The Research Council of Norway

Mixed Methods in Educational Research

Report from the March Seminar 2012

Programme

Norwegian Educational Research towards 2020 - UTDANNING2020

About the programme

Norwegian Educational Research towards 2020 - UTDANNING2020

The Programme for Educational Research – UTDANNING2020 (2009–2018) succeeds the Programme for Knowledge, Education and Learning – KUL (2003–2007) and runs parallel with the Programme for Practice-based Educational Research – PRAKUT (2010–2014). The programme has a total budgetary framework of approximately NOK 400 million.

The UTDANNING2020 programme conducts research on the education sector – from early childhood education and care to doctoral level education. The programme seeks to enhance the knowledge base for policymaking, public administration, professional education and professional practice and to promote research of high scientific merit. A variety of subject areas and research communities are encouraged to conduct research on issues related to the education sector as well as areas of overlap in other sectors.

Programme Board

Contents

Introduction

Mixed Methods in Educational Research

The quality of the educational system and infrastructure is central to every nation's economy, development, social integration and well-being. The quality of education depends and builds on the quality, rigour and relevance of available educational research. It is therefore of critical importance to secure and raise the standards for conducting research in order to improve education. The research programme Educational Research towards 2020- UTDANNING2020 is designed to address and challenge scientific merits, multidisciplinarity, rigour and relevance in educational research.

The long-term duration of the programme (10 years) facilitates the possibility of addressing factors which are critical to fostering high quality educational research, improving scientific merits within this field of investigation and enhancing the capacity of scholars, enabling them to produce such high quality research.

In order to promote scientific quality and merits, UTDANNING2020 uses different tools and strategic actions. Funding of high quality research relevant to the educational sciences holds a key position in this tool kit. Through a rich portfolio of varied and intriguing research projects the programme aims to contribute to new insight, accumulate knowledge, support methodological awareness and growth and contribute to fostering research capacity within the educational sciences.

Annual seminars and conferences as mechanisms for knowledge exchange and knowledge building are examples of other activities meant to foster quality in educational research. Within the programme these seminars and conferences are targeting different groups and audiences like policymakers and stakeholders, the teaching profession, researchers and other knowledge brokers. A special annual seminar held in March is devoted to addressing scientific and methodological quality in the educational sciences. The first March seminar took place in 2010, and the focus was on rigour and relevance in educational research. The seminar in 2011 focused on the role of theory in educational research.

This year, the seminar took a closer look at the use of mixed methods in educational research. Professor Stephen Gorard from the University of Birmingham gave a key-note lecture on challenges and possibilities in mixed methods. He reminded us of the fact that qualitative and quantitative methods are not incommensurable, and illustrated with real-life examples the need for mixing quantitative and qualitative data to answer any set of research questions relevant for todays' society.

Professor Lars-Erik Borge at the Center for Economic Research at Norwegian University of Science and Technology (NTNU) and Professor Bente Hagtvet at the Department for Special Educational needs at the University of Oslo commented on Gorard's argument. Furthermore, Project Managers representing research projects with funding from UTDANNING2020 shared their experience with combining different kind of data and using mixed methods in educational research. This report includes papers from the different contributors on this March seminar and we hope this report will evoke curiosity and interest in new developments in methods for doing educational research.

Oslo, October 2012

Kirsti Klette
Chair of the programme board

Stephen Gorard, University of Birmingham, UK

Mixed Methods Research in Education: Some Challenges and Possibilities

It was with great pleasure that I agreed to address the 2012 conference on mixed methods hosted by the UTDANNING2020 programme. My thesis was that what is usually called 'mixed methods' research in education is really just research in education. It is relatively easy to conduct, with many possibilities and few real-life challenges or barriers. What this paper tries to do is convey part of why this is so.

There are of course many different methods of investigation that could be said to be 'mixed' in any one study – interviews with documentary analysis, or multiple regression with inferential statistics, for example (Symonds and Gorard, 2010). However, for the purpose of this brief paper, the mixture is assumed to refer to those methods that have traditionally labelled 'qualitative' and 'quantitative'. For some reason, social scientists have long separated any data that involves counting or measuring from all data that involves anything else – text, conversations, observations, smells, drawings, acting, music and so on. I have no idea why. But such social scientists say that these two groups – numbers and everything else – are incommensurable, and require a completely different logic to use, and have un-matched criteria for judging research quality, and many other purported differences. Then, just to confuse things, some social scientists say that we can and should mix these forms of data – and that presumably they are not commensurable in combination, only in isolation if that makes any sense at all. It is no wonder that new researchers are confused, and that the potential users of social science evidence just ignore us. We live in a kind of la-la land.

In this paper, what I want to suggest to new researchers, and to remind more experienced ones about, is that none of the above is true. Methods are not incommensurable, and while they may legitimately be classified in a number of ways, these classifications should not become schisms. Starting with a consideration of a piece of real-life research, the paper argues that we should not separate numbers from every other form or data in the first place. Then, in terms of qualitative and quantitative data at least, we have nothing to mix. Because I do not separate the qualitative and quantitative approaches, what is termed mixed methods work just seems natural to me. It is, I contend, what anyone would do who wanted to answer any real set of research questions.

A real-life example

It is instructive to contrast how we, as researchers, sometimes behave when conducting research professionally with the ways we behave when trying to answer important questions in our personal lives. When we make real-life decisions about where to live, where to work, the care and safety of our children and so on, most of us behave very differently from the way we do as 'researchers'. If, for example, we were intending to purchase a house by paying most of our savings and taking out a mortgage for 25 years that is equal in size to many times our salary, then we would rightly be cautious. We would have many crucial questions to answer from the beginning, and would only go ahead with the transaction once assured that we had sufficiently good answers from what is, in effect, a serious piece of research. It is worth considering this example in some detail because it illustrates some fundamental issues about research in a very accessible way.

When purchasing a house, we will believe that the house is real even though external to us. And we will believe that it remains the same even when we approach it from different ends of the street, else why would we buy it? In these and other ways, we would un-problematically and without any trepidation just ignore the usual nonsense that is taught to new researchers as an essential preliminary to conducting research. In buying a house we would not start with epistemology, and we would not cite an 'isms' or Grand Theory. Nor would we need to consider the 'paradigm' in which we were working. We would not refuse to visit the house, or talk to the neighbours about it, because we were 'quantitative' researchers and did not believe that observation or narratives were valid or reliable enough for our purposes. We would not refuse to consider the size of the monthly mortgage repayments, or the number of rooms, because we were 'qualitative' researchers and did not believe that numbers could do justice to the social world. In other words, in matters that are important to us personally, there is a tendency to behave logically, eclectically, critically, and sceptically. We would collect all and any evidence available to us as time and resources allow, and then synthesize it quite naturally and without considering mixing methods as such. We are quite capable of judging whether the qualities of a house are worth the expenditure, for example.

If we really care about the research, as we would with buying a house, we naturally adopt what might be called a mixed methods approach. Why is it so different in academic social science then? One simple answer is that people do not care about their academic research in the same way. Another linked part of the answer is that many people purport to be doing research but in fact are doing something else entirely. I am not sure what game they are playing instead, as no one has told me the rules. But from the outside their research is similar to someone buying a house without having any idea of the price or size, or else buying it without any idea of its condition or location. Yet, education is an important applied field and the results of research, if taken seriously, can affect the lives of real people and lead to genuine expenditure and opportunity costs. So, it is quite clear that to behave like this in education research by eschewing one or more forms of data is unethical (Gorard 2002). The 'game' survives, I guess, simply because it is played by the majority, and so this majority also provides a high proportion of the peer-reviewers. Yet these reviewers are intended to prevent rubbish being published, public money being wasted and education suffering in practice, either by not having access to good evidence, or, worse, by having access to shoddy or misleading evidence.

Barriers to mixed methods

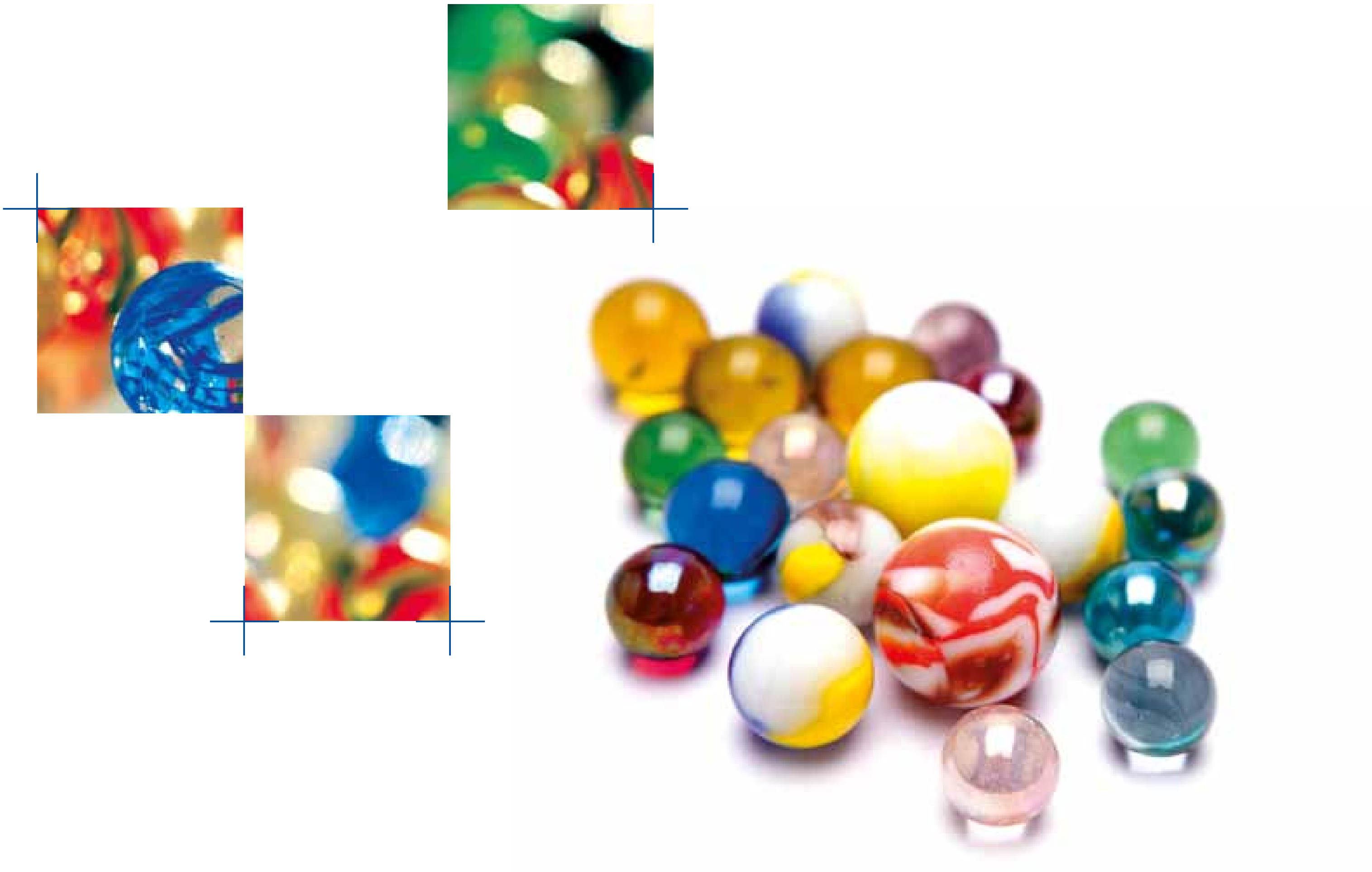
For me, that is the end of the matter, really. But I know from experience that readers will want more at this stage. So, the paper continues by briefly considering some of the self-imposed 'barriers' to using mixed methods, and why they should be ignored. One supposed barrier, the different way in which numeric data is usually analysed, is then used as an extended example of why these barriers are self-imposed and unhelpful. The final section of the paper suggests some models or approaches to synthesising numeric and non-numeric data. There is insufficient space here to deal with every supposed barrier and every forward-looking model. What are presented instead are selected examples, with references to further published examples.

First of all, the Q words are not paradigms. Types of data and methods of data collection and analysis do not have paradigmatic characteristics, and so there is no problem in using numbers, text, visual and sensory data synthetically in combination (Gorard, 2010a). Working with numbers does not, in any way, mean holding a view of human nature and knowledge that is different from when you work with text or shapes. In the sociology of science, the notion of a 'paradigm' is a description of the sets of socially accepted assumptions that tend to appear in 'normal science' (Kuhn, 1970). A paradigm is a set of accepted rules within any field for solving one or more puzzles – where a puzzle is defined as a scientific question to which it is possible to find a solution in the near future. An example would be Newton setting out to explain Kepler's discoveries about the motions of the planets. Newton knew the parameters of the puzzle and so was working within a paradigm. A more recent example might be the Human Genome Project, solving a closely defined problem with a widely accepted set of pre-existing techniques. The 'normal science' of puzzles in Kuhnian terms is held together, rightly or wrongly, by the norms of reviewing and acceptance that work within that taken-for-granted theoretical framework. A paradigm shift occurs when that framework changes, perhaps through the accumulation of evidence, perhaps due to a genuinely new idea, but partly through a change in general acceptance. Often a new paradigm emerges because a procedure or set of rules has been created for converting another more general query into a puzzle. None of this describes a schism between those working with numeric data and those working with

everything else. The notion of paradigm as a whole approach to research including philosophy, values and method is a red herring. It could be argued that commentators use the term ‘paradigm’ to defend themselves against the need to change, or against contradictory evidence of a different nature to their own. They damage social science by treating serious subjects like epistemology as though they were fashion items to be tried on and rejected on a whim.

The Q words do not define the scale of a study. It has been argued incorrectly, by Creswell and Plano Clark (2007) among others, that qualitative data collection necessarily involves small numbers of cases, whereas quantitative relies on very large samples in order to increase power and reduce the standard error. But this is not an accurate description of what happens in practice. The accounts of hundreds of interviewees can be properly analysed as text, and the account of one case study can properly involve numbers. Also, issues such as sampling error and power relate to only a tiny minority of quantitative studies where a true and complete random sample is used or where a population is randomly allocated to treatment groups. In the much more common situations of working with incomplete samples, with measurement error or dropout, or involving convenience, snowball and other non-random samples and the increasing amount of population data available to us, the constraints of sampling theory are simply not relevant (see below). The supposed link between scale and analysis is just an illusion.

The Q words are not related to research designs. What all rigorous research designs, and variants of them, have in common is that they do not specify the kind of data to be used or collected (Gorard 2013). No kinds of data, and no particular philosophical predicates, are entailed by common existing design structures such as longitudinal, case study, randomised controlled trial or action research. A good intervention study, for example, could and should use a variety of data collection techniques to understand whether something works, how to improve it, or why it does not work. Case studies involve immersion in one real-life scenario, collecting data of any kind ranging from existing records to ad hoc observations. The infamous Q words of qualitative and quantitative, and mixed methods approaches are therefore not kinds of research design. A study that followed infants from birth to adolescence, weighing them on 1st January every year, would be longitudinal in design. A study that followed infants from birth to adolescence, interviewing their parents about their happiness every year, would also be longitudinal. A study that did both of these would still be longitudinal, even though some commentators would distractingly and pointlessly categorise the first study as ‘quantitative’, the second as ‘qualitative’, and the third as ‘mixed methods’. In each example the design – ‘longitudinal’ or collecting data from the same cases repeatedly over a period of time – is the same. This illustrates that the design of a study does not entail a specific form of data to be collected, nor does it entail any specific method of analysis; nor does any method require a specific research design. These points are quite commonly confused in the literature, and even in many research methods resources. Such writings contribute to widespread misunderstanding of study design issues and their relationship to subsequent choice of methods. I wonder whether this confusion is sown deliberately to help the games-players evade the need for design in their own research, or to excuse their use of only qualitative methods.



One approach is not intrinsically more objective than another. Qualitative research, so its proponents argue, is supposed to be subjective and thus closer to a social world (Gergen and Gergen, 2000). Quantitative research, on the other hand, is supposed to help us become objective (Bradley and Schaefer, 1998). This distinction between quantitative and qualitative analysis is exaggerated, largely because of widespread error by those who do handle numbers (see below) and ignorance of the subjective and nature of numeric analysis by those who do not (Gorard, 2006). What few seem to recognize is that the similarities in the underlying procedures used are remarkable. Analytical techniques are not generally restricted by data gathering methods, input data, or by sample size. Most methods of analysis use some form of number, even if only descriptors such as 'tend', 'most', 'some', 'all', 'none', 'few', rare', 'typical', 'great' and so on. A claim of a pattern or relationship is a numeric claim, and can only be so substantiated, whether expressed verbally or in figures (Meehl, 1998). Similarly, quantification does not consist of simply assigning numbers to things (Gorard 2010b). Personal judgements lie at the heart of all research – in our choice of research questions, samples, questions to participants and methods of analysis – regardless of the kinds of data to be collected. The idea that quantitative work is objective and qualitative is subjective is based on a misunderstanding of how research is actually conducted.

The underlying logic of analysis is not different. The methods of analysis for text, numbers and sensory data are largely the same, consisting of searching for patterns and differences, establishing their superficial validity and then trying to explain them. Other commentators and methods resources may claim that there is a fundamental difference between looking for a pattern or difference in some measurements and in some text or observations. This unnecessarily complex view is based on a number of widely held logical fallacies that get passed on to new researchers under the guise of research methods training. I examine one of these very widespread errors in more detail.

A logical flaw in traditional statistics

At the conference, I asked the question: "What is the probability of being Norwegian if in this room?" Imagine that I was the only non-Norwegian among 100 people at the conference. Then the conditional probability of being Norwegian if in the room (pN|R) would be 99%. Anyone picked at random from the room would turn out to be Norwegian 99 times out of 100. I also asked the question: "What is the probability of being in this room if Norwegian?" Imagine that there were 99 Norwegians in the room from a total population of five million. Then the conditional probability pR|N would be 0.00002. I asked if these two probabilities were the same, and all agreed they were not. I asked whether if we were given one percentage in isolation we could work out the other percentage. All agreed that we could not. We would need also to know the number of Norwegians and the number of people in the room in total. That is, we would need complete information.

To make sure we agreed I conducted the same demonstration with a real bag of marbles. The bag contains 100 balls of identical size, of which 30 are red and 70 are blue. If someone picks one ball at random from the bag, what is the probability it will be red? This is a good example of a mathematical question that might appear in a test paper, and that has some applications in real-life, in gaming for example. We have perfect information about the size of the population of balls (there are 100), and the distribution of the characteristics of interest (30:70). Given these clear initial conditions it is easy to see that the chance of drawing a red ball from the bag is 30/100 (30%). It is almost as easy to see that the chance of drawing two red balls one after another (putting each back after picking it) is 30/100 times 30/100 (9%), or that of drawing two red balls at the same time is 30/100 times 29/99 (nearer 8.8%). Most people at the conference could either do these calculations or could see how they were possible.

Now consider a rather different problem of probability. The bag contains 100 balls of identical size, of two different colours (red and blue). We do not actually know how many of each colour there are. If someone picks a red ball at random from the bag, what does this tell us about the distribution of colours in the bag (beyond the fact that it must have originally contained at least one red ball)? It seems to tell us very little. There could be 30/100 red balls, or 70/100 or 99/100. The drawing of one red ball does not really help us to decide between these feasible alternatives. We certainly cannot use the existence of the red ball to calculate probable distributions in the population, because we do not have perfect information (unlike the first example). Yet this situation is much more life-like in being a scientific problem rather than a mathematical one. In social science we rarely have perfect information about a population, and if we did have it we would generally not bother sampling (because we already know how many balls are of each colour). The more common situation is where we have information about a sample (the colour of one or more balls), and wish to use it to estimate something about the population (all of the balls in the bag). No one in the audience was able to tell me anything secure or interesting about the balls remaining in the bag, under these conditions.

Put into the same terms as the first example, the conditional probability of drawing a red ball from the bag if there are 30 in the bag (pR|30) is nothing like the probability of there being 30 red balls in the bag if we pick one (p30|R). As in the first example, one could be large (99%) and the other very small (0.00002), or vice versa, or something in between. In the usual condition of research, rather than mathematical puzzles, where we do not know the number of red balls in the bag, the first probability is of no help in calculating the second. The audience agreed.

Yet, there seems to be almost a world-wide conspiracy to pretend that none of this is true when we conduct statistical analysis (Gorard 2010c). When social scientists conduct a significance test, they assume an initial condition about the prevalence of the characteristics of interest in the population and then calculate, in much the same way as for coloured balls, the probability of the observing the data they do observe. The calculation is relatively simple and can easily be handled by a computer. The analyst then knows, if their assumption is true, how probable their observed data is. For example, if they assume that there is no difference (the nil null hypothesis) between the scores of two groups in their population of interest, it is relatively easy to calculate the probability of achieving any level of apparent difference in a random sample of any size drawn from that population. This is the probability of the data given the null hypothesis (pD|H), and is what significance tests like t-tests compute. But who would want to know this figure? What the analysts really want is pH|D, the probability of the null hypothesis being true given the data they collected. As above, this is a completely different probability to the first. One could be small and the other large, or vice versa.

Yet statistical analysis as reported in education routinely confuses the two, by assuming that pD|H provides a good estimate of pH|D. So, the 'logic' goes, if pD|H is quite small, then pH|D must be also. But it is not true that a small value for pD|H must mean a small probability for pH|D. This step in significance testing is an error, and it remains an error however low pD|H is. The whole practice of significance testing from that stage on is incorrect and invalid. And this is true of all tests, and all other sampling theory derivatives, including standard errors, confidence intervals and complex modelling based on significance scores. Sampling theory itself, and the calculations derived from it, are not the problems here, as long as we are interested in pD|H. But no one is interested in that. As soon as we pretend that pD|H is equal to or closely related to the much more interesting pD|H, we have left the world of social science for that la-la land again.

Unfortunately for researchers there is no simple, push-button, technical way of deciding whether a difference or pattern observed in a sample would also hold for the wider population. But it does not really matter. We do not select random samples, or randomise cases to groups, in order to use statistical tests later. That would be like saying we use crockery when eating so that we can do the washing up later! We randomise in order to try and obtain an unbiased distribution of unknown variables, as well as measured ones, in the sample. If we have randomised in order to obtain unbiased sample(s), then we could later calculate pD|H (as above). But this is a largely fruitless exercise, partly for the reason already given,

but also because it does not answer the key question that is common to all analyses. This is: Is the difference, pattern or trend, large enough to be worth pursuing? This is the same question we would ask if we had population data, no sampling was involved, and we knew the population distribution without calculation of probabilities. It is also the same question we would ask if the sample(s) did not fit the requirements of sampling theory – where the sample is non-random in nature, or where there is any non-response or measurement error, for example.

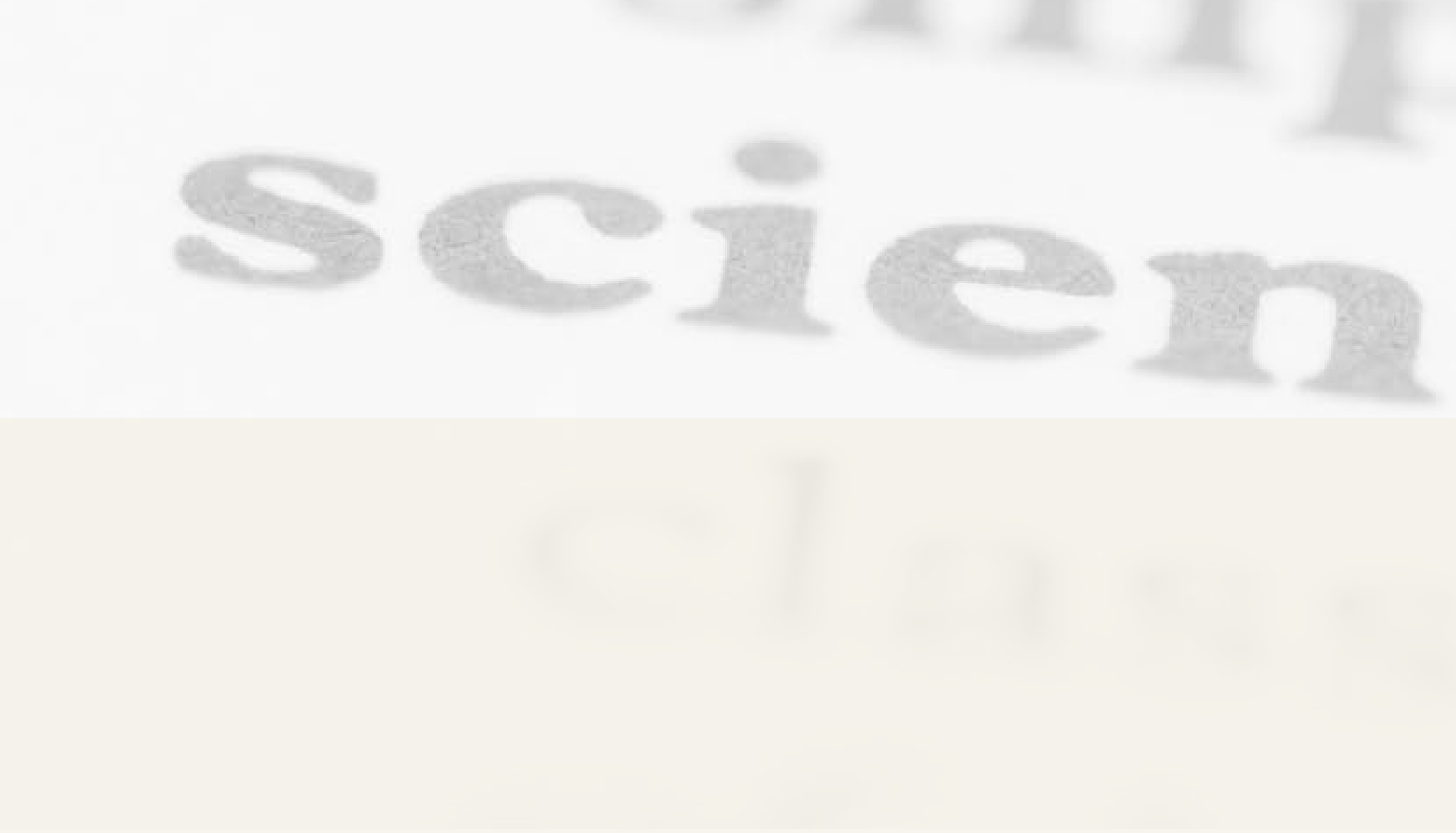
It is clear that, for any dataset, dividing the cases into two (or more) sub-groups will rarely yield exactly the same scores on all measures for both groups. It is unlikely a priori that the school pupils sitting on the left hand side of a classroom will have exactly the same average height as those sitting on the right. Their parents are unlikely to report drinking exactly the same average number of cups of coffee every day, and so on. A difference in scores or observations may, therefore, have no useful meaning at all. Whether a difference is more than this, and is actually substantial and worthy of note, can depend on a number of factors. It depends on the size of the difference in relation to the scale in which the difference occurs (an observed difference of two feet may be important in comparing the heights of two people, but not in comparing flight distances between Europe and Australia). It depends on the variability of all of the scores. It is harder to establish a clear difference between two sets of scores that have high levels of intrinsic variation than between scores in which each member of each group produces the same score as all other members of that group. The noteworthiness of a difference may also depend upon the benefits and dangers of missing a difference if it exists, or of assuming a difference if it does not exist.

All of these issues of scale, variability and cost are relevant even if the scores are measured precisely. But in reality, scores are seldom measured precisely, and common measures like test scores, self-esteem, aspiration, occupational class and ethnicity will be subject to a very high level of measurement error. Measurement error is nearly always a bias in the scores (i.e. it is not random). People who do not respond to questions accurately (or at all) cannot be assumed to be similar to those who do. Children for whom a school has no prior attainment data cannot be assumed to be the same as everyone else. A ruler that is too short and so over-estimates heights will tend to do so again and again, uncompensated by any kind of random under-estimates to match it. Even human (operator) error has been shown to be non-random, in such apparently neutral tasks as entering data into a computer. So knowledge of the likely sources of error in any score, and an estimate of the range of measurement errors, is an additional and crucial part of deciding whether a difference between groups is big enough (to justify a substantive claim). The harder it is to measure something, the larger the errors in measurement will tend to be, and so the larger the difference would have to be, to be considered substantial. We cannot specify the minimum size needed for an effect, nor can we use standardised tables of the meanings of effect sizes (Gorard 2006). Those tables showing an effect size of 0.2 as ‘small’ and 0.8 as ‘big’ and so on are a guide only. But we can say with some conviction that, in our present state of knowledge in social science, the harder it is to find the effect the harder it will be to find a use for the knowledge so generated. We need to focus our limited social science funding on developing effects that are big, sustained or have a high benefit:cost ratio.

Models for 'mixing'

The extended discussion of the flaw in statistical testing is just one example of the kinds of supposed barriers we have created to hinder ourselves in the collection and analysis of different types of data. Shorn of error, the logic of analysis using numeric data involves judgement of scale, variability, persistence, accuracy, and so on, laid bare for others to follow. This is the same logic as is used, or should be used, for all data. Similarly, the other purported barriers to treating different data in a similar way are false, but there is insufficient space to view them all here (see Gorard with Taylor 2004). Of course, this does not mean that different kinds of data are not differentially suitable for different tasks. Consider the simple paper by Gorard and See (2011), for example. It uses a large-scale dataset to establish a pattern, and then tries to explain the pattern using in-depth data drawn from a sub-set of the same participants as in the large-scale dataset. Typically, large-scale data (perhaps already existing from official sources) is used to define a problem, pattern, trend or difference. It is also used to select a representative subset of cases for in-depth research to investigate the reasons for the problem, pattern, trend or difference. The in-depth work is, therefore, generalisable in the sense that this term is traditionally used, and different datasets are used to define the pattern and its determinants. This is just one of a range of simple ways in which data of different types can be used in co-operation. Others include design-based approaches (design experiments), Bayesian synthesis (that also allows the inclusion of factors like professional judgement), new political arithmetic, and complex interventions. Again see Gorard with Taylor (2004) for others.

More basically, I wonder what the schism advocates do when synthesising the existing evidence base at the outset of any new project. When reviewing literature, do they just ignore any work not conducted by people clearly within their own camp? It seems so. They do not critique the other work in detail or show why it does not meet some specified inclusion criteria. In fact, there are usually no published inclusion criteria. The reviews, such as they, are usually very partial (meaning both incomplete and heavily biased). Ideally a synthesis is an inclusive review of the literature both published and unpublished, coupled with a re-analysis of relevant existing datasets of all kinds (including data archives and administrative datasets), and related policy/practice documents. It is impossible to conduct a fair appraisal of the existing evidence



on almost any topic in applied social science without drawing upon evidence involving text, numbers, pictures and a variety of other data forms. Anyone who claims to be conducting even the most basic literature review without combining numeric and textual data is surely misguided. For more on this, see Gorard (2013).

Conclusion

I wonder also if schism advocates are happy for potential research users like governments and practitioner bodies to adopt the same approach by recognising evidence of only one kind or another. I suspect not. In fact, in the US when the government mandated the preference for funding randomised controlled trials, most academic research departments complained vociferously. They were right to complain, because a full programme of genuine research requires a wide variety of designs and forms of evidence. However, they were wrong to do so by claiming that 'qualitative' work was in a minority, under threat, and the only work they were prepared to do. This is blatant hypocrisy. In fact, it was probably this kind of schismatic thinking that encouraged the US government to use legislation rather than incentives in the first place.

It is not clear why everything involving numbers is counted as one approach, and everything else including smells, drawings, acting, music and so on is treated as an alternate monolith called 'qualitative'. If researchers do, or should, naturally use whatever methods they need to answer their research questions, then there is no methods schism, and so no separate elements to be 'mixed'. If a researcher really cares about finding something out that is as robust as possible, they should consider ignoring the traditional two-camp research methods resources and behave in research as they would in real life. In real life, the use of mixed methods is natural – so natural, in fact, that we do not generally divide data in the first place. The question to be asked, therefore, is why research should be any different?

At present, the quality of social science research in education is threatened by widespread errors of the kind reported in this paper. Reviews of evidence, and the engineering of findings into usable forms, are often impoverished by adherence to a meaningless tradition of dividing data into the two Q word silos. This is unethical from the perspective of the funders of research, and that of the general public who will be affected by the results of research. There are no real challenges to mixing data of all kinds, except the barriers that we have created for ourselves. But these barriers are insubstantial and will fall simply through us ignoring them. We need therefore to remind existing researchers how they would behave if they wanted to find something out in real-life and actually cared about the results. We also need to prevent new researchers from being taught errors in their increasingly compulsory methods development courses. This is the approach being pursued in my UK ESRC-funded project on design as the basis for analysis (http://www.birmingham.ac.uk/research/activity/education/projects/quantitative-methods-teaching.aspx), of which one of the first products is the book - Gorard, S. (2013) Research Design: Robust approaches for the social sciences, London: Sage.

References

Bradley, W. and Shaefer, K. (1998). Limitations of Measurement in the Social Sciences. Thousand Oaks, CA: Sage.

Creswell, J. and Plano Clark, V. (2007). Designing and conducting mixed methods research. London: Sage.

Gergen, M. and Gergen, K. (2000). 'Qualitative inquiry, tensions and transformations'. In N. Denzin and Y. Lincoln (eds) The Landscape of Qualitative Research: Theories and Issues. Thousand Oaks, CA: Sage.

Gorard, S. (2002). Ethics and equity: pursuing the perspective of non-participants. In Social Research Update, 39, 1-4.

Gorard, S. (2006). Towards a judgement-based statistical analysis. In British Journal of Sociology of Education, 27, 1: 67–80.

Gorard, S. (2010a). Research design, as independent of methods. In Teddlie, C. and Tashakkori, A. (Eds.) Handbook of Mixed Methods. Los Angeles: Sage

Gorard, S. (2010b). Measuring is more than assigning numbers. In Walford, G., Tucker, E. and Viswanathan, M. (Eds.) Sage Handbook of Measurement. Los Angeles: Sage, pp. 389-408.

Gorard, S. (2010c). All evidence is equal: the flaw in statistical reasoning. In Oxford Review of Education, 36, 1, 63-77.

Gorard, S. (2013). Research Design: Robust approaches for the social sciences. London: Sage (forthcoming February).

Gorard, S. and See, BH (2011). How can we enhance enjoyment of secondary school?: the student view. In British Educational Research Journal, 37, 4, pp. 671-690.

Gorard, S. with Taylor, C. (2004). Combining Methods in Educational and Social Research. London: Open University Press.

Kuhn, T. (1970). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Meehl, P. (1998). 'The power of quantitative thinking'. Speech delivered upon receipt of the James McKeen Cattell Fellow award at American Psychological Society, Washington, DC, 23 May.

Symonds, J. and Gorard, S. (2010). 'The death of mixed methods?: or the rebirth of research as craft'. In Evaluation and Research in Education, 23, 2: 121–36.

Lars-Erik Borge, Norwegian University of Science and Technology (NTNU) and Center for Economic Research at NTNU

Comments on Stephen Gorard: Mixed Methods Research in Education

Stephen Gorard discusses the use of mixed methods in research on education. Methods can be mixed in numerous ways, but in his presentation and paper the mixture refers to “qualitative” and “quantitative” methods. It is not clear where to draw the line of demarcation between the two types.

Consider a project where the researcher conducts interviews with a number of respondents, codes the responses on an ordinal scale, and finally conducts a statistical analysis of the responses in combination with quantitative data on e.g. the respondents’ incomes. This project clearly makes use of mixed methods, but can it be exactly divided into qualitative and quantitative parts? In other words, where does the project pass the line of demarcation from using qualitative methods to using mixed methods (in the sense that quantitative methods are brought into the analysis)? When the responses are coded? Or where the responses are combined with quantitative data? As Gorard, I am skeptical to claims that the two Qs are incommensurable and that they need different criteria for judging research quality.

I understand Gorard not first and foremost as an advocate of mixed methods, but rather as a strong critic of researchers that rely on evidence from only one of the two Qs. It is difficult to disagree with his real-life example of purchasing a house. To me this example resembles the design of educational policy, e.g. in the Ministry of Education. It is obvious that educational policy must be based on all available evidence. It would be stupid to dismiss qualitative evidence just because it is qualitative, and to dismiss quantitative evidence just because it is quantitative. But in design of educational policy it is important to dismiss low quality research, irrespective of whether it is qualitative or quantitative, and to let the policy be guided by high quality research. Hopefully policy makers behave like the house buyer in Gorard’s real-life example.

While policy making should be based on all available evidence, it is less clear to me that a mixture of methods is warranted when it comes to individual research projects or individual researchers. Elsewhere Gorard has expressed that researchers who are unwilling to use a mixture of methods “do not care about the results, and are simply pretending to do research (and wasting people’s time and money in the process)”. This statement seems to ignore that there may be gains by specialization and division of labour in research (as in all other industries). It is my experience that large amounts of time and effort are required in order to learn state-of-the-art research methods and to apply them properly, and this is the case for both qualitative and quantitative methods. Specialization should be welcomed in order to avoid the many possible pitfalls and to make sure that the research process produces high quality research that can be trusted. Going back to Gorard’s real-life example: I doubt that the house buyer would rely on a single expert in order to get informed about conditions such as mortgage repayment, the technical condition of the house, traffic safety, local schools, etc. It is probably better to consult individual experts on personal finance, construction, traffic, and schools in order to reach a decision on whether to buy the house or not.

To sum up: Policy-making must be based on all available evidence, both qualitative and quantitative. The total pool of research must therefore include projects using a variety of methods. Moreover, high quality research and the desired variety in the total pool of research are best achieved if individual projects and researchers are allowed to specialize in specific methods.

Jens-Christian Smeby, Centre for the Study of Professions, Oslo and Akershus University College of Applied Sciences

How Can Qualitative and Quantitative Data Sets Be Linked?

All methodological approaches have strengths as well as weaknesses. This is an elementary insight from the literature on methodology. By combining methods in one and the same project it is possible to draw on the strengths of all of the methods used. However, combining different methodological approaches is time consuming and resource intensive. This is why we often advise Master's and Ph.D. students to focus their effort on a single methodological approach.

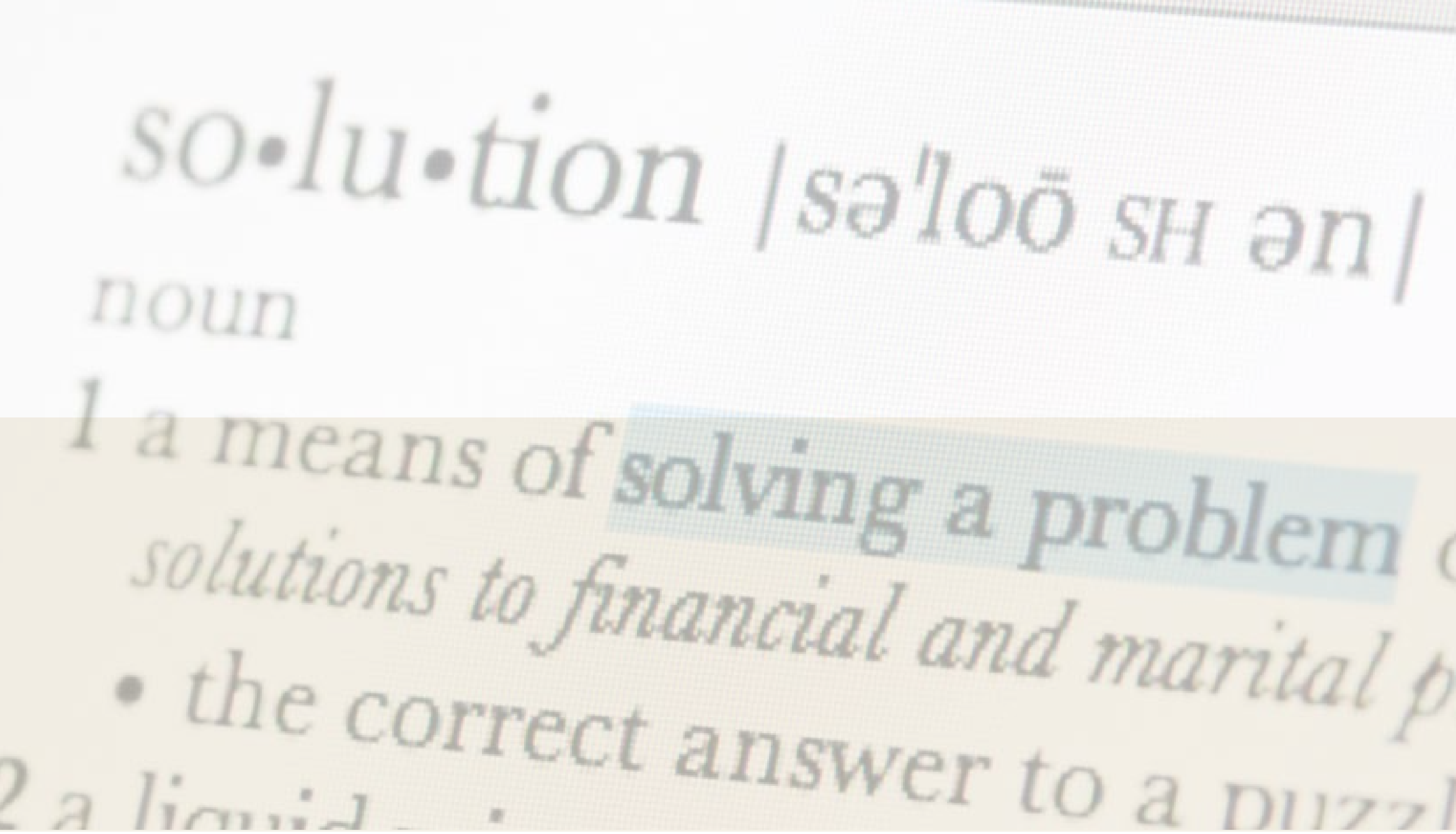
In larger projects involving cooperation among several researchers, the argument against combining methods does not apply. However, there are a number of practical challenges associated with combining methodological approaches and linking together quantitative and qualitative data sets. I will discuss these challenges in light of my own experience as well as give examples of how the use of different methods has produced interesting analyses and results. First, though, I will give a brief explanation of the various types of data and sources and the advantages and disadvantages of combining them.

Various combinations of types of data and sources

In his article on the relationship between qualitative and quantitative approaches in social research,[1] Sigmund Grønmo distinguishes between three main types of data sources in social science research: actors, respondents and documents. These may be studied using both qualitative and quantitative types of data; actors may be observed through participant observation and structured observation, respondents may be asked questions in informal interviews and on questionnaires, and documents may be the object of quantitative and qualitative content analysis. Some combinations of data sources are so common that we hardly ever think about them as methodological triangulation. For instance, it is extremely common to draw on documents regardless of the other types of data sources being used. It is also common to supplement observation of actors with respondent interviews. Qualitative and quantitative data may be combined in various ways. Qualitative studies may be followed up with quantitative studies and qualitative studies may be followed up with quantitative ones. A key reason for combining quantitative and qualitative data sets is that it may test validity of the methods and strengthen confidence in the results. Deviations in results may lead to new interpretations and interpretations may become more cohesive and nuanced.

A main disadvantage, as I alluded to above, is that this process is resource intensive. A stipulation to combine quantitative and qualitative data also limits the choice of research questions because some questions are best explored using only one method. Such demands may also limit the methodologi-

[1] Grønmo, S. (1996). *Forholdet mellom kvalitative og kvantitative tilnærminger i samfunnsforskning*. In: H. Holter og R. Kalleberg (eds.). *Kvalitative metoder i samfunnsforskning* (p. 73-108). Oslo: Universitetsforlaget.


so•lu•tion |səˈloō sн ən|
noun
1 a means of solving a problem
solutions to financial and marital p
• the correct answer to a puzzl
2 a liquid

cal design. Highly advanced statistical analyses and exploratory qualitative field studies may be difficult to combine with other methodological approaches in practice. A requirement that projects must draw on a variety of data may also result in sub-projects that are not well coordinated.

Although this type of methodological triangulation is recommended in the literature on methodology, it may also be difficult to achieve this within the framework of relatively time-limited projects. It takes time when the plan for the second phase of a project is supposed to be based on the results from the first phase. Thus one solution may be to implement the first phase in the form of pre-projects or to follow-up the results from previous projects with new methodological approaches. It is also possible to combine methodological approaches by, for example, quantifying qualitative data as part of the analysis. In the next section I discuss various challenges related to linking quantitative and qualitative data together in more or less parallel paths. This is a common challenge in many larger projects involving cooperation among several researchers.

Project organisation

To achieve good coordination between quantitative and qualitative data sets, it is crucial to incorporate this from the start when planning the project and formulating the project description. Rather than developing sub-projects based on various methodological approaches, it is my experience that sub-projects should be developed based on research questions that can be explored using a variety of methods. This is not always easy because researchers often have interests and areas of expertise related to specific methodological approaches targeted at particular research topics. When the project is then launched, management of the project is crucial for linking the analyses of the various types of data. It is important to emphasise that the project outline is binding for all participants in the project. This also applies to research fellows who usually implement much of the project. Those who apply for research fellowships often have their own research interests that do not completely correspond with the project. If the idea is that the research fellow will be responsible for large portions of the quantitative or qualitative data analyses, it is crucial that the research fellow is bound by the project's research question so that the data sets can be coordinated. There are also coordination problems related to the small number of applicants for this type of fellowship and time-consuming appointment procedures.

An effective way of linking various data sets is to involve the researchers in both the collection and the interpretation of the data across their individual activities. This gives the project partners vital insight into aspects of the data for which they themselves have not primarily been responsible, and this may enable individual researchers to draw on several data sets. A somewhat less demanding way to involve researchers in both processes is through seminars in which drafts of publications are presented and discussed in plenum. To achieve constructive linkage between quantitative and qualitative data sets, however, it is crucial to be somewhat realistic about what is the most effective approach. For one thing, methodological triangulation is not always the best way of investigating research questions. I also want to caution against linkages that look good on paper, but that can be difficult to achieve in practice. For instance, while it may look useful to include the

respondents who were interviewed among those who have answered a questionnaire, this may be difficult to accomplish for various reasons. Nor may it be of any research significance, either.

A long-term perspective

It should be emphasised that projects which achieve a good linkage between qualitative and quantitative data sets are usually the result of research cooperation that has taken place over time. Research groups that are characterised by methodological pluralism, in which the members develop in-depth knowledge of the field and some degree of a shared analytical frame of reference, are often a good basis for developing such projects. Personal contacts and networks are established over time, and projects that incorporate effective cooperation are often developed through collective processes. I also want to stress the importance of experience and knowledge from previous projects. As I noted previously, one way to facilitate linkages between various data sets may be to implement pre-projects. It may also be expedient to develop or draw on established databases so that quantitative data are already available before the project is launched. At the Centre for the Study of Professions we have conducted questionnaire-based longitudinal surveys (Studies of Recruitment and Qualifications in the Professions, "StudData")[2] in which we follow students from the start of their educational programmes until they enter the workforce. A number of other established databases and registry data are also available. The advantage of this type of database is that preliminary data analyses can be used as the basis for developing the project's research questions. These preliminary results may also be used to gain more in-depth knowledge through the collection and analysis of qualitative data.

Two examples

A key question in the research project Professional Learning in a Changing Society (ProLearn) was how recently graduated nurses, teachers, computer engineers and accountants tackled the need for new knowledge in their daily working life.[3] Among other things, we were interested in the degree to which they drew on various types of knowledge resources, such as colleagues, academic articles and the Internet. The project was based on questionnaires, individual interviews, focus group interviews and learning logs. All the interviews were transcribed, and we used a software program to encode and analyse this material. I was responsible for the quantitative analyses, but I also helped to prepare the interview guide and took part in the early phases of the qualitative analysis. Each of us presented a draft paper at the project meetings in which we familiarised ourselves with preliminary results based on various parts of the data. It was at these meetings we became especially aware of the major contradiction indicated in the results from the quantitative and qualitative data. The data from the questionnaires showed that teachers and nurses often sought out colleagues when they had a need for knowledge, but the teachers used considerably more time reading various types of academic literature. The qualitative data, however, indicated that many of the nurses

[2] For more information about the database, see http://www.hioa.no/studdata (in Norwegian only)

[3] For more information and the final report from the project, see: http://www.forskningsradet.no/prognett-utdanning/Artikkel/Professional_learning_in_a_changing_society_ProLearn/1224697828145

were very concerned about staying updated professionally by using various types of written material, while the teachers stated that they had very little time or capacity for precisely this. The question was whether different methods produced different results. However, a thorough comparison of the way in which the questions were formulated on the questionnaire and the interview guide, as well as a new review and analysis of the quantitative data, showed that the results complemented each other in an interesting way. The nurses and teachers read different types of academic literature. The teachers primarily read material that could be used as examples in their teaching, but focused little on literature relating to subject didactics or pedagogy. In contrast, the nurses read academic material directly related to their specific tasks or the patient groups they worked with. Thus the combination of the quantitative and qualitative data helped to verify and reveal important distinctions in the results.[4]

The second example is based on the ongoing project Qualifying for Professional Careers, funded under the Programme on Educational Research towards 2020 (UTDANNING2020).[5] We focus on four groups: school teachers, pre-school teachers, nurses and social workers. The project is based on registry data, various questionnaire-based surveys (including "StudData"), focus group interviews with final-year students, and individual interviews with recent graduates. One part of the project investigates recruitment to the educational programmes, completion of and drop-out from the programmes, and further career trajectories. This part of the project is based primarily on registry data. In the other part we draw on qualitative as well as quantitative data. A postdoctoral research fellow is mainly responsible for collecting and processing the qualitative data, but some of the people working with the quantitative data have also been involved in developing the interview guide and to some extent in conducting and analysing the interviews. We also have regular project meetings and workshops with international participants at which papers are presented. At this time we have no plans to write journals articles using both quantitative and qualitative data, but knowledge about the project results are an important backdrop and basis for the interpretation of much of the data. One of the key questions in the project is what constitutes research-based education and how wide is its scope, and what does this mean for the students in various professional study programmes. Access to various data sets is crucial in this context. We have data from questionnaires answered by teachers of the educational programmes, students who are nearing the conclusion of their studies and graduates who have been working professionally for two to three years. These questionnaires were administered partly before project start-up and partly in the project's early phase. Analyses of the quantitative data show some interesting patterns. For instance, a relatively small percentage of the students is directly involved in the research conducted by their teachers. However, we also find clear differences between the programmes. The interview guide was prepared partly on the basis of the preliminary quantitative results. In addition, the interviews themselves helped to provide depth and nuance to

[4] See Klette, K. & Smeby, J.C. (2012) Professional knowledge and knowledge sources. In: K. Jensen, L. C. Lahn & Nerland, M. (eds.) Professional Learning in the Knowledge Society. Rotterdam: Sense.

[5] Website for the QPC project: http://www.hioa.no/qpc

the quantitative analyses by emphasising that research-based education is an ambiguous term that is interpreted in different ways. As an example, the nursing students have a relatively clear idea of what this is and can give examples from their own studies, whereas the students in teacher education have a much more diffuse idea of what this term means. This is crucial to how the quantitative results are interpreted and presented. We decided to publish the results in book form in part to enable us to link the various data sources; compared to separate articles, this offers a much better opportunity to present the breadth and complexity of – and subtleties inherent in – the data.

Concluding comments

There are good reasons to draw on both quantitative and qualitative data sets, and I have described various ways in which this can be done. I have also pointed out some challenges related to how to achieve this in an effective manner. In many cases, it may be wise to use all the resources within the framework of a project to collect, process and analyse only one type of data. Today this is a completely accepted approach if the study is based on sound quantitative data. However, it should be kept in mind that many of the classical social science studies are based on extensive qualitative field studies. It would be unfortunate indeed if there were no longer room for this type of project.

Ingunn Størksen, Center for Behavioral Research, University of Stavanger

New and Inventive Approaches to Collect Qualitative and Quantitative Data among Young Children

Educational and psychological research on young children is often based on data collected from adults in the child's proximate surroundings, such as teachers or parents. This holds for both quantitative and qualitative research. The adults are seen as reliable informants when it comes to reporting children's behavior and adjustment.

In this article I argue that children could be more involved in qualitative studies in reporting on their own subjective feelings and experiences. After all, the child itself is the only one who has access to its own subjectivity. Furthermore, children could be more involved when we collect quantitative data too, and the data need not merely be based on teacher and parent report. I will give examples of how new and inventive approaches can make this feasible, and in the following I will present several approaches that we have applied by our research group at the Center for Behavioral Research at the University of Stavanger. I will also give some indications of how data from various data collection approaches may be integrated and connected in future research reports.

This article is based on examples from two research projects supported by the Research Council of Norway. The BAMBI project (Norwegian Daycare Centers Approach to Working with Children and Families of Divorce) was supported by PRAKSISFOU, and SKOLEKLAR (Preparing for School in Norwegian Daycare Centers) is supported by UTDANNING2020.

The BAMBI project

Educational and psychological research is often criticized for being more concerned with valid and reliable scales and scientific status than with children themselves (Greene, 2006) and for not taking into full account children as active social agents (Hood, Kelley, & Mayall, 1996). Children's general right and need to express themselves and their views is accentuated in the UN's Convention on the Rights of the Child (1989).

In the BAMBI project we wanted to include the voices of the children, and we therefore conducted a range of qualitative studies not only among daycare staff, parents, and family therapists, but also among the children themselves. The multi-informant design was set up to explore various challenges and solutions related to daycare centres' approaches to children and families experiencing divorce. All informant groups contributed with unique and important insight into this theme, and including the children themselves as informants resulted in some unexpected points that were integrated into the future development of material for daycare centres.

Our research approach when collecting data from the young children (Q methodology with visual images) will be described in more detail below. This approach was actually also adapted and included research results from the BAMBI project that were transformed to a material kit for daycare centres. More information about the research in BAMBI and the material kit "Ett barn – to hjem" or "One child – two homes" can be found at www.uis.no/bambi or in the reference list (Størksen & Skeie, 2012). See also illustrations of the material kit in Figure 1.

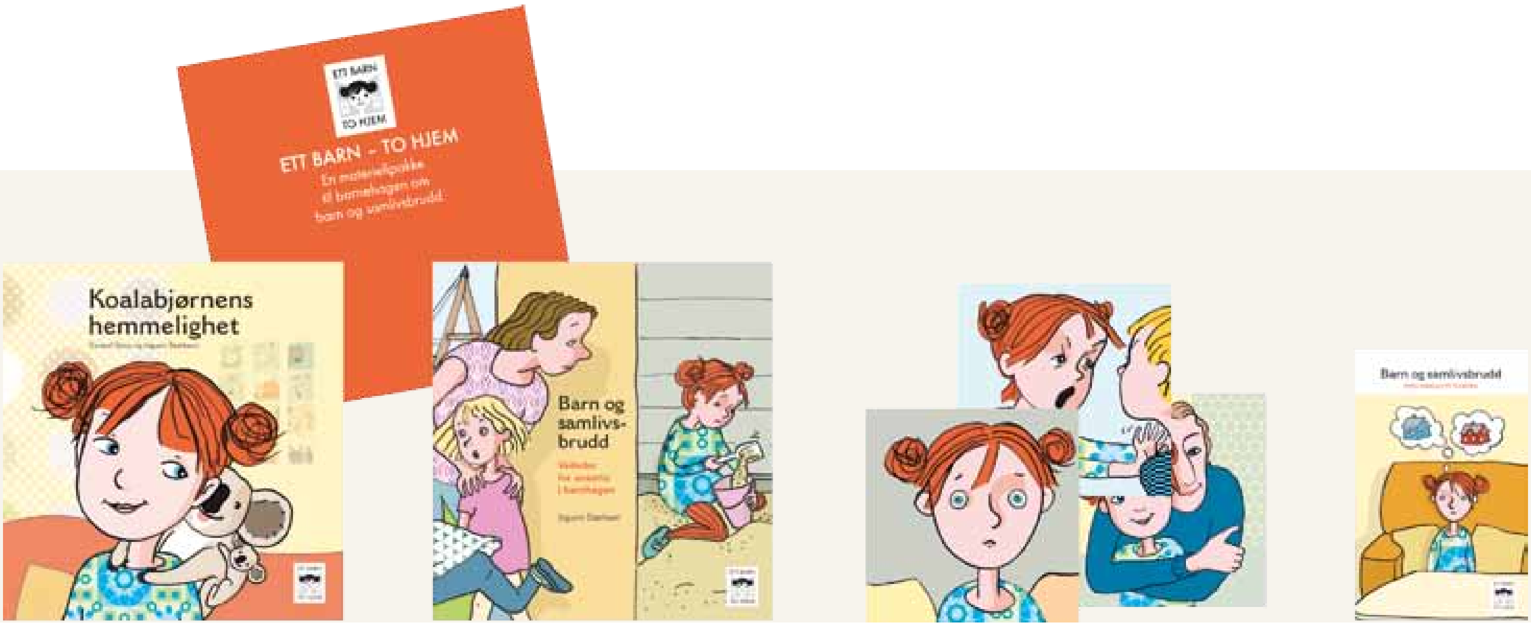


Figure 1. The material kit "Ett barn – to hjem" or "One child – two homes" (Størksen & Skeie, 2012). Pedlex norsk skoleinformasjon.

The SKOLEKLAR project

In SKOLEKLAR our aim is to study possible predictors and inhibitors of learning among young children as they move from daycare centres into the Norwegian school system (at age 6). The main activity in this project is centred on a large quantitative study that follows children from their last year in daycare centre to their first year of school. In addition to this we conduct supplemental qualitative inquiries to strengthen the total research design. The ultimate aim of this project is to detect skills among very young children that should be stimulated in daycare centers in order to facilitate for future adjustment and learning in school. A central hypothesis in this project is that daycare children's socio-emotional skills, such as children's abilities to establish and maintain positive and stimulating relationships with other children and adults and to self-regulate, are very important for future learning and adjustment in school. (There are also several other sub-themes in the SKOLEKLAR project, such as giftedness among young children and challenges related to minority background. For more information see www.uis.no/skoleklar.) In this project we collect data during spring of the last year in daycare, and during spring of the first year of school among approximately 250 Norwegian children. As the present article is being written (the summer of 2012) data from the first assessment has been collected (spring 2012) and data collecting from the second assessment is being planned (spring 2013). The data relate to relationship skills, self-regulation, early academic skills (knowledge of letters and numbers), verbal skills, general cognitive abilities and adjustment, demography and institutional characteristics of daycares and schools.

Q methodology with visual images

In the BAMBI project we were very interested in assessing daycare children's experiences of parents' divorce, and this resulted in a study were young children participated in a Q methodological study with visual images (Størksen, Thorsen, Øverland, & Brown, 2011). Q methodology was originally invented as a systematic approach to be used in the study of human subjectivity (Brown, 1993; Stephenson, 1953; Watts & Stenner, 2012). Still, Q methodology in combination with the use of visual images in a study of young children has never been seen before in Norway, and it is quite rare to see in international research literature too, although such studies have occasionally been conducted (e.g. Taylor & Delprato, 1994). In our study 37 children aged five years participated and almost half of them had experienced parents' divorce. The children were presented with 20 visual cards that illustrated various emotions and experiences that might be related to the divorce. The main contents of the cards could be either positive (e.g. joy or play) or negative (e.g. grief or anger). We took time to go through a carefully prepared routine that was established to make sure the children felt safe and to ensure that they understood the instructions. Generally, ethical questions related to this study were taken very seriously (Thorsen & Størksen, 2010). Subsequently, the children joined us in pointing out cards that they believed were "most like" and "most unlike" their everyday experiences and feelings. The cards where sorted into a predefined grid that indicated where "most like" and "most unlike" cards could be placed. Our experience was that the children managed to express their feelings and experiences in a reliable way through this research approach (Størksen & Thorsen, 2011). Using cards

Photo: Alexandra Halsan, University of Stavanger.

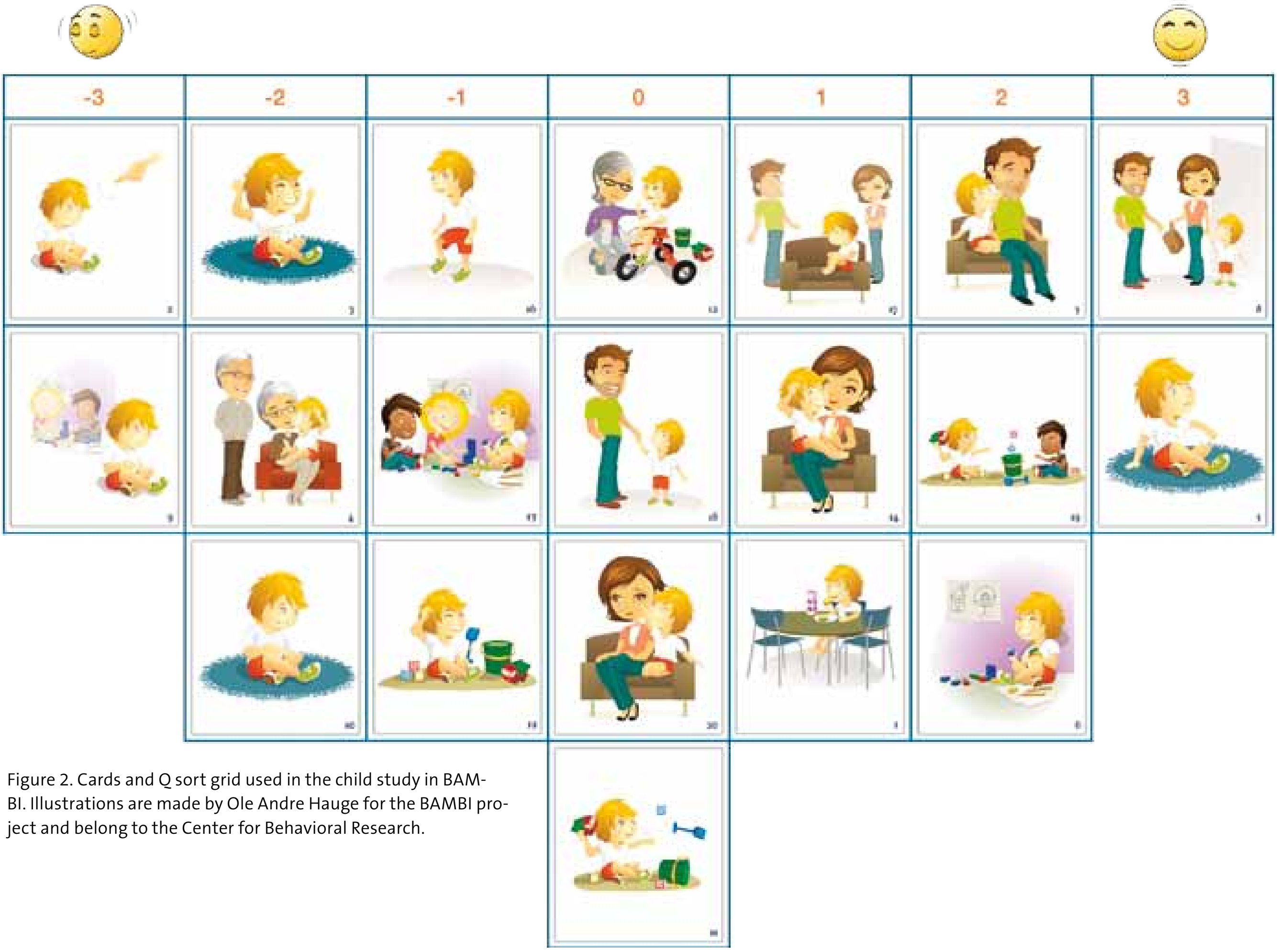


Figure 2. Cards and Q sort grid used in the child study in BAMBI. Illustrations are made by Ole Andre Hauge for the BAMBI project and belong to the Center for Behavioral Research.

instead of questions helped both children that were not verbally strong and children for whom this theme was emotionally challenging to express their experiences. All 37 Q sorts made by the children were analysed with Q methodological principals, and the results clearly enlightened new themes that would not have been detected by merely studying parents and teachers (see Størksen et al., 2011). As mentioned previously, the daycare centre staffs were so impressed by this way of communicating with children on such a sensitive theme, that they asked that a similar set of cards could be included in the material kit that was made for daycare centres in the summary of the project. See illustration of cards and Q sort grid that was applied in our study in Figure 2.

inCLASS observations

In SKOLEKLAR a main aim is to assess children's competences in daily interactions with adults, peers and tasks or learning activities. As mentioned previously, relationship skills are a main theme in this project, and we wanted to study these skills in naturalistic daycare with real-life child interactions. Through close collaboration with our research partners – the inventors of a unique assessment system - we were able to apply the Individualized Classroom Assessment Scoring System (inCLASS). SKOLEKLAR is the first Norwegian research project to utilize this system which has recently been elaborated by researchers at the Center for Advanced Study of Teaching and Learning (CASTL) at the University of Virginia (Downer, Booren, Lima, Luckner, & Pianta, 2010). The system enables systematic observations of children across three domains: interactions with adults, peers, and tasks or learning activities. These domains are divided into 10 more specific dimensions: positive engagement with teachers, teacher communication, teacher conflict, peer sociability, peer communication, peer assertiveness, peer conflict, engagement within tasks, self-reliance, and behaviour control. Each child is observed in natural settings in daycare four times in sequences that last for 10 minutes. The observers are carefully trained through personal reading of the manual, a two-day workshop with a representative from CASTL and post-training examination. This produces and secures a highly reliable observation system (Downer et al., 2010), and indications of reliability and validity have also been proven when applied in the SKOLEKLAR project (Haugerud, 2012). New findings related to children's interactions in Norwegian daycare settings have already been published through two master theses (Haugerud, 2012; Lunde, 2012). For more information on the inCLASS observation system see www.inclassobservation.com .

Assessing cognitive and academic skills through computer tablets

Cognitive and academic skills among young children are often assessed through teachers' reports of observed competencies (Reikerås, Løge, & Knivsberg, 2012). Such assessments rely on a very accurate observation and recollection from the teachers. An alternative way of assessing such skills among children is to let the children themselves participate in tasks and tests that tap these skills. Such tasks and tests are often administered in a traditional pencil-and-paper fashion. Children of today are more and more familiar with digital ways of replying to various cognitive and academic tasks. Therefore, in the SKOLEKLAR project, the tasks that where administered were converted to computer tablet applications in order to ease the process both for the children and for the adult assessors. Furthermore, such an approach eases the future processing of

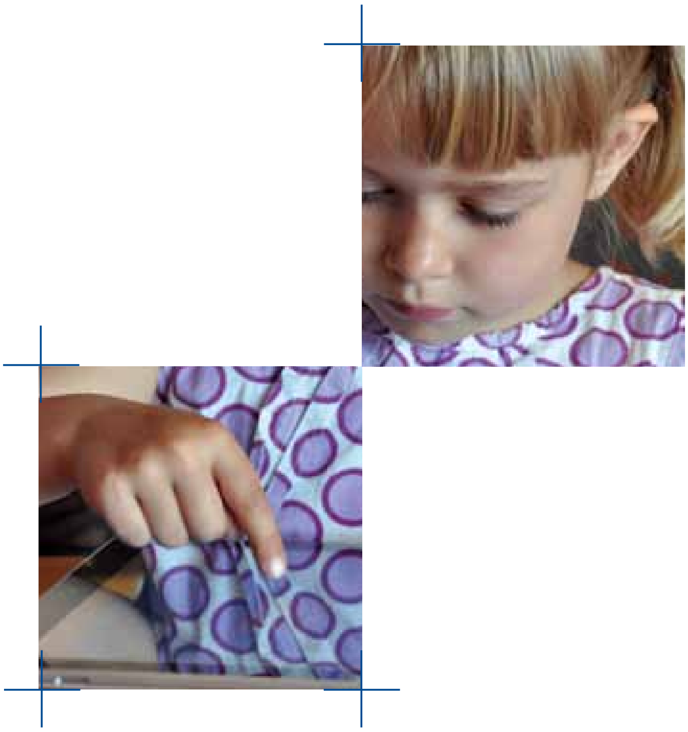


Photo: Alexandra Halsan,
University of Stavanger.

data, since the data are already digitalized. In the SKOLEKLAR project we measured self-regulation (two tasks), memory, vocabulary, early reading and early math skills through computer tablets. This approach seemed to engage the young children, and they were actually able to concentrate and complete all tasks through the 30-60 minutes that this battery required. The adult assessors that collected data from all 250 children were also very happy for this feasible way of collecting data.

Connecting data in SKOLEKLAR

In the SKOLEKLAR project – which is still in progress – we see several possibilities for connecting the rich data material that is being produced. The total data set will eventually result in data from two data assessment time points and will contain questionnaire data from teachers, parents, and daycare/school directors, in addition to inCLASS observations, computer tablet data and supplementary qualitative data. The possibilities for connecting data and exploring relevant research questions seem ample, and we are excited to move forward with analyses and publication of important research results.

Summary

In this article I have given examples of various approaches that may help to involve children more directly in the assessment of qualitative and quantitative research data. Our experience and beliefs is that such data will enrich research, enlighten new themes, produce new results and serve as a supplement to data assessed from adult caregivers. Our ultimate goal is that the research and recommendations that result from the BAMBI and SKOLEKLAR projects will affect Norwegian policies towards even better pedagogical practices in daycare centres and schools.

References

Brown, S. R. (1993). A Primer on Q Methodology. Operant subjectivity, 16(3/4), 91-138.

Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. Early Childhood Research Quarterly, 25, 1-16.

Haugerud, E. (2012). De psykometriske egenskapene til observasjonsverktøyet inCLASS - når det anvendes i en norsk setting / The psychometric properties of the inCLASS observation system when it is applied in a Norwegian setting. (Master thesis), The University of Stavanger.

Hood, S., Kelley, P., & Mayall, B. (1996). Children as Research Subjects: a Risky Enterprise. Children and Society, 10, 117-128.

Lunde, S. (2012). Positivt engasjement og kommunikasjon med voksne i barnehagen og små barns evne til selvregulering / Positive engagement and communication with teachers in Norwegian daycare centers and young children's self-regulatory skills. (Master thesis), The University of Stavanger.

Reikerås, E., Løge, I. K., & Knivsberg, A. M. (2012). The Mathematicla Competencies of Toddlers Expressed in Their Play and Daily Life Activities in Norwegian Kindergartens. International Journal of Early Childhood 44(1), 91-114. doi: 10.1007/s13158-011-0050-x

Stephenson, W. (1953). The study of behavior: Q-technique and its methodology. Chicago: University of Chicago Press.

Størksen, I., & Skeie, E. (2012). Ett barn - to hjem / One child - two homes. Veilederpakke for barnehager / material kit for daycare centers. Oslo: Pedlex - Norsk skoleinformasjon.

Størksen, I., & Thorsen, A. A. (2011). Young children's participation in a Q study with visual images. Some comments on reliability and validity. Operant Subjectivity: The International Journal of Q-methodology, 43(3), 146-171.

Størksen, I., Thorsen, A. A., Øverland, K., & Brown, S. R. (2011). Experiences of daycare children of divorce. Early Child Development and Care, DOI:10.1080/03004430.2011.585238

Taylor, P., & Delprato, D. J. (1994). Q-methodology in the study of child phenomenology. Psychological Record, 44(2), 171-184.

Thorsen, A. A., & Størksen, I. (2010). Ethical, methodological, and practical reflections when using Q methodology in research with young children. Operant Subjectivity: The International Journal of Q Methodolog, 33(1/2), 3-25.

United Nations. (1989). Convention on the Rights of the Child.

Watts, S., & Stenner, P. (2012). Doing Q Methodological Research. Theory, Method and Interpretation. Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE.

Marianne Ødegaard, The Norwegian Centre for Science Education[1]

Studying Science Classrooms – Linking Complex Data

Studying science classroom practice is a complex endeavour. There are many perspectives to have in mind. Science education in itself is multi-faceted. School science is a combination of several sciences: life sciences, earth sciences and physical sciences. In order to understand science, and become scientifically literate, students have to engage in a great variety of activities.

Designing for complexity

Sjøberg (2009) describes science education as having three balanced dimensions: products of science, processes of science and science in society. He argues that in order for students to obtain scientific literacy they have to appropriate more than the traditional scientific facts. They have to understand the processes through which scientific knowledge is produced. Other science education researchers expand this view and agree that students who are proficient in science should i) know, use, and interpret scientific explanations of the natural world; ii) generate and evaluate scientific evidence and explanations; iii) understand the nature and development of scientific knowledge; and iv) participate productively in scientific practices and discourse (Duschl et al, 2007). In our national science curriculum (KL06) several of these points are managed through a combination of the main subject area the Budding Scientist ("Forskerspiren") and basic skills ("grunnleggende ferdigheter"). The intention is that the Budding Scientist and basic skills should permeate learning activities in all the main subject areas in the science curriculum. Thus, it is possible for the students to learn all science content through a combination of inquiry-based science and literacy. In order to study science classroom, this means that a great variety of activities could be interesting to capture: practical activities, planning practical activities, discussing practical activities, field work, talking about theory, reading about theory, writing about the connections between theory and practice, mediating scientific knowledge orally or in a digital manner and so on. This is possible by linking a variety of multi-modal data.

Another part of the complexity of science classroom practice is that it includes both teaching and learning processes. For students this includes both affordances and experiences. In order to study student learning it might be interesting to look at the relationships between what the teacher offers of activities to the students, how students engage with them, and how this is understood by the students. This involves both observational data, like video observations, product data, meaning what students produce during the learning activity, and reflective data, like interviews or reflection notes. Another facet of the teaching learning process is the teachers' own reflective processes. How do they read the students in order to get hold of what they do and do not understand? How does this influence the teaching activities? Do the teach-

[1] This article is based on the work done by the Budding Science and Literacy research group that in addition to the author consist of Sonja M. Mork, Merethe Frøyland, Gard Ove Sørvik, Berit S. Haug and Kari Beate Remmen.

ers stick to their plan? Or do they adjust them together with giving feedback, or as a kind of feedback to students? In what ways do teachers offer formative assessment, and how do they know what to offer? Again, these research perspectives demand observational, reflective and product data.

Design-based research – maintaining complexity

The Budding Science and Literacy project (Ødegaard, Frøyland and Mork, 2009) is an example of a design-based research project (Sandoval and Bell, 2004) where a complex multimodal teaching model (Ødegaard and Frøyland, 2010) is developed and continuously improved through iterative research cycles. Several teachers are involved in trying out and adjusting the teaching model (intervention) to their students and teaching environment. In addition, a number of researchers are engaged in the study doing research with different approaches such as classroom activities: studying the range of multi-modality of the learning activities and how the inquiry processes interact with the literacy activities; student learning: studying students' conceptual learning and how students use and produce scientific texts in different genres; and teaching strategies: studying how teachers give formative assessment and how they initiate reading processes. These research approaches seek to contribute to the improvement of the teaching model in various ways depending on their research focus. Because of this combined design the project attempts to have a robust data collection consisting of data material from different sources.

The Budding Science and Literacy data material consists of:

- *Observational data:* video observations of whole classes, video observations of teacher movements and talk, videos of students' observations from head-mounted cameras (head cam), videos of teachers' observations from head-mounted cameras during outdoor field work and GPS-data of student movement during field work
- *Reflective data:* interviews of teachers (pre- and post-intervention), interviews with students (post-intervention), teachers' reflective notes (pre- and post-intervention), teacher questionnaires (pre- and post-intervention)
- *Product data:* students' written texts, students' drawings and students' field-notes
- *Reference data:* Seeds of Science Roots of Reading teacher guides

This range of data sources provides potential of many different connections of data. Some possibilities will be mentioned, but only a few ways of linking data will be described.

Analyzing complex data

Various data have different levels of complexity, video data being the most complex. The head-mounted cameras also introduce an extra dimension of complexity, where students' and teachers' direction of attention is included. It is possible to reduce complexity by transcribing the video files, but then you lose the opportunity to analyze the multi-modal reality that the students and teachers experience. This is especially important in science where exploration and inquiry skills are required, particularly in practical activities. In addition, the head cam videos capture the teachers' and students' perspective in a physical manner. What do the students look at when they are on a field trip? Which impressions motivate their talk? How do they develop and explore their own texts? These questions are more likely to be answered by analyzing head cam videos. The Budding Science and Literacy project has therefore chosen not to transcribe and thereby reduce the video observations, but rather to analyze the moving pictures using the video analysis software InterAct (Mangold, 2010).

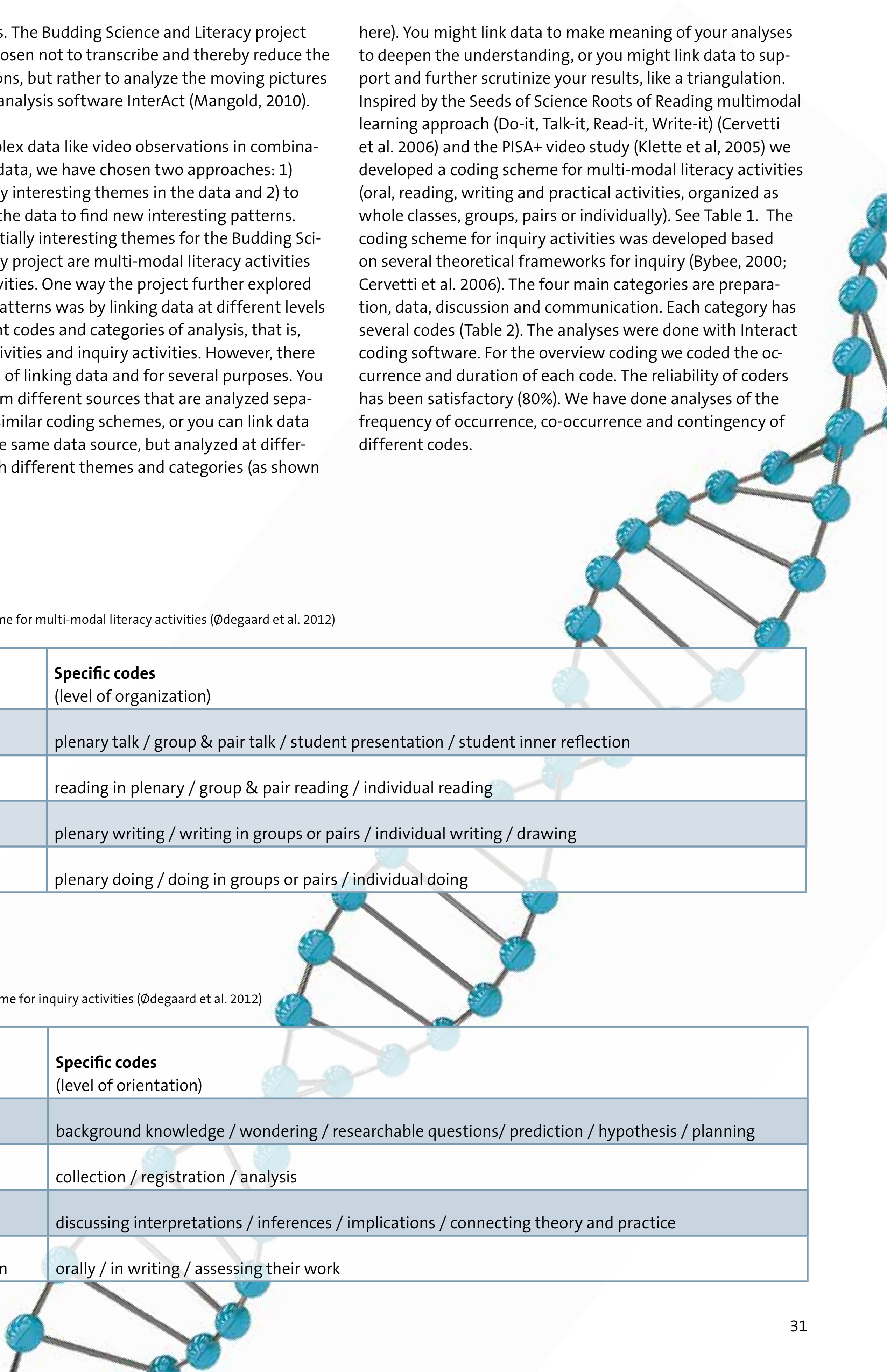
To analyze complex data like video observations in combination with other data, we have chosen two approaches: 1) to pursue initially interesting themes in the data and 2) to further explore the data to find new interesting patterns. Some central initially interesting themes for the Budding Science and Literacy project are multi-modal literacy activities and inquiry activities. One way the project further explored for interesting patterns was by linking data at different levels through different codes and categories of analysis, that is, multi-modal activities and inquiry activities. However, there are several ways of linking data and for several purposes. You can link data from different sources that are analyzed separately but with similar coding schemes, or you can link data that are from the same data source, but analyzed at different levels or with different themes and categories (as shown here). You might link data to make meaning of your analyses to deepen the understanding, or you might link data to support and further scrutinize your results, like a triangulation. Inspired by the Seeds of Science Roots of Reading multimodal learning approach (Do-it, Talk-it, Read-it, Write-it) (Cervetti et al. 2006) and the PISA+ video study (Klette et al, 2005) we developed a coding scheme for multi-modal literacy activities (oral, reading, writing and practical activities, organized as whole classes, groups, pairs or individually). See Table 1. The coding scheme for inquiry activities was developed based on several theoretical frameworks for inquiry (Bybee, 2000; Cervetti et al. 2006). The four main categories are preparation, data, discussion and communication. Each category has several codes (Table 2). The analyses were done with Interact coding software. For the overview coding we coded the occurrence and duration of each code. The reliability of coders has been satisfactory (80%). We have done analyses of the frequency of occurrence, co-occurrence and contingency of different codes.

Table 1. Coding scheme for multi-modal literacy activities (Ødegaard et al. 2012)

Category (activity)	**Specific codes** (level of organization)
Oral	plenary talk / group & pair talk / student presentation / student inner reflection
Reading	reading in plenary / group & pair reading / individual reading
Writing	plenary writing / writing in groups or pairs / individual writing / drawing
Practical doing	plenary doing / doing in groups or pairs / individual doing

Table 2. Coding scheme for inquiry activities (Ødegaard et al. 2012)

Category (activity)	**Specific codes** (level of orientation)
Preparation	background knowledge / wondering / researchable questions/ prediction / hypothesis / planning
Data	collection / registration / analysis
Discussion	discussing interpretations / inferences / implications / connecting theory and practice
Communication	orally / in writing / assessing their work



Linking video analyses on different levels

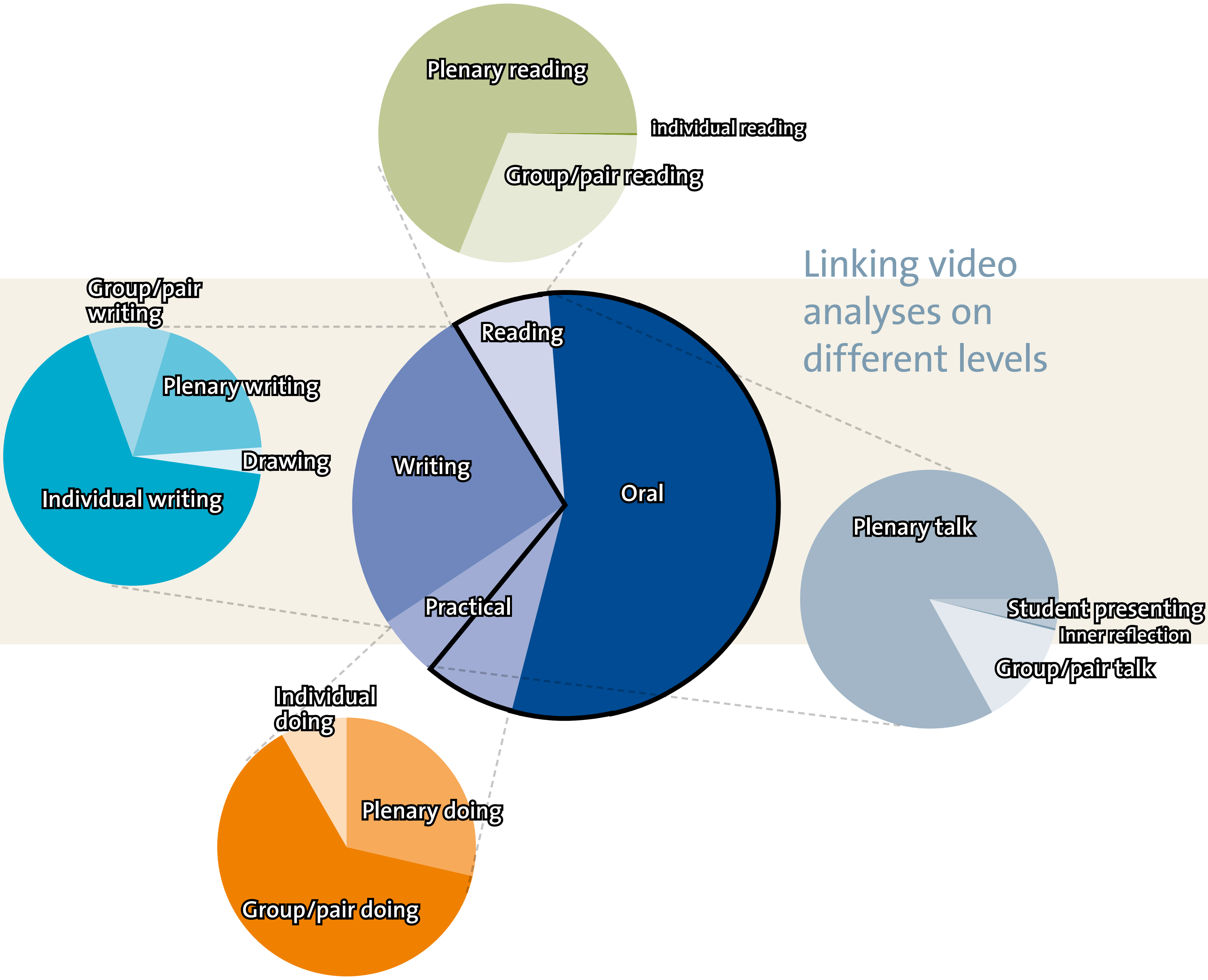
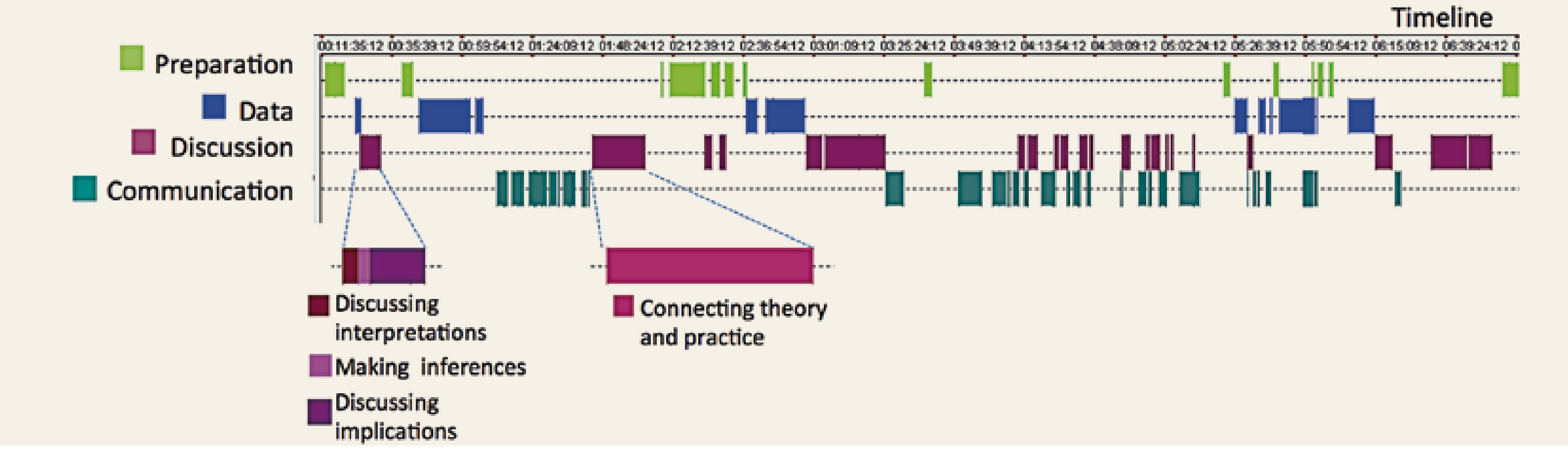


Figure 1. An example of linking different levels of analyses of the same data material. Linking activities with classroom organization.

Examples of linking data

When analyzing the classroom video observations in the Budding Science and Literacy project, the analyses show, as expected, variation in the occurrence of literacy activities. Summing up all videotaped and analyzed lessons, oral activity is the most frequent modality and occurs together with the other modalities. See Figure 1 (inner circle). Writing is the second most frequent before doing practical activities and reading, thus it may look like the science classroom is totally dominated by oral activity. However, when the analyses are linked to the next level of analysis (the organizing codes) additional information may allow a different picture of the situation to emerge. Additional analyzed data may provide contrasting or supplementary knowledge, according to Ødegaard and Klette (2012). Because the teachers often model the multimodal activities for the students in plenary, like reading and writing, the oral code dominates. But we see that in reality students are exposed to a variety of literacy activities (see Figure 1, the whole picture). Most of the reading is done in plenary sessions or in groups. Very little individual reading occurs in our material. Writing is a more individual activity. Practical activities are done mostly in groups. And as expected oral activities are most frequent in plenary sessions, but there are organized group discussions, which are often reported as lacking in Norwegian science classrooms (Ødegaard and Arnesen, 2009). These frequencies are not surprising. When we link these analyses with the analyses using the same code scheme, but used on different, but connected data – the teacher guide – we see that they largely agree with the activities recommended in the teacher guide in the Seeds and Roots material. However, scrutinizing each teacher, individual variations are identified, indicating how the teachers adapt the curriculum to the Norwegian science content and their personal teaching style.

Figure 2. Example of coding one teacher for inquiry activities. In addition, we see some examples of coding of the level of orientation. This gives a picture of a sequence of 7 hours of science lessons about body systems. The students are in 4th grade (9-10 years old).



Analyses of inquiry features show normal progression with preparation activities first, work with data and often alternation between discussion and communication. See Figure 2. Interestingly, we occasionally see that the pattern is interrupted, as when for instance a discussion is started in the middle of working with data. This indicates that students and teacher make small inquiries during the pace of the overarching one, as they take on and get used to an inquiry working manner. When you link the analyses to the level of orientation of the activities, again the picture is broadened. The analyses indicate how the students discuss different interpretations of the data, how students make inferences based on data, how they discuss implications of their findings and how the teacher helps the students connect theory and practice (see Figure 2). Thus the discussion phase shows potential for valuable learning moments.

When linking data from the multimodal activities and the inquiry activities, analyses of co-occurrence and contingency show that the data phase of inquiry is of particular importance for the dynamics of other classroom activities. We see that data is collected and handled using the whole range of literacy activities. Data might be collected when doing practical activities, but also when reading or writing. In addition, the data phase often initiates a change in the variation of classroom activities. It might initiate a reading activity because the students seek more information or a writing activity because they wish to communicate their findings. We were not able to see this pattern before we linked the analyses of the two types of activity.

Conclusion

Linking complex data can be done in various ways. The Budding Science and Literacy research group has worked with several approaches linking different data sources, such as linking headcam videos and student products (Remmen and Frøyland, 2011); linking teacher interview data, reflective notes and video observations (Haug, 2011); linking student products and interviews; and linking survey data with video observations (Mork, Erlien and Ødegaard, 2011).
However, in this article the examples are mainly of linking data of the same data source but analyzed in different manners. The mix of analyzing frequencies of different types of activities at diverse levels, and linking them in order to explore for co-occurrence and contingency illuminates parts of a complex picture that is necessary when doing design-based research. The coding schemes help us to reduce the complexity, so we are able to analyze what happens in the classroom. Linking analyzed data from the same or different sources helps us to see patterns that are otherwise hidden in the complexity.

References

Bybee, R. W. (2000). Teaching Science as Inquiry. In J. Minstrell & E. Van Zee (Eds.), Inquiring into Inquiry Learning and Teaching in Science (pp. 20–46). Washington, D.C.: American Association for the Advancement of Science, Washington, DC.

Cervetti, G., Pearson, P. D., Bravo, M. A., & Barber, J. (2006). Reading and Writing in the Service of Inquiry-based Science. In R. Douglas, et.al.(Eds.), Linking Science and Literacy (pp. 221–244). VA: NSTA Press.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). Taking Science to School: Learning and Teaching Science in Grades K-8. Washington D.C.: The National Academies Press.

Haug, B.S. (2011) Formative Assessment of Students' Understanding of Scientific Concepts. Conference paper presented at the ESERA International conference, August 2011 in Lyon.

Klette, K., Lie, S., Anmarkrud, Ø., Arnesen, N., Bergem, O.K., Ødegaard, M. et al., (2005). Categories for Video Analysis of Classroom Activities with a Focus on the Teacher. Oslo: University of Oslo.

Mangold (2010). INTERACT Quick Start Manual V2.4. Mangold International GmbH (Ed.) www.mangold-international.com

Remmen, K.B. & Frøyland, M. (2011). Patterns in Teacher-Student Interaction and Discourse in Two Settings: The Classroom and the Field. Conference paper presented at the ESERA International conference August 2011 in Lyon

Sandoval, W.A. & Bell, P. (2004). Design-Based Research Methods for Studying Learning in Context: Introduction. Educational Psychologist 39(4), pp. 199-201

Sjøberg, S. (2009). Naturfag som allmenndannelse en kritisk fagdidaktikk (3. utg. ed.). Oslo: Gyldendal akademisk.

Sørvik, G.O. (2011). Yes, They Can! 6-year-old Science Students Understanding and Using the Concept of Systems. Conference paper presented at the ESERA International conference, August 2011 in Lyon.

Ødegaard, M. and Arnesen, N. (2010) Hva skjer i naturfagklasserommet? Resultater fra en videobasert klasseromsstudie. PISA+. NorDiNa 6 (1) pp.16-32.

Ødegaard, M., Frøyland, M. og Mork, S.M. (2009) Budding Science and Literacy: A Longitudinal Study of Using Inquiry-based Science and Literacy in Comprehensive Schooling. Project description. Education2020 – Norwegian Research Council.

Ødegaard, M., & Frøyland, M. (2010). Undersøkende naturfag ute og inne. Forskerføtter og leserøtter. Kimen, 1/2010.

Ødegaard, M. & Klette K. (2012). Teaching Activities and Language Use in Science Classrooms: Categories and Levels of Analysis as Tools for Interpretation. In D. Jorde and J. Dillon (Eds.): The World of Science Education Handbook – Europe. Rotterdam: Sense Publishers.

Ødegaard, M., Haug, B.S., Mork, S.M. & Sørvik, G.O. (2012). Categories for Video Analysis of Multimodal and Inquiry Activities in Science Classrooms. Oslo: Norwegian Centre of Science Education (www. naturfagsenteret.no).

020

Torberg Falch, Department of Economics, Norwegian University of Science and Technology (NTNU) and Centre for Economic Research at NTNU

Pupils and Schools – Analyses of Data at Different Levels

Learning is a product of individual traits and the learning environment. The learning environment includes peers and teachers, and it is influenced by the facilitation of school principals and school authorities. Data available to researchers typically include information at several of these different levels. Data might include information on individual pupils, classroom and school characteristics, neighbourhood features, governance structure of school authorities (municipalities), or even country level information that can be explored in comparative studies. The data have a multilevel feature.

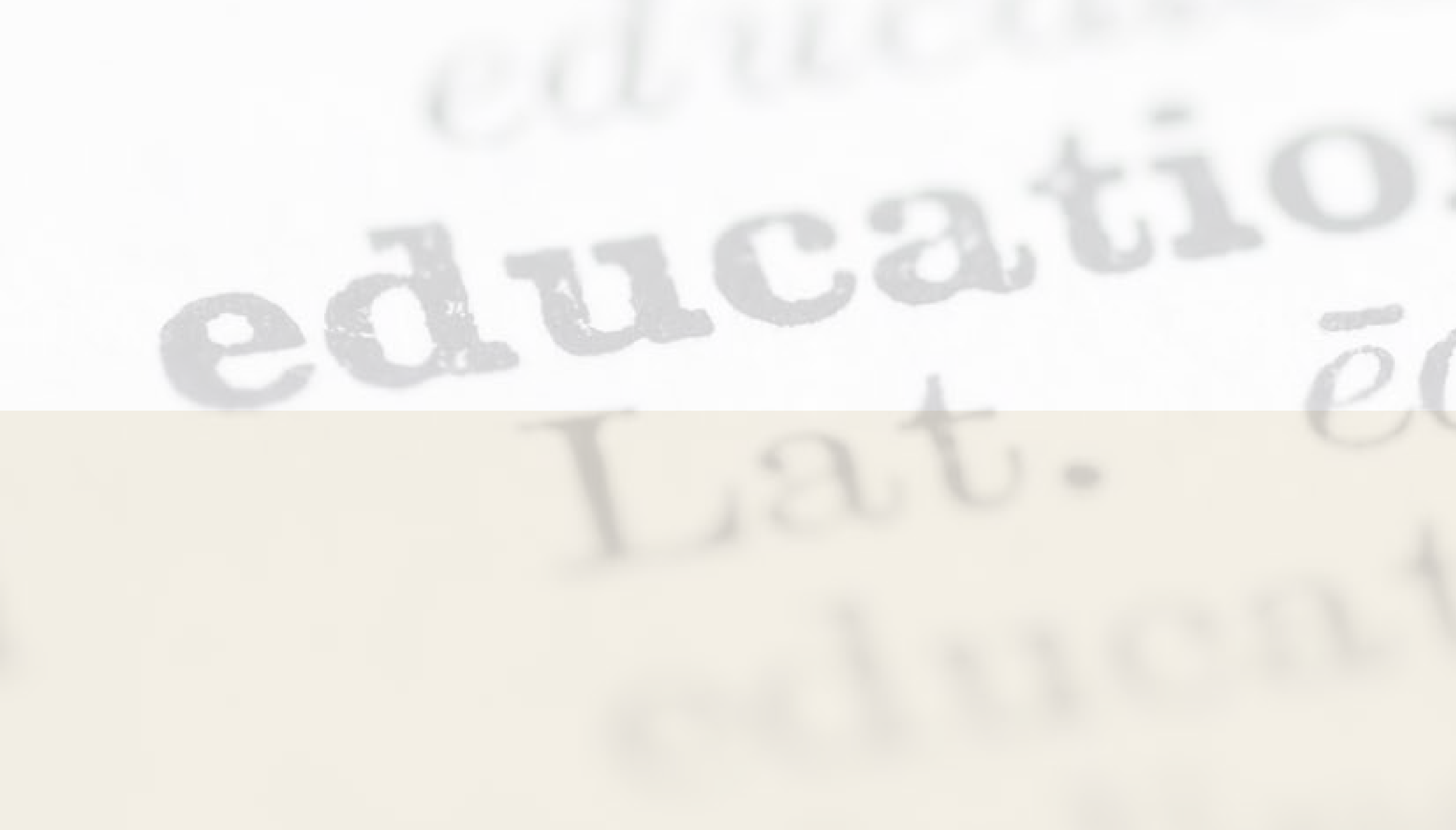
The ultimate goal for many empirical studies is to estimate the average effect of a policy on outcomes. There are multiple policies that need evaluation, for example increasing the teacher-student ratio, teacher assignment of homework, the location of schools, etc. In estimating policy effects, however, the estimates might be biased by unobserved factors that are (i) important for the outcome of interest and (ii) correlated with the policy. Only when the policy is implemented as a fully randomised intervention is the estimation of a causal effect straightforward. In all other cases the potential impact of unknown confounding factors must be considered.

Since random interventions are rare in education, analyses typically have to rely on statistical methods for inference. All non-experimental studies are plagued by potentially unobserved factors. Information on objectively measurable variables might be lacking in the data, and other important factors such as "culture for learning" are extremely difficult to measure.

Simply because the "correct" model is unknown, researchers cannot prove that they have estimated it. All estimates can in principle be biased estimates of the causal effect, and the size of the potential bias is at the outset unknown. Data rich on information are helpful, but however rich on information the data at hand are, empirical modelling requires careful judgments that build on the empirical and theoretical literature. I will argue that it is important to perform robustness analyses to investigate whether estimated relationships are robust to changes in model specification. Robust relationships are more trustworthy. On the other hand, when relationships turn out not to be robust, the researcher can learn about important features of the data, and, in the next round, be able to improve the empirical model. In the research process, multilevel features of the data might be highly valuable and should influence the modelling strategy. [1]

Next I discuss how the "relevant" data level relates to the research question and available data. I argue that it is useful

[1] An increasing literature exploits "quasi-natural" experiments of different kinds to estimate causal effects. Since such "experiments" are not based on fully randomised interventions, they also require careful judgements in the empirical modelling and robustness analyses (Angrist and Pischke, 2009).



to reduce the number of data levels for which there might be unobserved factors. Before the conclusion I also provide an example of how potential unobserved factors at different data levels can be handled.

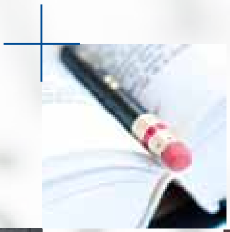
Research question, data level and empirical methodology
Important for the modelling strategy is at which data level the variation of interest occurs. It is this information that has to be exploited for identification of the effect of interest. To estimate effects of school level policies one needs variation across school observations, while information at a higher aggregation level is of less importance. The effect of school level policies might be estimated on data for a single (large) municipality. Consequently, when using data on several municipalities, the variation across municipalities is not necessary for the identification. To take one example, travel distance to school varies across pupils, and to estimate the effect of travel distance (as exemplified below) one does not need to rely on variation in travel distance across municipalities.

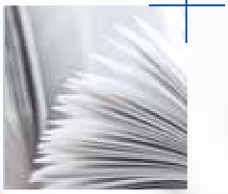
The main advantage of not relying on variation at a higher data level is that one avoids potential confounding effects from this level to bias the estimated effect of the variable of interest. Technically, the model is said to include fixed effects. These fixed effects capture all factors at the higher data level that influence the average student. There will be no omitted variables at this level in the analysis, only potentially unobserved factors at the data level with the variation of interest for the identification.

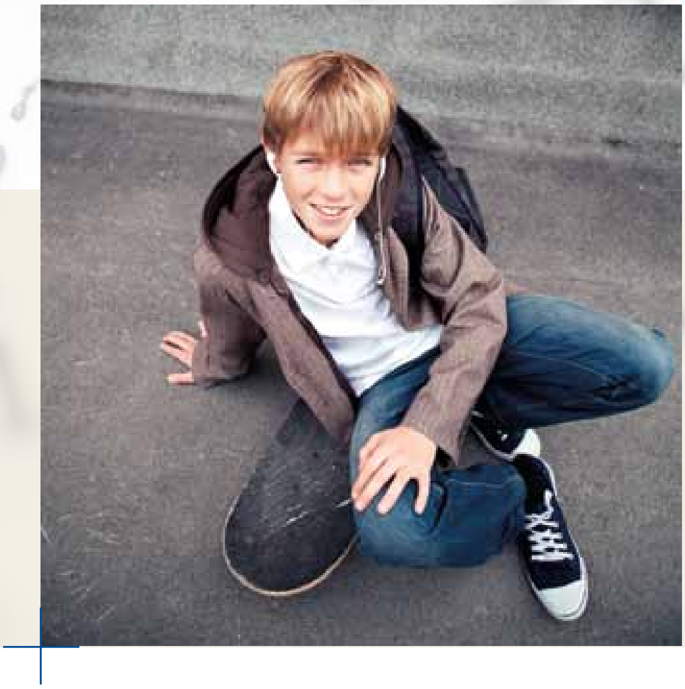
Often empirical work has several research questions related to different data levels. In this case the empirical modelling must take the multilevel features of the data into account in a different way. Since fixed effects capture all the variations at the higher data level it is not possible to identify the effect of any variable at this level. In order to identify the effect of variables at different data levels, the empirical model must allow for unexplained variation at several data levels.

One popular approach is so-called multilevel models, also denoted hierarchical models or nested models (Goldstein, 2011). There are two main features of this approach. First, statistical tests related to variables at the higher level take into account that the data are clustered, that is, there are fewer independent observations at the higher level than at the lower level. Second, and most important, some imposed structure is assumed on the variation across higher level units. These so-called random effects are most often assumed to be normally distributed across the higher level units. Some of the variation at the higher level is implicitly controlled for in the model.

This approach does not address the modelling challenge related to unobserved factors. Random effects have to be assumed random in the sense that they are not systematically related to any other factor in the empirical model. Unobserved factors at the higher data level might bias the effect of variables of interest at the lower data level by this modelling approach, a feature that this approach shares with models that do not take the multilevel structure of the data into account at all. Since the random effects are assumed to be random, they are designed to improve only on the statistical properties on the model, not to handle potentially unobserved factors (Wooldridge, 2009 Ch. 14).







This reasoning makes it attractive to focus on research questions at one data level at time. When the variable(s) of interest varies across pupils, variation at the school level might be controlled for by school fixed effects. When the variable(s) of interest is measured at the school level, variation at the municipal level can be controlled for by municipal fixed effects. When the former type of analyses is separated from the latter type of analyses one avoids that omitted variables at the school level bias the effect(s) of interest at the pupil level. The challenge that remains, of course, is related to potential unobserved factors at the data level of interest. This is always highly important for the interpretation of non-experimental empirical results.

An empirical example

Falch, Lujala and Strøm (2012) analyse the relationship between distance from parental residence to upper secondary schools and the probability of graduating upper secondary education on-time. Long travelling distances are likely to increase the costs of schooling, making it harder to graduate on-time.[2]

At which upper secondary school the pupils enrol is partly an individual choice. In particular, enthusiastic students might choose to enrol at schools some distance away from their home. As an objective measure of geographical constraints the paper focuses on minimum travel time as measured by driving time by car at speed limits from the residence at the end of compulsory education to the closest upper secondary school. For the mid-point in each ward ("grunnkrets") the travel time is calculated by use of ArcGIS Network Analyst.

The data include all pupils turning 16 years of age in 2002 and finishing compulsory education in 2002. The data is nested in the sense that pupils live in wards, wards belong to municipalities, and municipalities are located in regions. In the data there are 10,857 wards, 433 municipalities and 90 economic regions defined by Statistics Norway. The variation of interest is at the ward level, since travel time is only observed for the midpoint in each ward. The data is also clustered at the upper secondary school in which the pupils enrol, but they are not nested at this level because students from the same ward to some extent enrol at different schools.

Here I present the identification in Falch, Lujala and Strøm (2012) based on the fixed effects approach. This approach conditions on all variation at a higher data level than the ward, and thus identifies the effect of interest by exploiting variation within the higher data level.

Some empirical results are provided in Table 1. The first column shows that the simple correlation is negative. The point estimate implies that increased travel time to nearest upper secondary school by one hour is associated with 8.7 percentage points lower probability of graduating on-time. This relationship may of course be driven by e.g. regional differences in parental education or labour market conditions. Column (2) takes such factors into account by including a set of vari-

[2] Expected time to graduation varies across study tracks in upper secondary education; from 3 years in academic tracks and up to 4.5 years in some specialisations in vocational tracks. When calculating graduation on-time we have considered the expected time to graduation at the study track in which the students enrolled the first year in upper secondary education.

ables measuring socioeconomic status (SES) and region fixed effects. In this specification it is only variation in travel time within economic regions and for given SES that contributes to the identification. The effect of travel time drops to 4.6 percentage points, which indicates that the raw correlation in column (1) is inflated by omitted variables.

Column (3) replaces the region fixed effects by municipality fixed effects. All the variation across regions is absorbed by the municipality fixed effects because the municipality is at a lower data level than the regions. Column (4) also includes fixed effects for the upper secondary schools in which the students enrol. In this specification, only variation in travel time to nearest school across wards in the same municipality and at the same upper secondary school contributes to the identification of the effect of travel time. Thus, factors such as geographical conditions at the municipal level and school quality at the upper secondary level are fully controlled for. Table 1 shows that the estimated effect of travel time is not particularly sensitive to the inclusion of fixed effects at these data levels.

Table 1. The effect of travel time on on-time graduation from upper secondary education

	(1)	(2)	(3)	(4)
Travel time to nearest upper secondary school, measured in hours	-0.087* (-4.48)	-0.046* (-3.65)	-0.045* (-2.54)	-0.037* (-2.04)
Region fixed effects	No	Yes	-	-
Individual characteristics	No	Yes	Yes	Yes
Municipality fixed effects	No	No	Yes	Yes
Upper secondary school fixed effects	No	No	No	Yes
Observations	51,484	51,484	51,484	51,484
R-squared	0.002	0.324	0.331	0.354

Note. Source is Falch, Lujala and Strøm (2012). Individual characteristics included are GPA from compulsory education, gender, immigration status, birth month, mobility between ages 6-16, benefits due to disease or disability before age 18, parental education, parental income, and parental civil status. The models with individual characteristics also include ward information; average GPA, share of parents with at least upper secondary education, share of immigrant pupils, ward size, and rural vs. urban ward. t-values presented in parentheses are based on standard errors clustered at the regional level. * denotes statistical significance at 5 % level.

Even though the results are robust to the inclusion of fixed effects, this is not a proof that the model is not plagued by an omitted variable bias related to factors at the ward level. Falch, Lujala and Strøm (2012) also use other methods to further investigate the robustness of the finding, which indicates that the results are not driven by omitted variables at the ward level.

Conclusion

The interpretation of non-experimental empirical results must always be related to potential unobserved factors. I have argued that multilevel features in data make it possible to control for some unobserved factors in the empirical modelling. This does not, of course, guarantee that estimates are unbiased policy effects. However, it clarifies at which data level the relevant unobserved factors must be at play if the estimates should be biased.

References

Angrist, J. D., and J.-S. Pischke (2009). Mostly Harmless Econometrics. Princeton University Press.

Falch, T., P. Lujala and B. Strøm (2012). “Geographical constraints and educational attainment”. In Regional Science and Urban Economics, forthcoming.

Goldstein, H. (2011). Multilevel Statistical Models. John Wiley & Sons Ltd., 4th edition.

Wooldridge, J. M. (2009). Introductory Econometrics. South-Western Centage Learning, 4th edition.

cral plexu
sci•ence (sī'ə
tion, experim
phenomena.
nomena. c.
study. 2. M
so•lu•tion |sə'lo
noun
1 a means of solving a
solutions to financial an
• the correct answer
a liquid mixture in wh

MILJØMERKET
241 570
Trykksak

This publication may be ordered at
www.forskningsradet.no/publikasjoner

The Research Council of Norway
P.O.Box 2700 St. Hanshaugen
NO–0131 OSLO
Telephone +47 22 03 70 00
Telefax +47 22 03 70 01
post@rcn.no
www.rcn.no/english

Editor:
© The Research Council of Norway,
UTDANNING2020
www.rcn.no/utdanning

November 2012
ISBN 978-82-12-03151-7 (print)
ISBN 978-82-12-03152-4 (pdf)

Number of copies: 550
Printing: 07 Gruppen as

Design: Jentestreker as
Photo/illustration: Shutterstock