, when you look at the hills? 3rd or 4th graders – who has more games on the smartphone? (…)

223 Paula: 4th graders. (…) Because 4th graders moved farther to the right with their dots and if you take the exact center, now, only of the dots, not of the stripes [i. e. X-axes]. Because otherwise the median would be the same. Just from the dots, the median of the 4th graders is farther in the middle than of the 3rd graders.

224 Linda: Yes, definitely. But really, if you just take half somehow, then it really looks like the 3rd graders were furth/ had more (incomprehensible)

Describing how the distribution of fourth graders “moved” more to the right in comparison to the third graders, Paula describes the center as a point of reference (line 223). It is likely that Paula is not thinking of the correctly determined median (which - in contrast to her expectation - is smaller for the fourth graders than for the third graders), but rather her individually determined center of the range. However, her remark about the X-axis shows again awareness of the reference for this value. While both children are at first confident in their answer of the 4th graders having more games, Linda mentions doubts in line 224. By “taking half”, she could either refer to the left half of dots (e.g. the data between 0 and 16 games) or possibly to the lower “half” (e.g. the range of 0 to 36 games but cutting of each column above a certain frequency). This idea is not followed up upon, though.

When the girls determine the median in TinkerPlots, it seems to differ from their expectation:

238 Paula: They just took it of the hills. And in the hills, 3rd graders have more.

239 Linda: Right.
R: In the hills? I didn’t get that.
Paula: So the median of the hills.
R: The median of the hills? What is that?
Linda: They didn’t use all of them. They just drew it within the hills. (…)
Paula: Yes, but otherwise the [median] of the 4th graders would be much further in the middle and the one of 3rd graders, too. (…)
Linda: If you cut off the hill completely, then it should be sitting somewhere here, and of this one here at 14, 12, 16. Somewhere there. (…)
R: So, if you only look at the medians, who has more games on the smartphone?
Paula: Yes, in this case, where they used only the hills, it’s the 3rd graders. But if you had done it without the hills, it would be the 4th graders.

Paula explains the unexpected median by the way she thinks it was determined in the program, i.e. by determining the center of the hill’s width (cf. figure 1). Thus, she still interprets the TinkerPlots median as referential but differing from the learned definition. In line 247, Paula suspects the true median of the 4th graders to be farther in the middle (i.e. right). This indicates that she stays with her initial interpretation of the median as center of the range. In line 250, Linda suggests a manipulation of the data by “cutting of the hill”. As she says the median would then be between 12 and 16, she might refer to cutting of the dot plot horizontally, thus evening out the number of dots per number of games and making Paula’s procedure of determining the median correct. Thus, Linda twice offers ways to manipulate the data set and thinks about the effects on the median. Overall, the girls use the idea of two different medians in order to support her initial expectation that fourth graders have more games on the phone (line 272).

In regard to the research question, the girls developed conceptions of the median as the middle of the range and median as the middle of the modal hill. They are very consistent with their way of determining it, which might have its roots in the animated statistics activity. It is remarkable that they consider different options of manipulating the data set and their effects on the median, which hints that they perceive it as a value that is closely connected to the distribution and that has representative value. By assuming that TinkerPlots uses the hills’ boundaries rather than the whole distribution for determining the median, the girls establish a pre-formal connection between the visual approach and the calculated measure.

Additional interpretations of the median

The students in the interview study showed a variety of different conceptions of the median:

- Layla and Sandra as well as Burak and Tarik call the median “the middle (of the dots)” (LS_line 483, BT_line 103) but do not specify it further other than using the learned procedure to determine it → median as unspecified middle
- Johannes and Nils interpret it as the center with “roughly the same amount [of dots]” in each half (JN_line 53) → median as the middle of two halves
Lara and Lia seem to make a connection between the modal value and median “A median is the middle of the biggest. Of the biggest amount.” (LL_line 171) → median as the middle of modal clump

Conclusion

This paper aimed at identifying young students’ conceptions of the median and their relations to the conventional measure (Bakker, 2004). The analysis shows how the students are integrating visual and calculated aspects of the distribution in order to compare groups. Building on Mokros and Russell (1995), we could identify several conceptions of the median which draw on different references in the data set: median as middle of the range, median as middle of the modal hill, median as the middle of two halves, median as the middle of modal clump or column and median as unspecified middle. While all pairs were able to interpret the shift of the medians correctly, the in-depth analysis shows their struggle with the referential concept. Even though these conceptions were not yet in-line with the conventional conception of the median, they serve as important starting points for the development of more formal conceptions (Smith, diSessa, & Rochelle, 1991). Overall, the teaching-learning arrangement clearly supports young students in informal reasoning about statistical concepts while comparing groups. For instance, Linda’s considerations of manipulating the data set in order to change the median are promising starting points for further design cycles.

Another implication from the analysis is to be careful when introducing the median via animated statistics. When choosing this way, a scale should be prepared so that the distances between the children (i.e. data dots) become apparent and stacked distributions should be discussed. For a hypothetical learning pathway, we suggest to introduce first the modal clump as main area in the heart of the data distribution and then the median as a special value in relation to the modal clump, indicating the center of the distribution. Next discuss different distributions in which the position of the clump and the median differ. These ideas will be realized in the next design cycle of the teaching-learning arrangement.

References


