



# IQB Trends in Student Achievement 2018: A Large-Scale Educational Assessment Study in Germany

JOHANNA BUSSE

BENJAMIN BECKER 

SEBASTIAN WEIRICH 

STEFAN SCHIPOLOWSKI

\*Author affiliations can be found in the back matter of this article

SPECIAL COLLECTION:  
EDUCATION DATA

DATA PAPER

]u[ubiquity press

## ABSTRACT

The *IQB Trends in Student Achievement 2018* is a nation-wide large-scale educational assessment using a trend design. The study includes achievement tests on mathematics, biology, chemistry, and physics as well as background questionnaires for students, their parents, subject teachers, and school principals. The datasets include data of 51.511 grade-9 students from 1.473 German schools assessed in 2018, data of 5.026 teachers, and 1.264 school principals. The 6 datasets and the study documentation are available via the Research Data Centre (FDZ) at IQB and include a wide variety of constructs such as achievement estimates, indicators of general cognitive abilities, motivational and socio-emotional constructs as well as data on social networks.

## CORRESPONDING AUTHOR:

**Johanna Busse**

IQB Berlin, DE

johanna.busse@iqb.hu-berlin.de

## KEYWORDS:

Educational large-scale assessment; policy reporting; mathematical and science achievement; monitoring

## TO CITE THIS ARTICLE:

Busse, J., Becker, B., Weirich, S., & Schipolowski, S. (2022). IQB Trends in Student Achievement 2018: A Large-Scale Educational Assessment Study in Germany. *Journal of Open Psychology Data*, 10: 18, pp. 1–9. DOI: <https://doi.org/10.5334/jopd.71>

## (1) BACKGROUND

The IQB Trends in Student Achievement 2018-study (Stanat et al., 2019) is a nation-wide educational large-scale assessment that constitutes a central part of the educational monitoring in Germany as mandated by the Standing Conference of the Ministers of Education of the Federal States in Germany (KMK). The study evaluates whether (and to which extent) 9<sup>th</sup> grade students in the federal states (Länder) in Germany meet the proficiency expectations established with the educational standards introduced by the KMK. The assessment focuses on mathematics and science (i.e., biology, chemistry, and physics) achievement. Students' achievements are measured using a broad variety of standards-based test items. Stanat et al. (2019) report the results for Germany and the federal states using proficiency level models.

As education policy in Germany aims to reduce students' disparities in educational outcomes with regard to various socio-demographic factors (KMK, 2016; Stanat et al., 2019), the study also evaluates whether students' proficiencies vary on average with respect to gender, social background, and immigration background (Blossfeld et al., 2007). To investigate changes in educational achievement of 9<sup>th</sup> graders – including changes in disparities – over time, the results of the IQB Trends 2018-study were linked to a prior assessment in 2012 (Pant, Stanat, Schroeders, et al., 2013).

Using additional questionnaires for students, parents, teachers, and school principals, the study also evaluated students' motivational characteristics regarding mathematics and the sciences (e.g., interests and self-concepts), parents' educational background, and aspects of teacher training in mathematics and the sciences, to name only a few.

In total, 6 datasets are provided that contain (1) student-level data (e.g., achievement test results, responses from student and parent questionnaires), (2) social networks, (3) data from principals' questionnaires, responses from (4) general teachers' questionnaires and (5) class-specific teachers' questionnaires, and (6) matching IDs. The datasets contain data that is of interest for educational research but also for psychological research questions. The next follow-up study is scheduled for 2024. It will be possible to link the new data with the 2018 study and the study from 2012.

Detailed information on the study aims, assessed constructs, and the study design can be found in the report (Stanat et al., 2019; for an English summary, see Stanat et al., 2020).

## (2) METHODS

### 2.1 STUDY DESIGN

The study is part of a trend design, examining cohorts of 9<sup>th</sup> graders at regular intervals (i.e., every 6 years). The

target population of the study discussed here was the population of 9<sup>th</sup> grade students in Germany in the year 2018. All students were included regardless of potential special educational needs (SEN). The only exception were students with SEN in the cognitive development domain as well as recently immigrated students who had visited school in Germany for less than a year.

Several instruments were used in the study. Achievement tests, cognitive ability tests, and the student questionnaire were administered using paper and pencil, whereas parents could choose between a paper version and an online version of the questionnaire that was available in Arabic, English, German, Polish, Russian, and Turkish. Teachers in the test subjects and principals responded to an online questionnaire. The questionnaires for students and parents were designed to provide individual background information about the students and their learning environment in school and at home. Teachers and principals were asked about important characteristics of the composition of the learning groups, the schools (e.g., support and leisure activities provided by the school) and the lessons (e.g., instructional quality).

In the first part of the test session, students were given achievement tests with either mathematics items, science items, or both mathematics and science test items. A large number of different tasks and items was used for the achievement tests to enable broad construct coverage. As the resulting total number of tasks would have led to an enormous workload for individual test-takers, the tasks were divided into different test booklets which contain only a subset of all tasks according to a multiple matrix sampling design (Gonzalez & Rutkowski, 2010). SEN students were provided with adapted test booklets, containing less and easier tasks than regular booklets.

All students were given 120 minutes to work on the achievement tests. Afterwards, they worked on several tasks designed to measure general cognitive abilities and completed the student questionnaire. The complete testing session including the general cognitive abilities section and the student questionnaires was highly standardised and took about 4 hours including short breaks. The full schedule can be seen in Table 1. Organization and implementation were commissioned by the International Association for the Evaluation of Educational Achievement (IEA Hamburg). The administration was carried out by trained personnel with no prior affiliation to the respective schools.

The achievement test items were constructed to fit to the one-parameter item response theory (IRT; Embretson & Reise, 2000; Hambleton et al., 1991) Rasch model (Adams & Wu, 2007; Fischer, 2006) and were dichotomously scored as correct or incorrect. If sets of common items are repeatedly used in different studies (e.g., 2012, 2018, and in the follow-up 2024),

DURATION IN MINUTES	ACTIVITY
15	Beginning of the test session, distribution of the test booklets, general instructions for the students
60	Achievement tests part 1 (mathematics or science)
15	Break
60	Achievement tests part 2 (mathematics or science)
15	Break
20	General cognitive ability tests
45	Student questionnaire
5	End of the test session, collection of materials

**Table 1** Study Schedule for Students.

Note: SEN students had a shorter version of the questionnaire and adapted achievement tests.

student achievement in these cohorts may be linked to a common scale under the assumption that the difficulty of items is invariant across measurement points over time. The corresponding linking design is considered a “common-items non-equivalent groups equating design” (Holland, 2007; Kolen & Brennan, 2004; von Davier et al., 2008). In consequence, achievement scores of students can be compared between cohorts. The achievement test data (plausible values, see section 3.3) are already linked to the previous measurement (2012), which means that the achievement test data are comparable between 2012 and 2018. For further details, see also Becker et al. (2019).

## 2.2 TIME OF DATA COLLECTION

23<sup>th</sup> April – 22<sup>th</sup> June 2018

## 2.3 LOCATION OF DATA COLLECTION

Data were collected in each of the federal states in Germany within the facilities of the randomly sampled schools.

## 2.4 SAMPLING, SAMPLE AND DATA COLLECTION

The sampling procedure was designed to represent the population of 9<sup>th</sup> grade students in Germany by sampling public and private schools (including schools for SEN students, *Förderschulen*) from school lists provided by the statistical state offices. A multistage sampling procedure was implemented using federal state and school type (academic-track *Gymnasium* schools, non-academic schools, and SEN schools) as explicit sampling strata. For practical reasons, only SEN schools with students in the focus domains “Learning”, “Language” and “Emotional and Social Development” were sampled.

The second step in the sampling procedure was a random selection of one 9<sup>th</sup> grade class at each academic-track school and two 9<sup>th</sup> grade classes at each non-academic school (if available). At SEN schools, the complete 9<sup>th</sup> grade was selected for the study. All

students in each selected class were supposed to participate in the study.

Participation in the achievement tests and general cognitive ability tests was mandatory for both the schools and the students (except for some private schools) whereas the student questionnaire was voluntary in seven of sixteen states in accordance with the applicable state laws. Parent questionnaires were voluntary in all states. More information can be found in Mahler et al. (2019).

The datasets include data<sup>1</sup> of 51.511 students from 1.473 schools in all 16 German federal states, 19.027 parents, 1.264 principals, and 5.026 teachers.

As the sampling probabilities differed between federal states, the proportion of students per state in the sample does not correspond to the proportion of students per state in the population. Hence, the datasets also include weights on the student, class, and school level. Including weights into analyses is necessary if parameters should represent the population of 9<sup>th</sup> grade students in Germany.

Statistical inference in clustered surveys is often based on resampling techniques (Lumley, 2004), for example, bootstrap or jackknife. As the sampling design of the IQB Trends 2018-study allows using a jackknife procedure, the datasets include jackknife zone and jackknife replicate indicator variables which should be incorporated in secondary analyses to provide unbiased standard errors for parameter estimates (Weirich, Hecht, Becker, et al., 2022). Software packages suitable for incorporating sampling weights and jackknife procedures include, for example, Wesvar (Westat, 2000) as well as the R packages survey (Lumley, 2019), eatRep (Weirich, Hecht, & Becker, 2022) and BIFIEsurvey (Robitzsch & Oberwimmer, 2019).

## 2.5 MATERIALS/SURVEY INSTRUMENTS

The student-level dataset contains the data collected using the achievement tests, the general cognitive ability tests, the student questionnaire (excluding the social

network data), the parent questionnaire, and some variables provided by the schools and the statistical state offices (e.g., school type).

The proficiency data were collected via achievement tests in mathematics, biology, chemistry, and physics. The tasks were developed by experienced teachers from all states under the direction of the IQB and experts on didactics in the respective subjects from several universities. All tasks and items went through a rigorous pre-testing process prior to the study and the majority of the items had already been used in the first national assessment in 2012. 61% of the mathematics items and nearly 100% of the science items were already used in 2012 (Becker et al., 2019). New items for the 2018 assessment were developed and pre-tested using the same quality standards. Items with closed answer formats (e.g., multiple choice items) as well as items with open response formats were used. Mathematics items operationalise five proficiency domains (core themes) as described in the educational standards: *Numbers, measurements, space and shape, functional relations, and data and chance*. In each of the science subjects, proficiencies were assessed in the areas *subject knowledge* and *scientific inquiry*. Overall, 606 mathematics items and 386 science items were used in the current study.

The general cognitive ability tests comprised C-tests and a nonverbal reasoning test. C-tests are a variant of the cloze principle (Klein-Braley, 1997): In a short coherent text, every second half of every second word is missing; participants have to fill in the gaps in a meaningful and linguistically correct manner. C-tests are widely used as indicators of general language proficiency (e.g., Eckes, 2017). In the study at hand, two different texts were used with 30 gaps each. Reasoning was measured with the figural reasoning scale of the Berlin Test of Fluid and Crystallised Intelligence (BEFKI; Wilhelm et al., 2014). The scale is comprised of 16 items; each item consisted of a sequence of geometric shapes whose elements changed according to implicit rules. To complete the tasks, students had to infer these rules and choose the next two shapes in the sequence from a number of given alternatives. Some of the general cognitive ability tests were also used in the 2012 study and will be used in the forthcoming 2024 follow-up study.

The questionnaires for the students included socio-demographic questions about their personal background and their family, but also items related to subject-specific self-concepts, interest in the sciences and mathematics, instructional quality, and students' perception of their school.

The parent questionnaire was designed to gather socio-demographic information which is essential in analysing achievement disparities related to the social background and immigration background of the students.

This included information about the parents' occupations, their education, the birth countries of the parents and their perception of the school.

The social networks dataset includes students' assessment of their social relationships within the class context. Students were asked four questions regarding their social networks: (1) who their friends are, (2) with whom they are spending their breaks, (3) who they don't want to sit next to, and (4) who they ask for help in case of problems. For this purpose, lists of all students within the respective class were created and students were asked to tick boxes for all four questions and all of their class mates.

The questionnaire for school principals included questions about their personal background, general information about the school, the participation and implementation of comparison tests designed to improve the quality of instruction (*Vergleichsarbeiten* VERA; KMK, 2016), cooperation among teachers, all-day school, supportive, social, and other activities offered by the school, and the integration of refugee students.

Two datasets stemming from the teacher questionnaires are provided: A general one including information about the teachers' personal background, their work at school, their education, on-the-job training, their opinion about and their use of VERA, their teaching in general, their school, and cooperation with colleagues and a class-specific dataset including information about teaching in the specific class or course selected for participation in the study.

The sixth dataset includes matching IDs and is on the student subject level (a single student can have multiple rows in the dataset for different subjects). These matching IDs serve the technical purpose of allowing a matching of students and teachers via the class-specific teacher dataset. Note that all other datasets can be linked by ID variables already contained in the respective datasets.

Detailed information about all items and scales used in the study can be retrieved from the documentation (Becker et al., 2022).

## 2.6 QUALITY CONTROL

All items of the achievement tests were previously tested in multiple studies, such as the 2012 assessment and designated pilot studies (Mahler et al., 2020; Mahler et al., 2019; Pant, Stanat, Pöhlmann, et al., 2013). The contents of the questionnaires were also empirically evaluated prior to their use in the study. References for the questionnaire items and scales as well as some psychometric properties (e.g., internal consistency) are available in the documentation (Becker et al., 2022). The assessment itself was highly standardised to ensure objectivity and conducted by trained personnel. As for the data itself, all steps during data cleaning

and processing were conducted by two persons independently. Additionally, integrity of the data was thoroughly checked prior to publication by the Research Data Centre (FDZ) at IQB.

## 2.7 DATA ANONYMISATION AND ETHICAL ISSUES

An essential condition for the study is to protect the rights of all participating persons. For this, all data were collected using pseudonyms, that is, with a study-specific ID number which enables matching of different data sources. Key variables with low frequency counts for some or all categories were recoded and free text fields were removed from the datasets by the FDZ.

All study participants were informed about the study aims, contents, procedures, privacy protection, and their legal rights prior to the assessment. All questionnaires and procedures were also reviewed and formally approved by state authorities. In some federal states of Germany and certain school types, a permission of the parents was needed for their students to participate in the study or to fill out the student questionnaire. Only students with the necessary permission took part in the study or filled out the questionnaire.

## 2.8 EXISTING USE OF DATA

There are already some publications using the data. The comprehensive study report for the IQB Trends in Student Achievement 2018-study (Stanat et al., 2019) summarizes the main findings and is primarily intended for policy makers and the education administration, but

also for the general public. Schipolowski et al. (2021) investigated mathematics and science proficiency of young refugees in secondary schools in Germany and compared them to other students with and without an immigrant background. Furthermore, Schneider et al. (2022) published an article about gender differences in academic self-concepts and interests and how the temporal changes of these constructs could be explained by differences in academic achievement between cohorts.

## (3) DATASET DESCRIPTION AND ACCESS

### 3.1 REPOSITORY LOCATION

[https://doi.org/10.5159/IQB\\_BT\\_2018\\_v1](https://doi.org/10.5159/IQB_BT_2018_v1)

### 3.2 OBJECT/FILE NAME

The datasets can be requested online via the FDZ at IQB under <https://www.iqb.hu-berlin.de/fdz/Datenzugang>. More details on the datasets and other files are provided at [https://www.iqb.hu-berlin.de/fdz/studies/IQB-BT\\_2018](https://www.iqb.hu-berlin.de/fdz/studies/IQB-BT_2018) and <https://www.iqb.hu-berlin.de/bt/BT2018/Bericht/>. Table 2 lists the datasets and documentation files which are part of the IQB Trends in Student Achievement 2018 study. Note that the linking error data set only is relevant if competencies are compared between the studies in 2012 and 2018. The LV 2012 plausible value data set contains some modifications to the original LV2 2012 plausible values which are only relevant if subdomains in math are compared between 2012 and 2018.

FILE NAME	TYPE	DESCRIPTION	LANGUAGE
BT2018_SchuelerInnen.sav	Dataset	Student-level dataset for SPSS containing 1.403 variables	German
BT2018_SchuelerInnen_sozialeNetzwerke.sav	Dateset	Social network dataset for SPSS containing 146 variables	German
BT2018_Schulleitungsfragebogen.sav	Dataset	Principal dataset for SPSS containing 176 variables	German
BT2018_Lehrkraeftefragebogen_allgemein.sav	Dataset	General teacher dataset for SPSS containing 321 variables	German
BT2018_Lehrkraeftefragebogen_lerngruppenspezifisch.sav	Dataset	Class-specific teacher dataset for SPSS containing 61 variables	German
BT2018_Matching_SchuelerInnen_Lehrkraefte.sav	Dataset	Matching IDs dataset for SPSS containing 6 variables	German
BT2018_LinkingFehler.sav	Dataset	Linking errors for 2012-2018 trend for SPSS containing 15 variables	German
LV2012_PVs.sav	Dataset	LV12 updated math subdomains data for SPSS containing 151 variables	German
IQB_BT2018_Skalenhandbuch.pdf	Documentation	Codebook that contains information about all variables and datasets	German
IQB_BT2018_Berichtsband.pdf	Documentation	Policy report of the study	German
IQB_BT2018_Summary_English.pdf	Documentation	Summary of the report in English	English

**Table 2** Provided files.



### 3.3 DATA TYPE

Data from questionnaires are mostly primary data. Some information about the schools is secondary data provided by the statistical state offices. The datasets also include processed data for some variables based on primary data such as the results of the achievement tests. Specifically, the data contain achievement estimates as 15 plausible values (PVs) per person and domain (Mislevy et al., 1992; von Davier et al., 2009); raw test item responses are not available.

PVs result from a one-parameter logistic item response theory model which incorporates item responses as well as a multitude of demographic variables in a conditioning model and were generated using the R package TAM (Robitzsch et al., 2022). PVs can be used as dependent variables in analyses with achievement data. The data also contain proficiency levels for the students based on the PVs and proficiency level models developed by the IQB. Missing values on key variables were imputed using multiple imputation (Schafer & Graham, 2002) via the fully conditional specification (FCS) (van Buuren et al., 2006) implemented in the R package mice (van Buuren & Groothuis-Oudshoorn, 2011). In these cases, 15 imputations for the respective variables are included. Further information on the conditioning model can also be seen in the study codebook (Becker et al., 2022).

### 3.4 FORMAT NAMES AND VERSIONS

The datasets have the SPSS format .sav.

### 3.5 LANGUAGE

Variable and value labels as well as the study documentation are in German.

### 3.6 LICENSE

The IQB Trends in Student Achievement 2018-study data sets are archived at the FDZ at IQB. They can be obtained following this link: [https://www.iqb.hu-berlin.de/fdz/studies/IQB-BT\\_2018](https://www.iqb.hu-berlin.de/fdz/studies/IQB-BT_2018). The data is available for academic research which is ensured via a data usage agreement with the FDZ at IQB.

### 3.7 LIMITS TO SHARING

The data of the IQB Trends in Student Achievement studies are accessible for the scientific community. Data is available to registered data users only. For registration, interested researchers can follow the instructions on the FDZ homepage. The data can also be shared with university students, for example for the preparation of a Bachelor or Master thesis. Further guidelines regarding the data sharing policy of the FDZ at IQB can be found in the general terms and conditions of the FDZ at IQB: [https://www.iqb.hu-berlin.de/fdz/Datenzugang/fdz/20190131\\_FDZ\\_Ver\\_3.pdf](https://www.iqb.hu-berlin.de/fdz/Datenzugang/fdz/20190131_FDZ_Ver_3.pdf).

### 3.8 PUBLICATION DATE

30/06/2022

### 3.9 FAIR DATA/CODEBOOK

All data are following the FAIR data principles. Data can be officially requested at [https://www.iqb.hu-berlin.de/fdz/studies/IQB-BT\\_2018](https://www.iqb.hu-berlin.de/fdz/studies/IQB-BT_2018). The content and structure of all datasets is described on the website, as well as in a detailed codebook. All data have detailed meta data, and further information is available in the codebook, enabling reusability for any future research.

## (4) REUSE POTENTIAL

The data include a large variety of variables and constructs related to cognitive abilities, achievement, motivational characteristics, properties of schools, and classroom activities that can be used to investigate various research questions based on a large, representative sample.

The large student sample in particular allows for robust analyses of school achievement in small subpopulations such as refugees ( $n = 941$ ; see also Schipolowski et al., 2021) or students with special educational needs ( $n = 2.729$ ). In this regard, possible research questions include differences in achievement and motivational characteristics for SEN students in general education schools versus SEN schools or disparities in achievement for refugee students in comparison to other students with and without immigration background. Furthermore, research questions concerning different levels of the educational system (e.g., students versus classes versus schools) can be investigated using multilevel analyses. Finally, as key variables were previously assessed in the year 2012 (Lenski et al., 2016), the data allow for comparisons between two cohorts of 9<sup>th</sup> graders. Trend analyses may include changes in achievement, motivational characteristics, or classroom activities in grade 9 over a time period of 6 years.

Another unique aspect of the data is the assessment of social networks that can be used for in-depth analyses of social integration in different groups (e.g., based on school type, proficiency levels, gender, or social background). At the time of writing, the IQB Trends in Student Achievement 2018 provides a unique combination of social network data in conjunction with other constructs such as achievement for a very large sample that is not available in any other dataset. For example, research might address questions pertaining to which relations in social networks are associated with beneficial learning outcomes for specific groups such as students with an immigration background. This may help to identify conditions under which students with immigration background can profit from social support from their peers.

Aside from education-related variables, the data also include constructs commonly used in psychological research such as general language ability and reasoning. Therefore, the data have a large reuse potential not only for educational science, but also for psychological research.

Limitations to the reuse potential mainly refer to the high complexity of the data and the required statistical approaches. Due to the clustered sampling procedure resampling methods have to be used to compute unbiased standard error estimates (see section 2.4). If trend analyses are conducted using the 2012 and 2018 data sets, linking errors should be incorporated into standard error estimation (Sachse & Haag, 2017). Plausible values provide unbiased estimates for student achievement on group level but require pooling techniques (von Davier et al., 2009). However, most of these methodologies are readily available in standard statistical software, such as in the R package eatRep (Weirich, Hecht, & Becker, 2022), which was also used for the initial policy reporting at IQB. General advice on how to use large-scale assessment data can be, for example, found in Rutkowski et al. (2014).

## NOTE

- 1 Note that case numbers differ slightly from the reported numbers in Stanat et al. (2019). This is due to cases with zero weights being included in the full dataset reported here but not being used for the study report. These cases are nonetheless included in the dataset as they can be relevant for certain analyses that do not rely on achievement data.

## ACKNOWLEDGEMENTS

We acknowledge the work of the project managers in the federal states and at IEA Hamburg as well as our cooperation partners in mathematics and science education and the teachers who developed the assessment tasks.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Johanna Busse: Data Curation, Writing – Original Draft  
 Benjamin Becker: Software, Data Curation, Writing – Review Editing.  
 Sebastian Weirich: Methodology, Software, Data Curation, Writing – Review Editing.  
 Stefan Schipolowski: Project Administration, Data Curation, Writing – Review Editing.

Furthermore, we thank our colleagues at IQB for their assistance in conducting the study, data curation, and analyses, most notably Nicole Mahler and Marlen Holtmann. We also thank Aleksander Kocaj and Sara Techlin at the FDZ at IQB for thorough checks on the data and data anonymization.

## AUTHOR AFFILIATIONS

**Johanna Busse**

IQB Berlin, DE

**Benjamin Becker**  [orcid.org/0000-0003-3074-0918](https://orcid.org/0000-0003-3074-0918)

IQB Berlin, DE

**Sebastian Weirich**  [orcid.org/0000-0002-4493-4601](https://orcid.org/0000-0002-4493-4601)

IQB Berlin, DE

**Stefan Schipolowski**

IQB Berlin, DE

## REFERENCES

- Adams, R. J., & Wu, M. L.** (2007). The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models* (pp. 57–75). Springer. DOI: [https://doi.org/10.1007/978-0-387-49839-3\\_4](https://doi.org/10.1007/978-0-387-49839-3_4)
- Becker, B., Busse, J., Holtmann, M., Weirich, S., Schipolowski, S., Mahler, N., & Stanat, P.** (2022). *IQB-Bildungstrend 2018. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in Mathematik und den naturwissenschaftlichen Fächern*. Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen.
- Becker, B., Weirich, S., Mahler, N., & Sachse, K.** (2019). Testdesign und Auswertung des IQB-Bildungstrends 2018: Technische Grundlagen. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), *IQB-Bildungstrend 2018: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (pp. 411–426). Waxmann.
- Blossfeld, H.-P., Bos, W., Lenzen, D., Müller-Böling, D., Oelkers, J., Prenzel, M., & Wößmann, L.** (2007). Heterogenität und Bildungsstandards. In *Bildungsgerechtigkeit: Jahresgutachten 2007* (pp. 87–91). VS Verlag für Sozialwissenschaften. DOI: [https://doi.org/10.1007/978-3-531-90341-5\\_7](https://doi.org/10.1007/978-3-531-90341-5_7)
- Eckes, T.** (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing*, 34(3), 383–411. DOI: <https://doi.org/10.1177/0265532216672703>
- Embretson, S. E., & Reise, S. P.** (2000). *Item response theory for psychologists*. Erlbaum. DOI: <https://doi.org/10.1037/10519-153>
- Fischer, G. H.** (2006). Rasch Models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*, 26, pp. 515–585. Elsevier. DOI: [https://doi.org/10.1016/S0169-7161\(06\)26016-4](https://doi.org/10.1016/S0169-7161(06)26016-4)

- Gonzalez, E., & Rutkowski, L.** (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, 3, 125–156.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J.** (1991). *Fundamentals of item response theory*. Sage.
- Holland, P. W.** (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30). Springer. DOI: [https://doi.org/10.1007/978-0-387-49771-6\\_2](https://doi.org/10.1007/978-0-387-49771-6_2)
- Klein-Braley, C.** (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, 14(1), 47–84. DOI: <https://doi.org/10.1177/026553229701400104>
- KMK.** (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. Beschluss der 350. Kultusministerkonferenz vom 11.06.2015.* [https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen\\_beschluesse/2015/2015\\_06\\_11-Gesamtstrategie-Bildungsmonitoring.pdf](https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf)
- Kolen, M. J., & Brennan, R. L.** (2004). *Testing equating, scaling, and linking: Methods and practice*. Springer. DOI: <https://doi.org/10.1007/978-1-4757-4310-4>
- Lenski, A. E., Hecht, M., Penk, C., Milles, F., Mezger, M., Heitmann, P., Stanat, P., & Pant, H. A.** (2016). *IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. DOI: <https://doi.org/10.18452/3125>
- Lumley, T.** (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1–19. DOI: <https://doi.org/10.18637/jss.v009.i08>
- Lumley, T.** (2019). *survey: analysis of complex survey samples*. In (Version R package version 3.35-1).
- Mahler, N., Kölm, J., & Werner, B.** (2020). Entwicklung von Mathematiktestaufgaben für Schüler\*innen mit einem sonderpädagogischen Förderbedarf im Lernen – Konzeption und empirische Ergebnisse. In C. Gresch, P. Kuhl, M. Grosche, C. Sälzer, & P. Stanat (Eds.), *Schüler\*innen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen: Einblicke und Entwicklungen* (pp. 109–146). Springer Fachmedien Wiesbaden. DOI: [https://doi.org/10.1007/978-3-658-27608-9\\_5](https://doi.org/10.1007/978-3-658-27608-9_5)
- Mahler, N., Schipolowski, S., & Weirich, S.** (2019). Anlage, Durchführung und Auswertung des IQB-Bildungstrends 2018. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (pp. 99–130). Waxmann.
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M.** (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133–161. <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1992-41607-001&site=ehost-live> (The National Assessment of Educational Progress). DOI: <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>
- Pant, H. A., Stanat, P., Pöhlmann, C., & Böhme, K.** (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (pp. 13–22). Waxmann.
- Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C.** (Eds.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Waxmann.
- Robitzsch, A., Kiefer, T., & Wu, M.** (2022). *TAM: Test Analysis Modules*. In (Version R package version 4.1–4) <https://CRAN.R-project.org/package=TAM>
- Robitzsch, A., & Oberwimmer, K.** (2019). *BIFIEsurvey: Tools for survey statistics in educational assessment*. In (Version 3.3-12)
- Rutkowski, L., Von Davier, M., & Rutkowski, D.** (2014). Handbook of international large-scale assessment. Background, technical issues, and methods of data analysis. DOI: <https://doi.org/10.1201/b16061>
- Sachse, K., & Haag, N.** (2017). Standard Errors for National Trends in International Large-Scale Assessments in the Case of Cross-National Differential Item Functioning. *Applied Measurement in Education*, 30(2), 102–116. DOI: <https://doi.org/10.1080/08957347.2017.1283315>
- Schafer, J. L., & Graham, J. W.** (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. DOI: <https://doi.org/10.1037/1082-989X.7.2.147>
- Schipolowski, S., Edele, A., Mahler, N., & Stanat, P.** (2021). Mathematische und naturwissenschaftliche Kompetenzen von Jugendlichen mit Fluchtbiografie am Ende der Sekundarstufe I. *Journal for Educational Research Online*, 13(1), 78–104. DOI: <https://doi.org/10.31244/jero.2021.01.03>
- Schneider, R., Gentrup, S., Jansen, M., & Stanat, P.** (2022). Kohortentrends in schulfachbezogenen Selbstkonzepten und Interessen bei Mädchen und Jungen. *Zeitschrift für Pädagogische Psychologie*. DOI: <https://doi.org/10.1024/1010-0652/a000346>
- Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S.** (2019). *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich*. Waxmann.
- Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S.** (2020). *IQB Trends in Student Achievement 2018. The Second National Assessment of Mathematics and Science Proficiencies at the End of Ninth Grade*. Waxmann.
- van Buuren, S., Brands, J. P. L., Groothuis-Oudshoorn, K., & Rubin, D. B.** (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064. DOI: <https://doi.org/10.1080/10629360600810434>



- van Buuren, S., & Groothuis-Oudshoorn, K.** (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. DOI: <https://doi.org/10.18637/jss.v045.i03>
- von Davier, A., Carstensen, C. H., & von Davier, M.** (2008). Linking Competencies in Horizontal, Vertical, and Longitudinal Settings and Measuring Growth. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 121–149). Hogrefe & Huber.
- von Davier, M., Gonzalez, E., & Mislevy, R. J.** (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, 2, 9–36.
- Weirich, S., Hecht, M., & Becker, B.** (2022). *Educational Assessment Tools for Replication Methods*. In (Version R package version 0.14.5) <https://cran.r-project.org/web/packages/eatRep/index.html>
- Weirich, S., Hecht, M., Becker, B., & Zitzmann, S.** (2022). Comparing group means with the total mean in random samples, surveys, and large-scale assessments: A tutorial and software illustration. *Behavior Research Methods*, 54(3), 1051–1062. DOI: <https://doi.org/10.3758/s13428-021-01553-1>
- Westat.** (2000). *WesVar*. Westat.
- Wilhelm, O., Schroeders, U., & Schipolowski, S.** (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe*. Hogrefe.

---

**TO CITE THIS ARTICLE:**

Busse, J., Becker, B., Weirich, S., & Schipolowski, S. (2022). IQB Trends in Student Achievement 2018: A Large-Scale Educational Assessment Study in Germany. *Journal of Open Psychology Data*, 10: 18, pp. 1–9. DOI: <https://doi.org/10.5334/jopd.71>

**Published:** 12 December 2022

**COPYRIGHT:**

© 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Journal of Open Psychology Data* is a peer-reviewed open access journal published by Ubiquity Press.