# EMPIRICAL RESEARCH ON UNDERSTANDING PROBABILITY AND RELATED CONCEPTS – A REVIEW OF VITAL ISSUES

Manfred Borovcnik
Alps-Adria University Klagenfurt, Klagenfurt, Austria
manfred.borovcnik@uni-klu.ac.at

*The 2007 ASA report renewed the debate on qualitative vs. quantitative statistical methods in educational research and promoted the randomized controlled experiment (RCT) as standard. While quantitative methods have their value, the target of the two approaches is completely different and RCTs will not provide reliable evidence in cases when the researcher is interested in how students understand concepts and why some fail, or which teaching methods are helpful for whom. This paper focuses on issues for improving qualitative educational research. Task analysis is shown to be one major factor. In probability and statistics such an analysis is more basic than in other branches of mathematics as tasks are often – unintentionally and unnoticed – ambiguous.*

## INTRODUCTION

The ASA report 'Using Statistics Effectively in Mathematics Education Research' (Scheaffer & Smith, 2007) suggests the randomized controlled experiment as gold standard. Randomized means the units are randomly selected or randomly attributed to different "treatments"; controlled means that an intervention group has to be compared to a control group to filter out the treatment effect by (group) differences. Item response models are recommended to analyse items used for assessment (p.33); we discuss a contextual item analysis (task analysis) and its purpose in a section below. At first sight the recommendations summarize good practice of research. The ASA report is a challenge for qualitative research as it criticizes its basic positions and propagates a new gold standard instead. It dismisses the predominant qualitative approach in education research as inferior and relocates its role to exploratory studies (p.6). "Once small-scale [qualitative] research studies have examined phenomena through observation or intervention, more comprehensive [quantitative] studies can be mounted that seek to generalize [...]." The goal of this paper is to take up this challenge but find solutions *within* a qualitatively dominated paradigm. Besides general quality criteria, task analysis will play a major factor to ensure authenticity of research. This approach was suggested in the 1990's and there has been an extensive elaboration of concepts. In probability and statistics, task analysis pursues a different purpose. It is not about how to solve a problem and investigate the complexity of the methods used to derive a solution. It is more crucial because tasks often are ambiguous even though this is neither intended nor noticed. This resembles – to some extent – the mutual relation of probability to its (subjective and objective) interpretations.

Scientific methods are prone to trends. Quantitative methods in education were extensively used in the 1960' and 70's. It was not so much use of RCTs but of factor analysis and regression methods. Statistical methods work also for non-random samples as any bias resulting from the present sample could – in theory – be attributed to covariates. From the 1980's quantitative methods lost favour to qualitative methods, which seemed more appropriate to researchers to cope with research questions of why and how students do understand concepts learned in a specific form or another approach and not only of that approach *x* is better than *y*. (By the way, better for whom?)

From the ASA report (Scheaffer & Smith, 2007, p. 4), we can identify the *No Child Left Behind Act* as an impetus for quantitative methods: "[This act] calls for scientifically based research, which the act defines as 'research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs.' " Funding organizations and official boards have pursued an active policy of funding projects that comply with the new guidelines (What Works Clearinghouse, 2008; National Mathematics Advisory Panel, 2008). In the UK, Goldacre expressed concern about bad quality in education research and propagated the paradigm of quantitative research instead (Haynes, Service, Goldacre, & Torgerson, 2012). Goldacre extensively compared education with health. "RCTs are the universal means of assessing which of two medical treatments works best […] RCTs play a vital role in demonstrating not only the effectiveness of an intervention, but also value for money." (pp.13). However, quite a few measurements in medicine are not evaluated by RCTs but are ex-

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org 

changed between experts (e.g., favourable operation techniques) and we learn only recently that many medical drugs have quite a different impact on gender or various ethnicities.

Researchers in education reacted reluctantly but expressed their concern about improving the quality of qualitative research. For probability and statistics, a special issue of Statistics Education Research Journal was devoted to qualitative research (Petocz & Newbery, 2010; Groth, 2010; Kalinovski, Lai, Fidler, & Cumming, 2010). General arguments against a pure quantitative approach with RCT as golden standard and quantitative methods are amongst others:

i. There is no best teaching strategy or intervention for all! The search for a best teaching intervention is somehow comparable to the Nurnberg trichter. ii. The conditions of a test always influence all stakeholders; things are different under everyday conditions. iii. Short-term effects may differ from long-term implications. iv. Why not develop a qualitative view of what intervention is better, is it only instrumental what we learn? Maybe interrelated knowledge bases will be an advantage only later but not help to succeed in a short-term test. v. How to test the success of any intervention? One establishes criteria by a test construct; this test construct defines success. As it could be different, the question is whether the success remains stable under various criteria? vi. How to convince teachers of an intervention, which has been found best? vii. Learning is a social and interactive process; for a long time we learn in a class and interaction with peers has a great impact on the effect of any teaching intervention. Surely, individuals cannot be randomly selected and put together to an experimental class. viii. The alternative is to randomly select classes and use multi-level statistical models to account for that. The statistical models used get quite complicated and do not allow for detailed research questions. Furthermore, the experiment could then only be performed on a large scale with much funding, which reduces the circle of people who can do research. It would take people too long to "learn" the model and the underlying assumptions. Moreover, without this expertise, it seems impossible to criticize the results from outside. ix. Research will be pooled within a few expert groups who gather the "required" expertise.

We do not argue against RCTs per se. Nevertheless, it is a clear fact that most of what we know about teaching probability and statistics, we know from philosophic, historic, and mathematical analysis and – from qualitative studies. There are few quantitative studies using factor analysis to detect communalities in students' perceptions. Maybe a mixed form of qualitative and quantitative methods will provide an alternative. However, qualitative research has to accept the challenge, explain its positions, and improve the approach. Badly performed qualitative research is no argument for switching to quantitative methods in the same way as deficiencies of the quantitative approach are no excuse to perform a badly designed qualitative study. We try to improve the qualitative approach by our present discussion.

## A "DISCUSSION" ON THE QUALITY OF QUALITATIVE RESEARCH

The following text is a fictional but illustrative discussion between two academics on a research paper. One is the editor of a journal and the other a referee. The design of the involved study – as far as it is relevant – will be clear from the details. The views on quality revealed in this discussion indicate difficulties to arrive at a consensus on criteria to judge qualitative research.

*Editor to author.* We *reject* the paper. We see your methodology as being well grounded. You clearly state the context for your study, so we do not see any problem with it being based primarily on open-ended responses to only two items. The main issues are, however, that there are viable alternative explanations for the observed responses. As a side remark, we do not share the referee's critique that item 2 is ambiguous. All studies have limitations but your manuscript does not identify them properly. If you address these issues you might produce an acceptable paper.

*Referee.* The present study is a qualitative study with seven test persons on two items (investigated by an interview) and links to another study with the same persons where two equivalent items have been posed in written form. One major research question is whether incorrect responses on the written test are maintained in an interview situation.

Item 2 of the interview is open to several *feasible* task reconstructions. The confrontation of the interviewees with a fictional solution may recall perceptions of the task different from that what was expected by the researcher. From a given interview passage one can see that the test person in fact became aware that such a different perception of the task is possible but obviously could not solve this new task. From then on, the interviewee only complained about the difficulty of

probability and the confusion that this task has caused. It cannot be excluded that other participants have been confused too. Thus, to establish links between the written and oral test can only be based on mere speculation. Of course, it is difficult for test persons to speak about their task perception and it is difficult for the researcher to recognize the actual task perception. However, an interpretation of the answers may be meaningless if the status quo of the test persons is unknown.

*Referee's reaction to editor's decision letter*. For the task of interpreting the ideas behind any answer to an item, the researcher has to make sure that interviewee and interviewer are *communicating about the same issue*. The researcher is fixed on *one and only one* solution (while there are several); this confuses the test person more and more and blurs further statements. We may call it a breakdown of (rational) communication. The research design uses *pairs of items* in a different format and investigates, whether patterns of behaviour (in a written test) change if the test persons are confronted with a hypothetical statement of a third person in the interview. This makes sense only if the elements of the pairs are equivalent. As this is not the case, a comparison gets doubtful. As the flaws have had *their impact on the progress of the interview*, there is no way to repair the study. How to filter out how the students would have reacted without the methodological flaw?

*Editor on qualitative studies*. The authors conducted a primarily *qualitative* study, and they have not provided a thorough consideration of the restrictions of their study. In order to produce a publishable manuscript, they need to attempt to effectively recognize these considerations, use them in their analyses, and also in their reflections on the results. As I wrote, *every study*, even one with quantitative measures, a balanced design, a clear control group, and random assignment, *has limitations*. The results often do not have broad generalization or applicability because the well-controlled design is too artificial. All of these limitations need to be recognized when the researcher interprets findings. We can never have perfect information and from which unquestionable conclusions can be drawn for any study of human behaviour. As for qualitative research, there are many aspects to commend in the design of the author's study. However, I do not agree that the study being has too many flaws to produce meaningful information. I believe you misunderstand the purpose of the interview items. They were not designed to be isomorphic with the items on written test. They were designed to elicit the participants' reasoning and thinking.

*Reviewer on the difficulty to interpret answers*. First, if there are more ways to reconstruct, the researcher has to be well aware which one the student is actually using and *not* influence the interview on the basis of his or her wrongly restricted perception of the situation. Alongside a change of perception as caused by the ambiguous hypothetical statement, the test person did recognize that the task could be read completely different from previous reading and this new task was more than he could cope with (in fact, this new task is quite difficult). Second, the answers to the written and oral tasks *are analysed in pairs* as was also formulated in the research question. How can such a link sensibly be established if the items are far from being equivalent to each other? Such issues apply whether the research is qualitatively done or inferential methods are applied.

*The editor on the difficulty to design unambiguous tasks*. Whether one is conducting quantitative or qualitative research, the researcher needs to know how the participant is interpreting the task. In developing assessment items, I have found how difficult it is to create an item so that every single person has exactly the same interpretation of what is asked by the item. Even among experienced scholars you can find slight differences in how the same context, problem, or question is interpreted. Usually, the *different interpretations indicate a misunderstanding or lack of understanding*. Therefore, the unexpected interpretations, as revealed through students' responses, surely provide insightful information on students' thinking and understanding.

*A final rejoinder of the referee*. It is not about unambiguous tasks. It is about biasing the interview if the researcher is not prepared for other task reconstructions. This has a negative effect on the "rationality" of the communication. Both speak in a different world with no bridges between; both think of the other to be irrational. Or, the interviewee judges questions from the interviewer as inadequate. What can one expect from such an interview and how to link this extraordinary situation to the student's understanding of the concepts? While the tasks may always be ambiguous, the researcher has to be prepared for various different perceptions of it. This is especially important for the interview technique where the interviewer interacts with the test person. Finally, in the concluding phase of interpreting the behaviour of test persons, there has to be a clearer interpretation of the answers. Task analysis prior to the testing phase may enrich the perception of the researcher. In

probability, a lack of consensus on the task is quite usual. By the way, if experts arrive at different perceptions, is it due to their lack of expertise or does it reflect their different approaches?

## TOWARDS BETTER STANDARDS IN QUALITATIVE RESEARCH

Journals should establish and promote stricter criteria for qualitative research. A starter could well be the ASA guidelines on quantitative research. A review like Gal & Ograjenšek (2010) can offer orientation on research design and methods. From his experience as an applied statistician and reviewer for journals, the author considers the following issues as vital for qualitative research.

*Research problem*. The study question is often vague. By an intervention the students will understand a concept, or they can solve specific tasks better. Which kind of tasks, in which formulation, in which context, using which tools? Short- or long-term effects? A systematic study of potential influence factors on the target variable is often missing. An impact on the target variable can be due to (uncontrolled) confounders so that the conclusions are wrong; e.g., if a new type of tasks is used for learning, its effect on success in exams may be due to monitoring the students (by online quizzes) and get them to permanent work during the semester and not to the superior format of the tasks. Obviously, continuous surveillance of students is a confounder that might outweigh the influence of the format. The effect of an intervention is a change of behaviour, which need not necessarily establish a success. E.g., in studying the effect of a teaching unit supported by specific software, the students might be able to solve certain tasks better (than before or better than a control group) and show that their thinking is influenced by the software. Vital questions are: Are the students able to transfer such skills to other software or are they "hindered" to use it effectively? Does sequencing of a task into steps 1, 2 etc. to solve it by this software help to understand the problem, the solution method, the restrictions on interpretations, and the concepts involved?

*Design*. The design of a study establishes the benchmark for success. To measure the success of an intervention, the study group could be compared to a control group. This design is too rarely used in qualitative research. However, learning is also signified by group effects and it is hard to make the groups comparable. The success might also be measured by the difference in scores between post intervention tests and prior tests. An intervention, especially if it takes a longer time, should have an impact on thinking and solving behaviour. Is the difference of achievement between post and prior tests higher than could be expected? The question is how to judge that.

*Selection of subjects*. The relation of an investigated group to a target population is often unclear. The difficulties increase if self-selection is operating; volunteers might cause serious restrictions on judging how the study group compares to the target population. Participants could be better in objective capacities, more interested in the subject, better motivated, etc. Participants could also be attracted by some payment (or gift), which would make matters more complicated.

*Systems analysis: target and influential variables and instrumentation.* Measuring success of an intervention by self-reporting lacks external benchmarks. If concepts should be acquired, if capacity to solve specific tasks should be enlarged, this has to be checked by adequate measures. Are the tasks used to measure success representative (valid)? Are the tasks solved better because *similar* tasks have been used in the intervention while a transfer to other formulations or contexts would not be feasible? Does a correct solution indicate that the concepts and results are well understood? The effect of a longer intervention should be visible shortly after the intervention: the participants have got accustomed to some way of expressing the ideas. What about long-term effects?

*Interpreting and generalizing the results*. The (teaching) implications of a study often coincide with the current views and the authors miss to ground them on their empirical results.

## TASKS AND TASK ANALYSIS

In qualitative research, the demand is much higher than in the quantitative approach as it is aimed at investigating why people think and behave as they do. It goes beyond the question that one approach is "better" than another. A main challenge is to investigate how the test persons think when they answer questions or solve tasks. This could be done by audio-taped screen-videos, or, e.g., by semi-structured interviews. If the view on test persons' thinking is restricted in the analysis, then researchers are prone to misunderstand and misinterpret the actual thinking processes. In an interview, even worse, they influence the course of the interview and the thinking processes of the test persons. That is why task analysis gets so important in qualitative research.

Tasks play an eminent role in mathematics education research as well as in teaching, and are even more important in empirical research on understanding the concepts. If specific tasks are solved by the majority of test persons it is interpreted as an indication that the concepts and strategies involved in solving the task are familiar to them. Tasks establish a benchmark for testing whether a teaching intervention has been successful in an empirical study; they operationalize the researchers' hypotheses of *what* constitutes understanding the underlying concepts. Tasks may be analysed statistically by item analysis (reliability, difficulty, etc.). They also may be analysed conceptually, i.e., whether they are valid instruments for conceptual understanding and whether they allow for predictions of adequate solving behaviour for other tasks related to the concepts involved.

Tasks and task systems have been analysed in educational research since long. The main focus of investigations has been on their function within a specific learning situation. The approach goes back to Brousseau (1984). Quite a few papers have been written by Stein (e.g., Stein, Grover, & Henningsen, 1996) and the results are used to improve teaching practice or teachers' expertise. Task analysis has found attention more recently with a stronger access on tasks as key elements for teaching. Conceptual relations are often too complex and generally oriented (include all incidences where the concepts apply, investigate all relevant properties of a concept, etc.) – logical considerations become a key for investigating and learning the concepts. Tasks are used as local surrogate for the concepts, building up these concepts in the learner's cognitive system in a bottom-up direction from specific features, contexts, and situations to general relations and overall strategies, i.e., to mathematical properties of the concepts and the concepts' position within a theoretical framework. The ICMI study 22 (Margolinas, 2013) represents the latest status of task design. It ignores the purpose of task analysis of Borovcnik, & Bentz (1991), which we will outline below.

A thorough task analysis might have prevented the breakdown of communication in the interview above when the test person recognized that the task could be perceived quite differently from his previous understanding and the interviewer did not recognize that. The interviewer insisted on the simple and unique task while the interviewee saw a legitimate reconstruction of the task, which makes the offered explanation by a fictional student plausible but an explanation or even solution of the "new task" was far off the limits of the interviewee. The conflict in the interview and the inadequate interpretation of it later as conceptual error of the test person would have been avoided if a task analysis had been performed prior to the test.

Methods may be applied routinely to get correct answers without any conceptual understanding. Likewise, idiosyncratic conceptions may lead to correct answers. Tasks are often more difficult if the text is longer; if redundant information is given it gets worse; if there is no solution then the confusion is more pronounced. If wording is varied slightly, the difficulty of the task may change drastically. All this may seem well-known from educational research. The usual approach investigates the difficulties of tasks under the overall assumption that there is a unique task. A number of "mathematical tasks used [...] was analyzed in terms of (a) task features (number of solution strategies, number and kind of representations, and communication requirements) and (b) cognitive demands (e.g., memorization, the use of procedures with [and without] connections to concepts, the 'doing of mathematics')" (Stein, Grover, & Henningsen, 1996). In stochastics, often there is no unique task. This is an ongoing source of confusion despite a careful analysis in Borovcnik & Bentz (1991). Even the simplest tasks have no normative view. Moreover, language plays a more subtle role than in other branches of mathematics (Kapadia, 2013).

Task analysis in probability has also to face the consequences that probability deals with ideas in a *virtual* world with sparse connections to the material world. A communication on probabilities between a test person and an interviewer might easily lack rationality. Sometimes the test person feels "forced" to justify something for which one cannot give reasons. As a typical example, we refer to the action-reflection conflict from Borovcnik & Bentz (1991). Which orientation – *action* (decision) or *reflective* – is (unconsciously) guiding the answering behaviour. For example, for the fifth toss of a coin, the probability is the same for heads and tails; despite such a reflective insight, the test person might still strongly prefer tails (or heads) after an assumed result of four heads in a series. Neither the person nor the interviewer might be aware of this clash between modes of thinking. If asked for a justification for tails (heads) in the fifth toss, the test person might *switch back* from action (decision) to reflection and see that there is no argument for the actual choice and regard any demand to *justify* the first choice as irrational. The subsequent communication probably

lacks rational criteria that would allow for an interpretation indicating this or another conception or misconception. Usually it is interpreted as misconception of the law of large numbers.

Even harder to handle are the different concepts of probability. With the competing conceptions there are irreconcilably different intuitive perceptions associated, which are often confused (Borovcnik, 2012; or Borovcnik & Kapadia, 2014). Some intuitions are better received within a frequentist framework while others are close to the Bayesian concept. They lead to a different perception of tasks, interpretations of concepts used and results arrived at. Even simple tasks are ambiguous. In Borovcnik & Bentz (1991) feasible task reconstructions are developed. Coin tossing is usually modelled as independent trials with the same probability of ½ in each trial. However, a longer series of heads establishes empirical evidence that ½ should not apply here. There is a trade-off between a strong hypothesis (as ½ for the coin) and the evidence from a sample. If the latter "contradicts" the first, then it is reasonable to change the model. However, such a test person fails in a normatively perceived task. Borovcnik & Bentz also discuss a bag item, which leads to an unsolvable task. Depending on the framework, the same objects get a completely different meaning. This complicates an evaluation of behaviour and statements even in the simplest tasks.

*Final statements*. Rather than refute the ASA guidelines as inappropriate we should use them to improve qualitative research. General criteria may be developed from the ideas indicated here and from Gal & Ograjenšek (2010). For the purpose of investigating how test persons think, we recommend a conceptual analysis of the used tasks in the sense of Borovcnik & Bentz (1991).

REFERENCES

Borovcnik, M. (2012). Multiple perspectives on the concept of conditional probability. *Avances de Investigación en Didactica de la Mathemática*, 2, 5-27.

Borovcnik M. & Bentz, H.-J. (1991). Empirical research in understanding probability. In R. Kapadia, & M. Borovcnik (Eds.), *Chance encounters* (pp. 73-105). Dordrecht: Kluwer.

Borovcnik, M. & Kapadia, R. (2014). A historical and philosophical perspective on probability. In E.J. Chernoff, & B. Sriraman (Eds.), *Probabilistic thinking. Presenting plural perspectives*. *Advances in mathematics education*. New York: Springer.

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H. Steiner (Ed.), *Theory of mathematics education* (pp. 110–119). Bielefeld: Institut für Didaktik der Mathematik.

Gal, I., & Ograjenšek, I. (2010). Qualitative research in the service of understanding learners and users of statistics. *International Statistical Review*, *78*(2), 287–296.

Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. London: Cabinet Office.

Kapadia, R. (2013). The role of statistical inference in teaching and achievement of students. *Proceedings of the 59th World Statistics Congress*. Voorburg: ISI. (6 p.).

Margolinas, C. (Ed.) (2013). Task design in mathematics education. *Proceedings of ICMI Study 22* (Vol. 1). Oxford, Oxford University.

Groth, R.E. (2010). Situating qualitative modes of inquiry within the discipline of statistics education Research. *Statistics Education Research Journal, 9*(2), 7-21.

Kalinowski, P., Lai, J., Fidler, F. & Cumming, G. (2010). Qualitative research: An essential part of statistical cognition research. *Statistics Education Research Journal*, *9*(2), 22-34.

National Mathematics Advisory Panel. (2008). *Foundations for success*: *The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Petocz, A., & Newbery, G. (2010). On conceptual analysis as the primary qualitative approach to statistics education research. *Statistics Education Research Journal,* 9(2), 123-145.

Scheaffer, R., & Smith, W. B. (2007). *Using statistics effectively in mathematics education research: A report from a series of workshops organized by the ASA with funding from the National Science Foundation*. Alexandria, VA: American Statistical Association.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455–488.

What Works Clearinghouse. (2008). *Procedures and standards handbook.* Washington, DC: U.S. Department of Education Institute of Education Sciences.