Assessing the re-use potential of research data in empirical educational research

Neuendorf, Claudia; Jansen, Malte; Pegelow, Lisa

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Empfohlene Zitierung / Suggested Citation:
Neuendorf, C., Jansen, M., & Pegelow, L. (2020). *Assessing the re-use potential of research data in empirical educational research.* (RatSWD Working Paper Series, 270). Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.49

RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

270

Assessing the re-use potential of research data in empirical educational research

Claudia Neuendorf,
Malte Jansen,
Lisa Pegelow

April 2020

SPONSORED BY THE

Federal Ministry
of Education
and Research

Working Paper Series of the German Data Forum

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the German section lists the complete contents of all issues in the series in chronological order.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD or of the Federal Ministry of Education and Research.

The RatSWD Working Paper Series is edited by:

since 2014 Regina T. Riphahn (Chair of the RatSWD)
2009–2014 Gert G. Wagner
2007–2008 Heike Solga

Assessing the re-use potential of research data in empirical educational research[1]

Claudia Neuendorf, Malte Jansen, and Lisa Pegelow

Berlin, April 2020

Introduction

In the last decade, the call for sharing research data has intensified within the educational research community in Germany. This development has taken place within the professional communities and it has been spurred by research funding organizations mandating researchers to share their research data. However, researchers and data centers alike are aware that not all data might be fit for re-use. Therefore, research data should be evaluated with respect to their analytical potential for re-use. Yet, criteria and processes for identifying data with high re-use potential are lacking. Thus, a workshop on the topic "Re-use potential of research data" was held on June 19th, 2018, at the Institute for Educational Quality Improvement, which was organized within the German Network for Educational Research Data (Verbund Forschungsdaten Bildung, VerbundFDB). Participants were ten researchers from different disciplines of empirical educational research in Germany. Representatives from the educational sciences, psychology, economics and sociology were present.

The aim of the workshop was to develop and discuss quality criteria for research data from the perspective of secondary users of research data.

doi: 10.17620/02671.49

[1] A German version of this report was published as: Neuendorf, C., & Jansen, M. (2020). Bericht vom Workshop "Nachnutzungspotenzial von Forschungsdaten" des Verbund Forschungsdaten Bildung (VerbundFDB). *forschungsdaten bildung informiert 8.* https://www.forschungsdaten-bildung.de/files/fdb-informiert-nr-8.pdf.

1 First part: Input lectures and discussion

1.1 Input lectures

In the morning, Malte Jansen and Claudia Neuendorf introduced the topic. This included 1) reasons for the need of an appraisal of research data, 2) the description of the selection process, 3) previous approaches to the appraisal of research data and the assessment of the potential for re-use.
The reasons why the evaluation of research data is necessary were explained: With an increasing amount of data available and the need to invest resources wisely, an assessment should be made at an early stage as to whether research data is suitable for subsequent use by a larger community and whether the effort of comprehensive documentation and preparation is justified. Furthermore, the process of data selection and evaluation was discussed. Research data centers typically apply a multi-stage selection process in which the fit of the data with the collection guidelines of the respective research data center (RDC) as well as its archivability (with regard to technical, legal and documentary requirements) are evaluated. If these criteria are met, an assessment of the potential for subsequent re-use is also carried out, which leads to a decision on the investment of resources for data publication.
A process that is consistent, comprehensible, transparent and efficient is desirable. This process must be clearly defined and documented and it must be based on clearly formulated and operationalizable selection and evaluation criteria. Standards already exist for the formal-technical criteria. However, criteria for evaluating the re-use potential from a content-based perspective are yet to be developed.

1.2 Discussion

In the ensuing discussion, it was suggested that the current secondary use of research data should be analyzed in order to empirically determine characteristics of widely used data sets. The GESIS data archive, for example, offers possibilities for such analyses. This presupposes, however, that operationalizable criteria already exist. It was pointed out, however, that such an empirical approach can only ever refer to the past, but that this does not cover the prediction of a future scientific interest in data sets.
Furthermore, the question was discussed as to who should assess the potential for the subsequent use of research data. The tendency was that scientists collecting the data themselves should make the assessment, for example in the project proposal.
Furthermore, it was discussed whether a high number of subsequent uses should be the goal of data provision - even data sets that have only been re-used in one study can be important for a research field.
It was important for the participants to emphasize that high-quality data sets that were generated with public funds in the past should be made available for re- use in order to grant all researchers fair and free access to these data. In this context, the datasets TIMSS transition, BIJU and ELEMENT-8 were mentioned as examples. Likewise, it should be easier to access data from the German national comparison tests (VERA-3 or VERA-8) for scientific purposes.

During the discussion, it already became clear that different criteria are used for judging the quality of data sets: For example, representative time series with frequent measurements were particularly interesting for sociologists in Germany (e.g. data from comparison tests). Sample sizes of less than 1000 were not considered re-usable. Other participants disagreed and said that even smaller sample sizes of high-quality studies can be interesting for some research questions.

2 Second part: development of criteria for subsequent use

In the second part, the participants were asked in a brainstorming session to name aspects that come to their minds when assessing the quality of research data. For this purpose, the categories 1) documentation, 2) preparation, 3) content and 4) methodology were specified.
The workshop participants were then divided into three groups, with each group representing a different area of expertise: educational science, psychology and sociology/economy. The groups had one hour to discuss aspects related to their field of expertise for assessing the quality of research data, collect their points on posters and, if possible, arrange them according to importance.
Afterwards, the results of the group work were discussed in a poster session. Some of the topics were mentioned for several disciplines; these will be assigned and summarized accordingly in the following. The poster illustrations are in the appendix.

2.1 Documentation

The participants agreed that a **short summary** or a data manual is necessary, in which core information (sample, study design (e.g. sampling, survey details) and topics/contents) of the study are clearly presented.

Codebooks or scale documentation are considered important. As minimum information for each scale 1) the name and a short description of the construct, 2) the source of the scale and 3) the item wordings are named.

This means that the items must be assigned to a scale. Also, inversions of items and other re-codings that have already been made must be indicated. Additional information that often appears in the scale documentation is the number of cases, information about missings, mean values and internal consistencies. However, information that researchers themselves can calculate from the data is not absolutely necessary in the eyes of the researchers. The format of the documentation (whether Excel, Word or in SPSS labels) is not decisive.

Documentation also includes that the naming of scales in the data set or the scale manual is theory-based and, if possible, uniform across studies. However, this is seen as a challenge, since there are nuanced differences in the meaning of certain constructs depending on their theoretical provenance (e.g. cognitive activation according to Klieme vs. Seidel). However, an

entry or cross-referencing in the "Database on School Quality"[2] could help here. The social science item and scale compilation[3] and PSYNEX tests[4] also provide information about scales and tests that could be referenced in the documentation.

In addition, the **original questionnaires** are considered important.

Both scale documentation and original questionnaires pose a challenge for the documentation if they are protected by copyright law.

It would also be desirable to have a **file documenting potential pitfalls** or FAQs for each dataset, pointing out special features of the dataset (e.g. special features in linking student and teacher data, high rates of attrition, other **anomalies** and unusual features that primary researchers or other users may have encountered during preparation and evaluation).

Uniform documentation (with regard to the information contained and the format) across studies would facilitate subsequent use (e.g. through automated routines at the RDC).

On several occasions, good **keywording** of the studies was called for in order to improve the retrieval of relevant data sets. However, a mapping of the terms searched for in different disciplines would be important (in this case, techniques and expertise for keywording from library and information sciences should be used).

It would be helpful if an RDC offered a **description of the depth of documentation** on the study website so that researchers would know transparently what to expect when applying for the study.

2.2 Preparation

In the discussions, it became clear that documentation and preparation are closely related, since data preparation is carried out partly for documentation purposes and, conversely, the documentation of the processing steps themselves is also important.

In the case of preparation, the researchers demand

- consistent logic in the naming of variables
- a uniform documentation of missing values
- information on the nested data structure (school ID, class ID, student ID)

A major point of discussion was the question in which processing state data sets should be delivered. Basic data cleansing should already have taken place. Apart from this, however, the scientists would like to see both raw data and already processed data (e.g. indicators already formed (e.g. HISEI, class membership), scaled data, possibly imputed data) or a

[2] DaQS (https://daqs.fachportal-paedagogik.de)

[3] ZIS (https://zis.gesis.org)

[4] https://www.psyndex.de/index.php?wahl=PSYNDEXTests

corresponding processing syntax. This would improve the traceability of the processing, provide maximum flexibility for secondary researchers and at the same time save time-consuming and error-prone recoding.

In the case of longitudinal studies, it is also desirable to have a uniform designation of waves as well as to indicate the participation status and new additions to the data.

Finally, providing a well usable file format including a syntax for importing the data into different statistical programs is desirable.

More important than the question of good formatting, however, is that the data is available at all.

2.3 Data access

An additional point raised by researchers is data access. On the one hand, this concerned differences in **data protection standards** between data centers. Some researchers would like to see a more **uniform approach** here. The tension between data protection and Open Data for the scientific community was perceived and the scientists would like the RDC to champion the interests of data users.

Further, **low-threshold, free** access to data was important to the researchers. If remote computing is the only option, online access, such as remote desktop, should be used instead of the JosuA portal, which leads to time delays.

2.4 Contents of the data

Regarding data content, there were numerous requests to improve data quality. Researchers would like to have a high **data density**: as many 1) measurement points, 2) study participants, 3) variables and 4) instruments as possible.

With regard to variables and instruments, there was consensus on some points, but there were also differences between the different subject-fields. For example, everyone wanted comprehensive **background information** on the study participants (see illustrations). Here, it is important that the constructs are recorded in a **standardized** way so that different studies can be related to each other in terms of content and, for example, trend analyses can be carried out. This also applies to different waves of a longitudinal study. There were also participants who went even further and referred to the examples of different countries and states in which it is possible to obtain very detailed information from official statistics for scientific purposes (e.g. school entrance examinations, VERA data, "core data set", social security numbers). To this end, the researchers would like to see closer cooperation between the RDC and the authorities.

For the participating social scientists, the topic of **regional information** was particularly important. This information makes it possible to investigate research questions on school transitions, effects of streaming, teaching and educational reforms. Also, by providing

regional information that is as detailed as possible, the combination of datasets with external information becomes feasible. Examples of regional information are very rough subdivisions (such as city-country), information that allow the comparison of different educational systems within the Federal Republic of Germany (e.g. federal states), or very small-scale subdivisions (such as district and municipal level). One way to make this possible while maintaining data protection would be routines that allow external information (e.g. from official school statistics[5]) on the school location to be linked to the data set via syntaxes, without the regional identifier itself having to be issued by data centers.

Psychologists and educational scientists also wanted to see the presence of **cognitive, psychosocial, emotional and motivational constructs** as well as information on **teaching quality and teaching processes**. In this context, the use of scales with **multiple indicators** (items) was preferable to the use of single item scales.

A discussion arose on the question of whether classic scales (e.g. test anxiety) or innovative scales (e.g. use of digital media) have a greater potential for subsequent use. Ultimately, there was agreement that **originality** and **connectivity** (to international research, e.g. use of established constructs that are recorded relatively similarly in different studies) should be balanced.

2.5 Methodology

The density of data also plays a role in the methodology of the study: **longitudinal studies** are particularly well suited for re-use, especially longitudinal studies at individual level (panel surveys). The continuity and comparability of the survey over time (same operationalization/recording of constructs) is of particular importance. Examples of large, longitudinal studies in Germany are NEPS and SOEP, where the researchers stress the importance to the connection to the scientific community, which is guaranteed by the structure of these two panel studies. Potential is also seen for longitudinal links in school achievement studies: for example, between students participating in IQB Trends in Student Achievement both in primary and secondary school, in the sampling of a school panel in the IQB Trends in Student Achievement, or in establishing links between the IQB Trends in Student Achievement and NEPS. The publication of PISA-I-Plus-2012/2013 data is also highly anticipated by the scientific community. In this context, the question of whether a data set allows causal inferences was also mentioned as a criterion. In the eyes of some, however, this was a rather difficult criterion, since disciplines also differ in their understanding of causal inference.

Indicators related to the **sample** were mentioned as criteria of data quality: the size of the sample, its representativeness (elaborate sampling) and a high sampling rate. With regard to representativeness, however, there were also voices who felt that this aspect was not so important.

[5] Their availability could still be improved.

One aspect of the methodology of a study that considerably improves the potential for re-use is the **combination of different survey instruments** (tests, questionnaires, observations, etc.) and **different observers** (e.g. teachers, pupils, parents, etc.).
The question of whether experimental studies would in principle have a lower potential for subsequent use tended to be answered negatively. On the one hand, many experimental studies focus on a very specific issue, so that their analysis potential is often already exhausted during primary use, but on the other hand, the potential for re-use could be improved if the experimental design was combined with questionnaires enriched with standard indicators. There are examples of large experimental field studies with a high potential for re-use (e.g. on teaching methods: IGEL study at the DIPF).

2.6 Further discussion themes

One criterion was that a data set should enable high-ranking publications, and the question arose how this could be predicted.

When asked about other easily operationalizable criteria (such as the size of the consortium, interdisciplinarity, funding amount), the researchers felt that these had a rather small influence on the potential for re-use.

The researchers also discussed that the potential for re-use should not be confused with the importance of the study. There are many small-scale studies that are very important, but not very suitable for re-use.

Different aspects were discussed with regard to the question of whether **actuality** is an important criterion. While, on the one hand, a study using older data is often accused in the review process that the results may already be outdated due to the age of the data and the change in the educational context, historical data, on the other hand, allow comparisons over time. It is therefore worthwhile to prepare and provide both current and older data sets (e.g. FIMSS, Fend's studies). It would be ideal if earlier studies could be linked to current data, e.g. by building on old studies to continue them.

The researchers agreed that there is **historical data** that is of great value as a basis for an entire research area, but that have never been made available to the scientific community for re- and secondary analysis. It was argued in the discussion that these data were also collected with taxpayers' money and therefore should not be regarded private property by those who collected them. In the interests of fairness, transparency and good scientific practice, efforts should be made to prepare and make these data available.

While data sets from some projects are known to scientists (e.g. LIFE-Study (Fend), SCHOLASTIK incl. the highly gifted add-on by Heller, PALMA, TOSCA, BIJU, PISA-Plus), it can be assumed that further data sets still exist, but these are, for example, held by emeritus professors or on servers to which hardly anyone has access. There is a desire that the entire community be called upon to raise such old "data treasures" and to invest resources, prepare them and make them available.

Attendees

Experts

- Dr. Katrin Arens (DIPF, Frankfurt a. M.)
- Dr. Gwendolin Blossfeld (University of Bamberg)
- Prof. Katja Görlitz (German Institute for Economic Research (DIW), Berlin and FU Berlin)
- Prof. Marcel Helbig (University of Erfurt, Social Science Research Center Berlin (WZB))
- Dr. Lydia Kleine (Leibniz Institute for Educational Pathways (LIfBi), Bamberg)
- Dr. Susanne Kuger (German Youth Institute (DJI), Munich)
- Prof. Rebecca Lazarides (University of Potsdam)
- Prof. Martin Neugebauer (FU Berlin)
- Dr. Rolf Strietholt (TU Dortmund University)
- Dr. Sebastian Wurster (Johannes Gutenberg University Mainz)

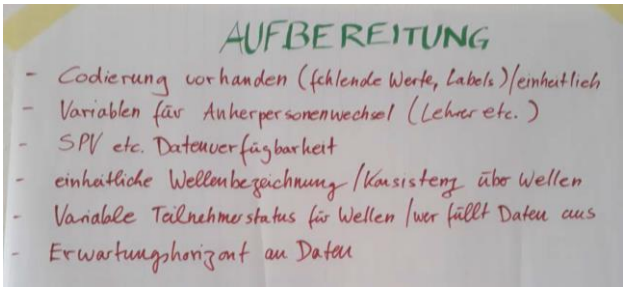
IQB Staff

- Dr. Malte Jansen (Head of Research Data Center at the Institute for Educational Quality Improvement (FDZ at IQB))
- Claudia Neuendorf (Research Associate, German Network for Educational Research Data)

Annex

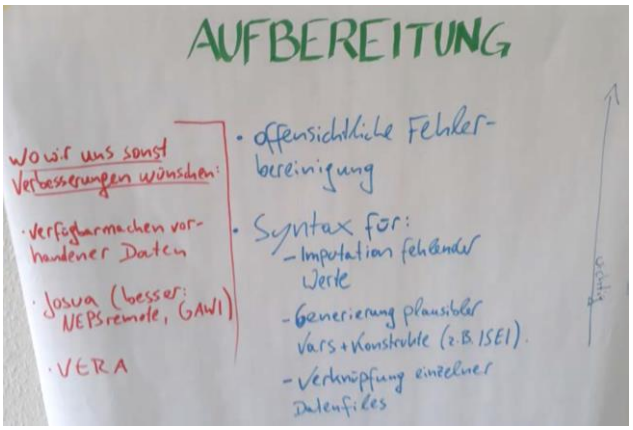
The following is an overview of the posters developed by the participants on the four topics preparation, documentation, content and methodology. The poster in the upper row was created by the group of psychologists, the poster from sociology/economy is in the middle row and the results from educational science are shown at the bottom.

Poster on the topic of preparation



Poster 1

- Coding available (missing values, labels)/uniform
- Variable for change of personnel (teachers etc.)
- SPV etc. Data availability
- Uniform designation/consistency across waves
- Variable "participant status for waves" / who fills in data
- Expectation horizon for data

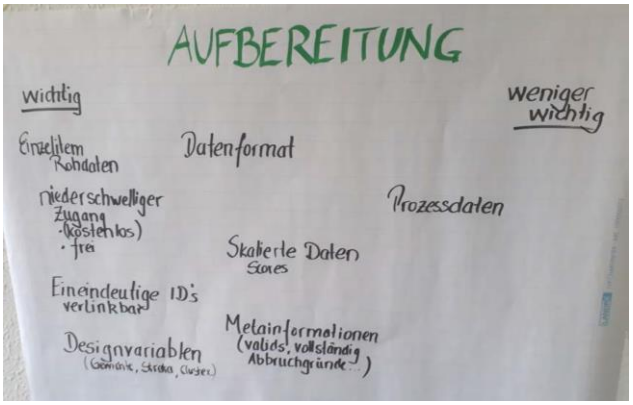


Poster 2:

- obvious fixes and data cleaning
- syntax for
  - imputation of missing values
  - generation of plausible variables and constructions (e.g. ISEI)
  - Linking of individual data files

Where we would otherwise wish for improvements:

- Making existing data available
- Joshua (Better: NEPSremote, GAWI)
- VERA

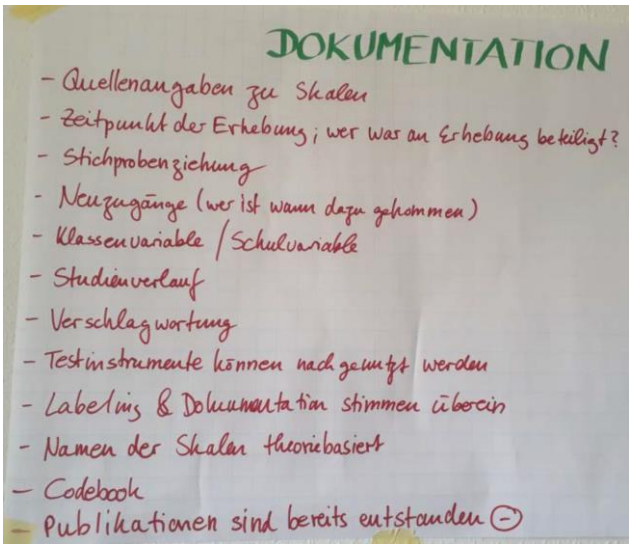


Poster 3
(top is most important, bottom least):
single item raw data
low-threshold access (free, open access)
unique IDs, linkable
Design variables (weights, strata, clusters ...)

scaled scores
meta-information (valid, complete reasons for termination...)

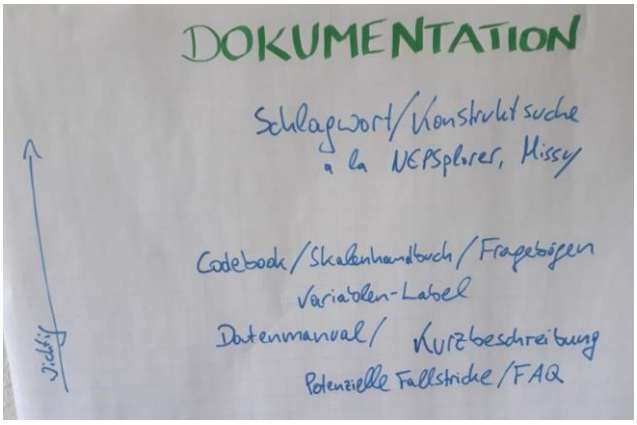
Process Data

Poster on the topic of documentation



Poster 1

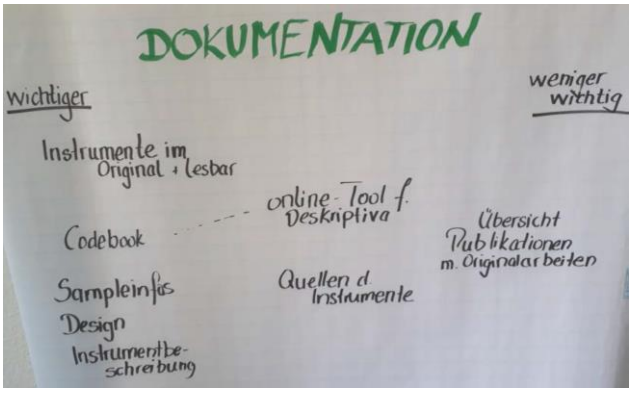
- Source information on scales
- Time of the survey; who was involved in the survey?
- Sampling
- Information on refresher samples (who has joined when)
- Class variable/school variable
- Course of studies
- Keywording
- Test instruments can be reused
- Labeling & documentation match
- Names of the scales theory-based
- Codebook
- Publications have already been produced



Poster 2

(top is most important, bottom least):

- Keyword/construct search a la NEPSplorer, Missy
- Codebook/scale handbook/questionnaires/variable label
- Data manual/short description/potential fall lines/ FAQ

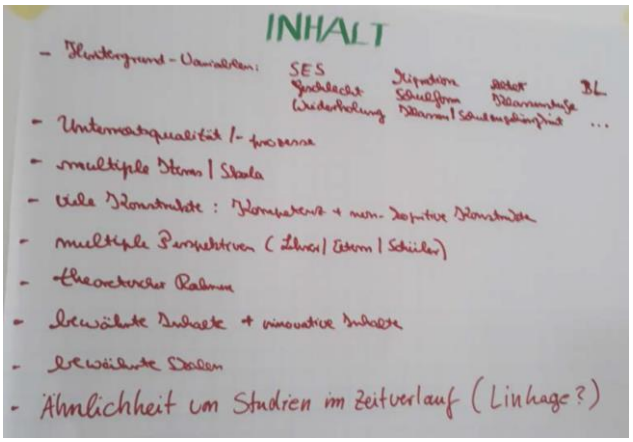


Poster 3

(top is most important, bottom least):

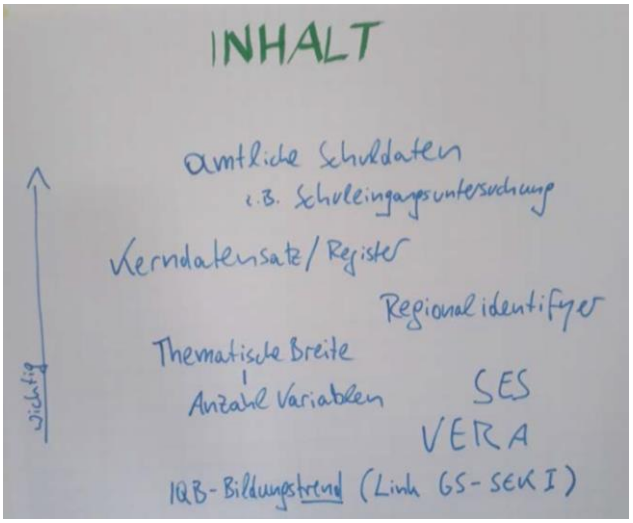
- Original instruments + readable
- Codebook --- online tool for descriptives
- Sample information
- Design
- Instrument description
- Sources of the instruments
- Overview of publications with original works

Poster on the topic of contents



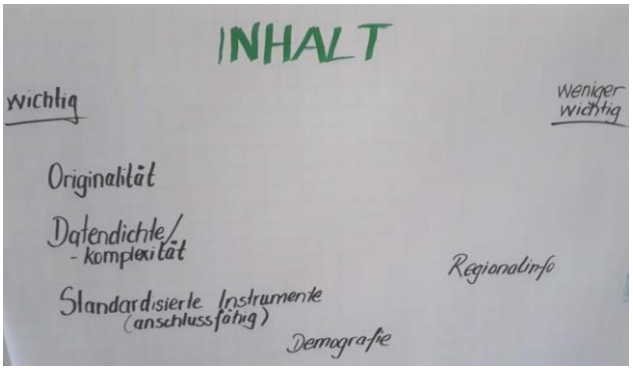
Poster 1

- Background variables: SES, Migration, Federal State, Repetition, Type of School, Grade Level, Gender, Classes/School Affiliation ...
- Teaching quality/processes
- Multiple Items/Scale
- Many constructs: Competence + non-cognitive constructs
- Multiple perspectives (teachers/parents/students)
- Theoretical framework
- Proven content + innovative content
- Proven scales
- Similarity around studies over time (Linkage?)



Poster 2
(top is most important, bottom least):

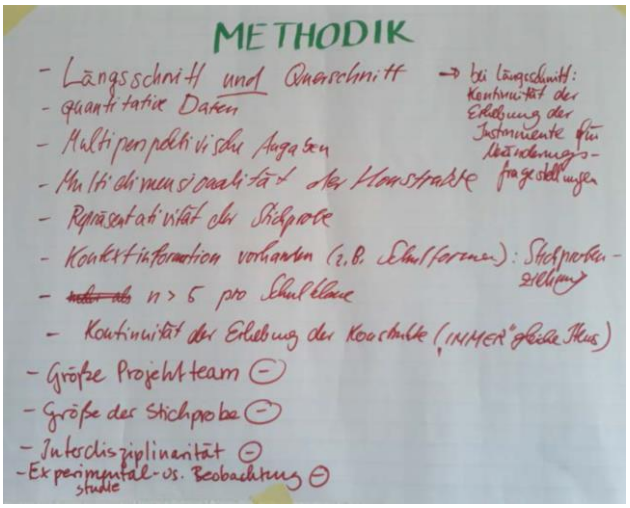
- Official school data (e.g. school entrance examination)
- Core data set/register
- Regional Identifier
- Thematic breadth --- Number of variables
- SES
- VERA
- IQB Trends in Student Achievement (Link GS-SEK I)



Poster 3
(top is most important, bottom least):

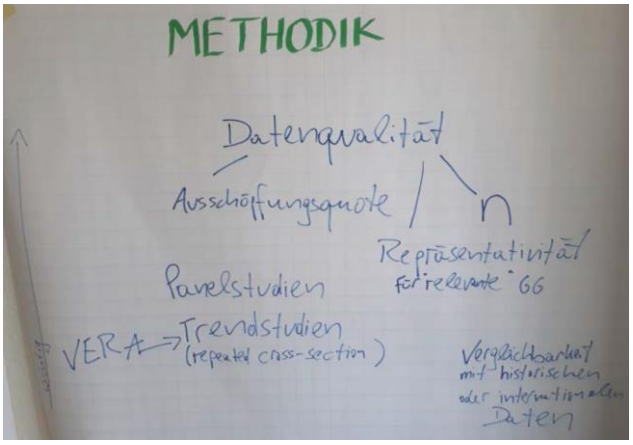
- Originality
- Data density/complexity
- Standardized instruments (connectable)
- Demography
- Regional information

Poster on the topic of methodology



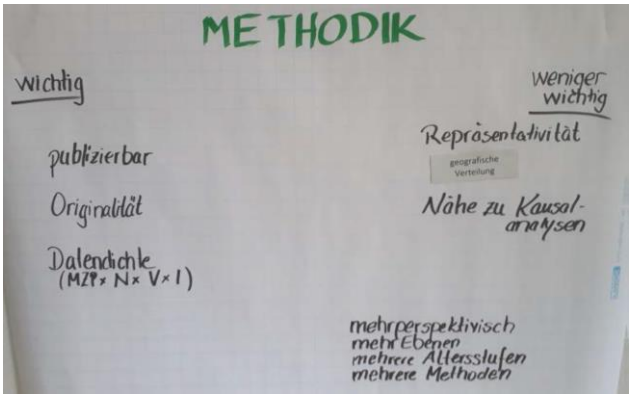
Poster 1

- Longitudinal section and cross-section -- for longitudinal section: continuity of the survey of instruments for change
- Quantitative data
- Multi-perspective tasks
- Multidimensionality of the constructs
- Representativeness of the sample
- Context information available (e.g. types of school): Sampling
- n > 5 per school class
- Continuity of the survey of constructs ("Always" same items)
- Size of project team –
- Size of the sample –
- Interdisciplinarity –
- Experimental vs. observational study –



Poster 2
(top is most important, bottom least):

- Data Quality
  - Response rate
  - Representativeness for "relevant" population
  - n
- Panel studies
- VERA → Trend studies (repeated cross-section)
- Comparability with historical or international data



Poster 3
(top is most important, bottom least):

- Publishable
- Originality
- Data density (TxNxVxI)
- Representativeness
- Geographic distribution
- Proximity to causal analysis
- Multi-perspective
- More levels
- Several age groups
- Several methods