STATISTICAL THINKING AND LANGUAGE – A QUALITATIVE ANALYSIS

Ute Sproesser, Sebastian Kuntze

Ludwigsburg University of Education, Germany

This paper presents results of a qualitative study on the role of language for statistical literacy and statistical thinking. As statistical literacy is targeted in competencies necessary for social participation, communication skills related to basic statistical models appear as crucial. Language thus comes into play at various levels when learners make sense of situational contexts statistically or when they present solutions. Misconceptions might be connected to a development need of domain-specific language. The results of this study suggest that such language-related knowledge merits focused support in corresponding learning environments. The development of statistical literacy and statistical thinking might thus be framed, i.e. enhanced or restricted, by a corresponding development of domain-specific language.

1 INTRODUCTION

Language plays a role for statistical thinking—not only abilities of data-related reading appear as crucial, but it is also highly important to be able to express statistical observations or ideas in an adequate language. Whereas competency aspects of data-related reading can be assessed with test instruments and have been integrated in competency models for statistical literacy, the area of data-related speaking or writing merits more focused empirical research.

Consequently, this paper reports corresponding preliminary findings from a study with secondary students. The students were given tasks related to statistical contexts and asked to give written answers. An analysis of these answers focused on the way students expressed their ideas and on the role of language for the statistical understanding shown in the answers. The results of a case-based analysis indicate that some students have problems that appear to stem from the area of language use rather than from their statistical understanding. Moreover, existing intuitions could be further developed by learning opportunities targeted on language.

In the following second section, we will give a short overview of the theoretical background, which leads to the research interest of this study presented in the third section. We will then describe methods and design in the fourth section, present results in the fifth section, and conclude with a discussion in the sixth section.

2 THEORETICAL BACKGROUND

Curcio's (1987) metaphores 'reading the data', 'reading between the data', and 'reading beyond the data' indicate that reading competency is related to statistical literacy

(Wallman, 1993; Watson & Callingham, 2003). We understand statistical literacy (SL) as a competency dimension (cf. Watson & Callingham, 2003) rather than a curriculum-like catalogue of elementary contents in statistics. More precisely, extending Curcio's (1987) approach and integrating the thoughts of Watson and Callingham (2003), the metaphor of "data-related reading" can be used as a core thought of a competency model in the area of SL (Kuntze, Lindmeier, & Reiss, 2008; cf. Kuntze, Engel, Martignon & Gundlach, 2010), so that the competency of learners in SL can be described not only on an elementary level but also on higher levels of complexity. According to this competency model, elements of statistical thinking (Wild & Pfannkuch, 1999; Chance, 2002; Ben Zvi & Garfield, 2004) intervene at various levels of complexity and contribute to SL (cf. Kuntze et al., 2010).

Reading data from representations such as bar diagrams is not only relevant for SL, it is also addressed e. g. in the PISA reading competency tests. Reading competency encompasses dealing with information given in texts and other graphical representations like diagrams. Correspondingly, reading is not only a language processing activity: when dealing with data from situational contexts, processes of making sense require modelling activities. For instance, when reading data from a diagram, the graphical elements have to be interpreted against mathematical model(s) such as scales, proportions or absolute values, or even translated into a particular non-given mathematical or statistical model which can be used for representing or analysing the data. Such modelling activities make the idea of reading in representations of data very meaningful e. g. for describing competencies of learners. Extending Curcio's approach and integrating the thoughts of Watson and Callingham (2003), the metaphor of "data-related reading" can be used as a core thought of a competency model in the area of SL (Kuntze, Lindmeier, & Reiss, 2008; cf. Kuntze, Engel, Martignon & Gundlach, 2010).

Even though data-related reading is hence an active process involving specific modelling activities, a potentially even more demanding active process consists communicating related to data in adequate ways, regardless whether in an oral dialogue or in written form. Learners have to use their language skills in specific ways when expressing their understanding of situations in which statistics can play a role.

Further, language plays a role for many areas of statistical problem solving. It already comes into play when students have to decode a complex statistical situation and formulate the problem (Shaughnessy, 2007). When passing through PPDAC cycles (Wild and Pfannkuch, 1999), language may frame the corresponding statistical thinking process. Gal (2003) identified the ability to express someone's opinion concerning statistical information as crucial for statistical thinking and SL. Gal as well as Watson (1997) emphasised the importance of the adequate use of statistical terms as well as the ability to (critically) communicate one's reaction to statistical information. Even though Watson and Gal underlined the importance to use standard statistical language and although Biehler (1997) points out that some problems in data-related communication are caused by a lack of formal language, we see the need to focus also on pre-formal language use. In line with Makar and Confrey (2005), this study hence examines students' abilities to express their ideas and their understanding of a statistical situation regardless the level of abstraction of the language used. By accepting all types of language in analyses of cases of students' answers, we strive to access their statistical understanding of the situation. This corresponds to a research need, as the questions such as how statistical understanding develops together with language or how statistical understanding can be fostered in learning environments focusing on language have not be answered fully in prior research.

Concerning our interest related to language use and statistical understanding, we would like to recall that we keep distinct the (potentially non-verbal) statistical understanding of learners from the form this understanding may be expressed by means of language. It may appear as relatively obvious that a good statistical understanding may coexist with a high statistics-specific language mastery, or that difficulties in the statistical understanding can coincide with difficulties in statistics-related verbal expression. But there could also be the following cases: For example, a student might have a good statistical understanding of a phenomenon (e.g. in the sense of intuitions, cf. Fishbein, 1975) coupled with a low ability to express this understanding with domain-specific vocabulary. Conversely, a learner might have acquired a "language toolbox" related to statistics but as a consequence of a non-optimal statistical understanding, he or she might be unable to use this vocabulary adequately in a corresponding argumentation (cf. Makar & Confrey, 2005). Even if we assert that all four cases are basically possible, we expect that the development of specific language supports statistical understanding and that non-verbal statistical intuitions may promote the process of making sense of statistical vocabulary. Students' abilities to express their statistical understanding as well as the role of specific language development for aspects of statistical thinking thus merits in-depth empirical attention.

3 RESEARCH INTEREST

The theoretical considerations of the previous section highlight a need of research about how students express their understanding of statistical problems. In particular, the role of language for statistical thinking needs to be examined.

Consequently the study focuses on the following research questions:

How do students express their ideas and their understanding of statistical problems?

How can the relationship between students' language skills shown in verbal statements and their statistical understanding be described and analysed?

4 DESIGN AND METHODS

This study was carried out as part of the project ReVa-Stat, ("Developing concepts of data-related reduction and statistical variation as a support for building up SL"). In the first phase of this research project, 83 students of grade 8 (39 girls and 44 boys, mostly 14 years old) were asked to work on tasks related to different statistical problems. These tasks addressed basic aspects of statistical thinking like organizing and representing data as well as modelling and reflection activities about variation or data reduction – hence aspects of statistical thinking relevant for SL. In their usual mathematics lessons, students worked on these tasks so that they were able to cope with the presented tasks. As the students had not attended any specific statistics course prior to participating in this study, they were asked to describe their ideas in their own words. We did not stimulate students to use a specific type of language. During four lessons, the students worked on this specific learner-centred material, which affords exploring their knowledge, views and ways of expression. In the material, students were encouraged to first discuss the presented statistical problems in pairs and then to write down their answers. Data was gathered by collecting their written work. During the work phase of the students, a member of the research team observed the implementation and use of the material in the classrooms.

The written results of the work of the students were analysed using an qualitative bottom-up methodology based on Mayring's qualitative content analysis (2000). An interpretative analysis of the data was carried out by two researchers. In a first step, particular cases of students' answers were selected against the background of the research interest in order to identify types of answers inductively. In a second step, theses cases were analysed in more depth so as to further develop categories. I a third step, the developed categories were used to obtain an overview of the whole sample. For this purpose two researchers rated all cases and were able to classify them according to these categories in a top-down coding procedure. We will give further information in the following section together with the presentation of results.

5 RESULTS

In the following, we focus on answers the students gave when working on tasks in a relatively early phase (first lesson) of their work with the learning material. The task (see Fig. 1) refers to the context of frequencies of colours in packages of chocolate lentils (cf. Engel & Vogel, 2005, and Eichler & Vogel, 2012, for the task context, Watson & Callimgham, 2003, for the question format). Prior to this task, students had been informed that a sweets company produces the same number of chocolate lentils of every colour, they had then been given a package of 24 chocolate lentils as well as data about 9 more packages. To see the balancing nature of the mean, they also had calculated the average for every colour within the 10 given packages.

As shown in Figure 1, the students were asked to comment on their decisions for the three given diagrams. A sample answer of a student is presented in Figure 2 („Peter“). In his answer, Peter uses the word "distribution". This notion is described as "uniform" and "non-uniform" in the case of a) and b). According to Peters answer in b), in "real", the distribution is "fairly non-uniform" in his view. This expression appears to describe his understanding of statistical variation in this situation. The ways he uses the notion "distribution" in a) and b) enables him to characterise the data shown in the diagrams and to develop short and relatively dense argumentations.

Interestingly, Peter's use of language in the answer to c) shows that he has made sense of the term "distribution" in an individual way: He argues that the distribution "is simply too big". In our analysis, two possible interpretations have emerged: Peter either uses the term "distribution" in a different way here than in a) and b), which may suggest that the connection of the word "distribution" to a statistical concept is still relatively loose, or he sees the word and concept of "distribution" like shifting balls between rows in the diagram, so that a perfectly uniform distribution would then be a 'small distribution' (because there would then be little concentration of balls in any of the categories), whereas an accumulation of all balls in one category would then be the 'highest possible distribution'.

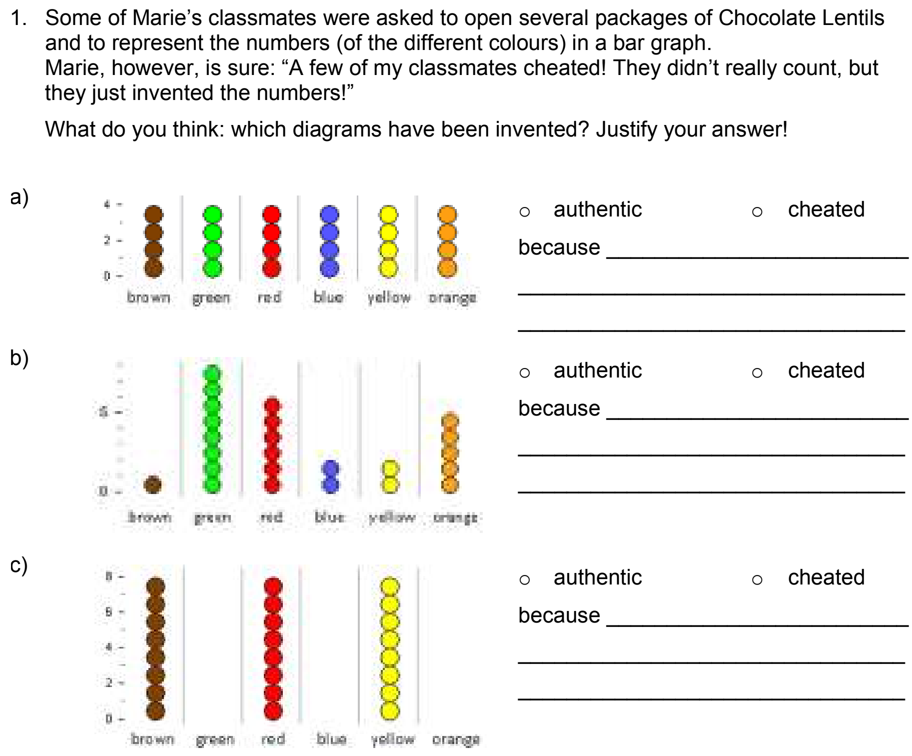
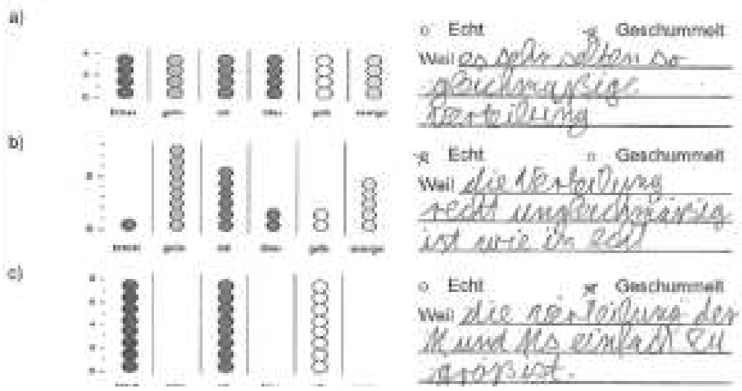


Figure 1: One of the „Chocolate Lentils" tasks

In the case of Peter's answer to question c), it is very probable that he has a correct statistical understanding of the situation without however perfectly mastering the language side in the sense that he would be able to share his thoughts using the domain-specific vocabulary adequately. Both interpretation alternatives further suggest



[cheated] because "very rarely such a uniform distribution" [occurs]

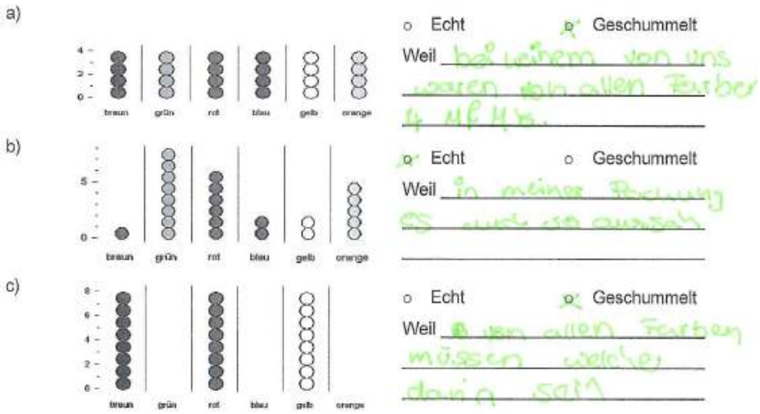
[authentic] because „the distribution is fairly non-uniform like in real".

[cheated] because "the distribution of the M&Ms is simply too big".

Figure 2: Peter's answers

that the language use in his answer not only points to a need of some further refinement of conceptual knowledge related to the notion of ‘distribution’, but also that knowledge about the use of language itself should be supported in order to enhance conceptual knowledge.

The example of Peter’s answer also gives insight into the development process of domain-specific language together with conceptual knowledge. Whereas Peter uses the notions of uniform and non-uniform distribution showing and applying corresponding conceptual knowledge, his answer to c) (“big” distribution) shows that the development of his domain-specific language concerning the term “distribution” has not been completed yet. In equal measure, he might not be familiar with the underlying concept, which supports the assumption that the development of conceptual knowledge goes together with a development of language.



[cheated] because “nobody of my group had 4 Chocolate Lentils of every colour.”

[authentic] because „ because my package looked like this. “

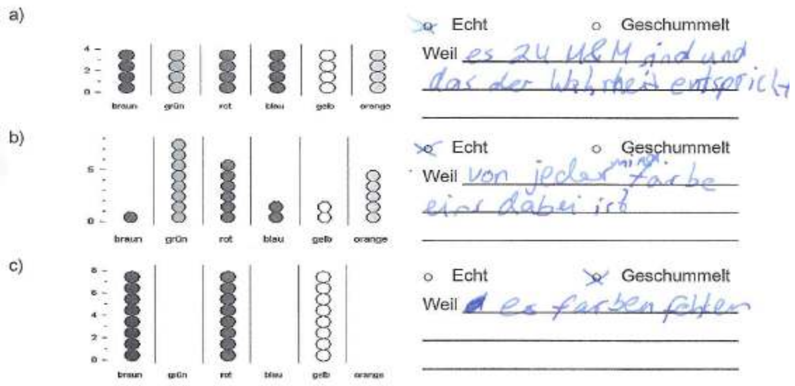
[cheated] because “ there have to be some (Chocolate Lentils) of every colour.”

Figure 3: Anne’s answers

Anne (see Figure 3) answers question a) and b) by referring to her own experience with packages of chocolate lentils. She gets into the problem in a very elementary way—we do not know whether she has the statistical understanding related to the concept of probability, i. e. whether she sees that it is quite improbable to encounter a package with an exact uniform distribution without any variation like in the case of a). She might have an intuitive and experience-driven pre-concept in this area but is not able to express her ideas through more specific terms. Comparable to this example, she appears to have recognised a similarity between her package and the distribution shown in case b) but she lacks describing this phenomenon. Very probably, she had not found identical frequencies in her package in the prior experiment, but the similarity she sees is a structural one, with frequencies showing statistical variation. Anne obviously has the language ability for describing her experience, however she might still not completely be able to express her thoughts beyond the specific experience.

At the first sight, Anne’s answer to c) seems to reveal a lack of conceptual knowledge, as she claims that all colours must be there in each package. In particular, prior to working on this task she had access to packages with missing colours. This makes it more likely to interpret her answer as an expression of irritation by the balanced distribution that might not have appeared “typical” to her. In contrast, in her written answer, she focuses on a different aspect. This utterance might hence be a case in which the statistical understanding diverges from the written answer, possibly as a consequence of a non-ability of expressing the statistical understanding in words. However, the analysis yielded this interpretation as one possibility among other possible interpretations.

Anne’s answers reveal a certain understanding (at least in the sense of intuitions, cf. Fishbein, 1975) of the statistical situation. However, she is not able to communicate her ideas neither using standard language nor using more general terms.



[authentic] because “ there are 24 Chocolate Lentils and this corresponds to the truth.”

[authentic] because „ there is one Chocolate Lentil of every at least colour.“

[cheated] because “ colours are missing.”

Figure 4: Tim’s answers

Figure 4 shows Tim’s answers. In contrast to the previous examples, Tim might have misunderstood the problem or have used an inappropriate logical structure. In case a) he argues that the distribution is authentic because the number of chocolate lentils in the shown package is 24, as stated previously in the learning material. His emphasis that “this corresponds to the truth” shows that this observation predominates over other possible considerations, so that he appears not to seek for other possible criteria for his judgement. Although he explains in a) his decision by the number of chocolate lentils, in b) and c) he changes his justification. However, again one single criterion is dominant, and Tim does not question whether the aspect is relevant. Now, his reason for the authenticity of the distribution is that “there is at least one chocolate lentil of every colour”. The order of the words in Tim’s answer (“at least” appears to have been inserted in the wrong place) indicates that he had later seen the need of inserting the “at least”, probably because he saw that the frequencies were not all equal to one. In this case, Tim has obviously tried to make the language of his answer more exact, without however developing a more deepened understanding e. g. of statistical variation.

Finally, Tim justifies the decision that c) is invented with the statement that “some colours are missing”. We do not know if he thinks that the author made a mistake and “forgot” these colours or if Tim considers such a distribution as atypical.

Both from the point of view of statistical understanding and language, Tim’s answers suggest a need of further developing SL. In this case, both the level of being able to communicate with statistical terms and the level of having corresponding conceptual knowledge show deficits.

As a result of this case-based qualitative bottom-up analysis, we derived categories from the interpretation process of selected cases. Figure 5 gives an overview of these categories. In these categories, qualifiers like “good” and “high” refer to the sample of the cases in our study and hence will have to be adjusted in potential adaptations to other samples.

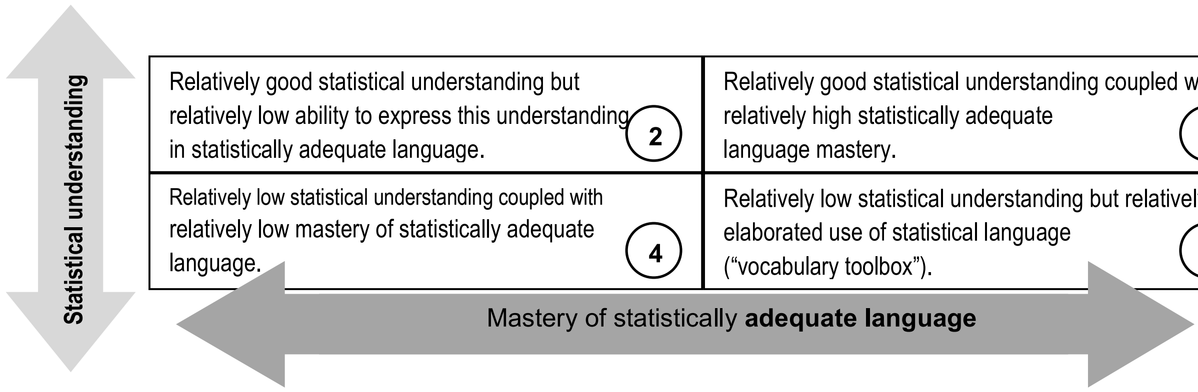


Figure 5: Bottom-up categories

In the third step of this study, a top-down coding procedure was done with all 83 cases of answers usíng the categories in Figure 5. The analysis was done by two raters independently (inter-rater reliability: $\kappa=0.86$). In all cases with different codes an a-posteriori agreement of the raters could be reached in a subsequent joint interpretative analysis. The analysis yielded that 52 out of 83 cases were assigned to category 2, which means that they have at least an intuitive understanding of the statistical situation, being however unable to express this understanding in adequate language. Fewer students were classified into the categories 1 (13 out of 83), 3 (1 out of 83), and 4 (17 out of 83). Whereas Peter’s answers (Fig. 2) can be seen as a case illustrating category 1, Anne’s (Fig. 3) and Tim’s answers (Fig. 4) have been classified into categories 2 resp. 4.

6 DISCUSSION AND CONCLUSIONS

Consistent with the findings of Makar and Confrey (2005) the examples presented above as well as the results of the top-down analysis of the whole sample show that the students frequently use every-day language to express their ideas related to the statistical problem considered. Often, the non-availability of adequate notions affected the quality of their statements. Moreover, the use of appropriate language and the statistical understanding appear to be interrelated and both appear to play an important role for the ways students approach statistical problems. In the sample of this study, the language development tended rather to have lagged behind the development of an intuition or understanding. In more general terms, both areas—the refinement of statistical understanding on the one hand and the domain-specific language development on the other—appear as almost two sides of a medal, for their complementary importance for conceptual learning. Figure 6 gives a schematic overview of this complementarity.

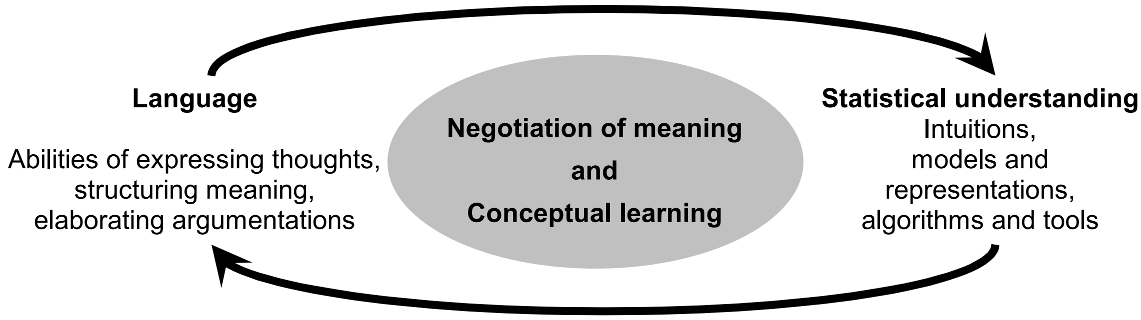


Figure 6: Language, statistical understanding and conceptual learning

Expressing statistical thoughts by means of written language often requires reorganising and deepening these thoughts—and it is hence a learning opportunity. In particular, misconceptions and incongruencies in statistical knowledge often get apparent and can inform teachers and students about the stage of the ongoing learning process.

For building up statistical thinking and hence competencies in the area of SL in the classroom, the findings raise the question how a combined instructional focus on the development of conceptual knowledge in statistics and language can be implemented. We conclude that an explicit awareness of language can also foster the students' statistical understanding. In a follow-up study, effects of corresponding learning materials with an emphasis on language use will be examined.

ACKNOWLEDGEMENTS

The research project ReVa-Stat ("Developing concepts of data-related reduction and statistical variation as a support for building up SL") is supported by research funds of Ludwigsburg University of Education.

REFERENCES

Ben-Zvi, D. & Garfield, J. (2004). Statistical Literacy, Reasoning, and Thinking: goals, definitions, and challenges. In Ben-Zvi, D. & Garfield, J. (Eds.) *The Challenge of Developing Statistical Literacy, Reasoning and Thinking,* 3-15. Kluwer.

Biehler, R. (1997). Student's difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In Garfield, J. & Burril, G. (Eds.) *Research on the role of technology in teaching and learning statistics*, 169-190. Voorburg, The Netherlands: ISI.

Chance, B.L. (2002). Components of Statistical Thinking and Implications for Instruction and Assessment. *Journal of Statistics Education, 10(3)* [retrieved on 23th November 2012 from www.amstat.org/publications/jse/v10n3/chance.html].

Curcio, F.R. (1987). Comprehension of Mathematical Relationships Expressed in Graphs. *Journal for Research in Mathematics Education, 18(5)*, 382-393.

Eichler, A. & Vogel, M (2012). Basic modelling of uncertainty: young students' mental models. *ZDM Mathematics Education, 44*, 841-854.

Engel, J. & Vogel, M. (2005). Von M&Ms und bevorzugten Farben: ein handlungsorientierter Unterrichtsvorschlag zur Leitidee Daten und Zufall in der Sekundarstufe I. *Stochastik in der Schule 25(2)*, 11-18.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children.* Reidel: Dordrecht-Holland.

Gal, I. (2003). Expanding conceptions of statistical literacy: An analysis of products from statistics agencies. *Statistics Education Research Journal, 2, 3-21.*

Kuntze, S., Lindmeier, A., & Reiss, K. (2008). "Using models and representations in statistical contexts" as a sub-competency of statistical literacy – Results from three empirical studies. *Proc. of the 11th Int. Congr. on Math. Educ. (ICME 11)*. [retrieved on 5th September 2012 from http://tsg.icme11.org /document/get/474].

Kuntze, S., Engel, J., Martignon, L. & Gundlach, M. (2010). Aspects of statistical literacy between competency measures and indicators for conceptual knowledge – Empirical research in the project RIKO-STAT. In C. Reading (Ed.), Proc. of the 8th Int. Conf. on Teach. Stat. (ICOTS8, July, 2010), Ljubljana, Slovenia. Voorburg, NL: ISI. www.stat.auckland.ac.nz/~iase/publications.php [Refereed paper].

Makar, K. & Confrey, J. (2005). "Variation-Talk": Articulating Meaning in Statistics. *Statistics Education Research Journal, 4(1)*, 27-54.

Mayring, P. (2000). Qualitative Inhaltsanalyse. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [Online Journal],1*(2). [retrieved on 23th November 2012 from http://qualitative-research.net/fqs/fqs-d/2-00inhalt-d.htm].

Shaughnessy, J.M. (2007). Research on statistics learning and reasoning. *Second Handbook of Research on Mathematics Teaching and Learning*, 957-1009. Greenwich, CT: Information Age Publishing, Inc. and NCTM.

Wallman, K. (1993). Enhancing Statistical Literacy: Enriching our Society. *Journal of the American Statistical Association, 88(421)*, 1-8.

Watson, J. (1997). Assessing Statistical Thinking Using the Media, In I. Gal, & J. Garfield, (Eds.), *The Assessment Challenge in Stat Education* (pp. 107-121). IOS.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2(2)*, 3-46.

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical inquiry. *International Statistical Review, 67*, 223-265.