Mannheim inventory for testing judgment accuracy – MITU

Karina Karst, Meike Bonefeld & Jan Siebert

Note: This paper is written in German. If a German version of this paper is required, please contact the author via meike.bonefeld@ezw.uni-freiburg.de

Goal

The Mannheim Inventory for the Testing of Judgment Accuracy (MITU) is a virtual classroom. Within this virtual environment, judgments by teachers or prospective teachers about students are to be recorded in an approximation of a real classroom setting.

The characteristics of the students included in the current form of the virtual classroom (N = 12) are experimentally varied in order to be able to determine effects of different student characteristics (performance level, gender, migration background) on judgment formation and on the accuracy of the judgment.

In addition, the inventory can also be used to experimentally vary characteristics of the judge and/or characteristics of the judgment as determinants of judgment accuracy.

Potential extension

The student descriptions explained below can be arbitrarily expanded to include factors that can be assumed to (falsely) influence teacher judgment.

The school class [classroom]

The virtual school class consists of twelve students (6 boys, 3 of them with migration background; 6 girls, 2 of them with migration background, all varying in their achievement level). The names of the students presented in the virtual school class are fictitious and piloted for various characteristics in pre-tests.

The information that the teachers receive about the students, apart from their real names, test scores from a 15-minute test on multiplication that 2nd grade students completed immediately after a videotaped double lesson on "Introduction to Multiplication." The test collection as well as the videography of the lesson were part of the larger elementary school project PERLE (Persönlichkeits- und Lernentwicklung von Grundschulkindern; Lipowsky, Faust & Kastens, 2013).

After the introductory, the focus of the test on multiplication lesson was on the conceptual understanding of multiplication (e.g., switching between addition and addition and

multiplication; simple multiplication tasks), which is important precursor skill for multiplication tasks.

Eight days after the introductory lesson, the students were tested again on the multiplication topic.

The 12 students represented in the virtual classroom were randomly selected from 579 students who participated in both days of testing. It was noted that the 12 students were from the same class. The 12 students are roughly evenly distributed across the achievement domains.

The assignment of the names presented in the virtual classroom to the respective test booklets was random and fictitious names were chosen for this purpose. The names used were pretested in a pretest for a clear assignment to a gender and migration background.

The two-window view

First, the subjects are presented with a brief explanation of the following two-window view, the core of the Mannheim Inventory for testing judgment accuracy (cf. Fig.1)

Figure 1: Explanation of the two-window-view

Translation of figure 1:

On the next page, two windows will open. There is the window "classroom" (green/left) and the window "student assessment" (gray/right).

In the green window "classroom" we show you a virtual school class with 12 students at the beginning of the second term of the 2nd grade. Your task will be to view the different information about each student. This information will appear by clicking on the school's name. The information is test results from a 15-minute multiplication test that students took immediately after a double lesson on "Introduction to multiplication. Students were not allowed to use other aids to solve the tasks.

In the next step, please assess in the gray (right) window how the students in the virtual class would perform on a different test. This is a test on multiplication that students worked on eight days after the test, which will be shown in the green window. Please give your assessment for all 12 students.
Even if you have limited information available, we ask that you assess students based on this information when they take a similar mathematics test.

In the next step, two windows (the actual two-window view) appear on the subjects' screens. (The actual two-window view), one on the left and one on the right (cf. Fig. 4-Fig.10).
The test subjects see both windows throughout and can therefore refer to the student information during every information during each judgment.

The student information

In the left window "classroom" there is a view of a virtual school class with the 12 students throughout the judgment. The first view contains the class overview with the names: Murat, Emre, Farid, Max, Lukas, Jonas, Elif, Seda, Julia, Anne, Leonie and Anna (cf. Fig.2). The names were pretested as part of a study and can be clearly assigned to a nationality and gender.

The test subjects are informed that the students are second grade students. Furthermore, the test subjects are informed that they can obtain further information on the respective student by clicking on the name.

By clicking on the name, a scan of the test booklet (22 items), the total score achieved by the student and the potentially achievable points (cf. Fig. 3) are then displayed.

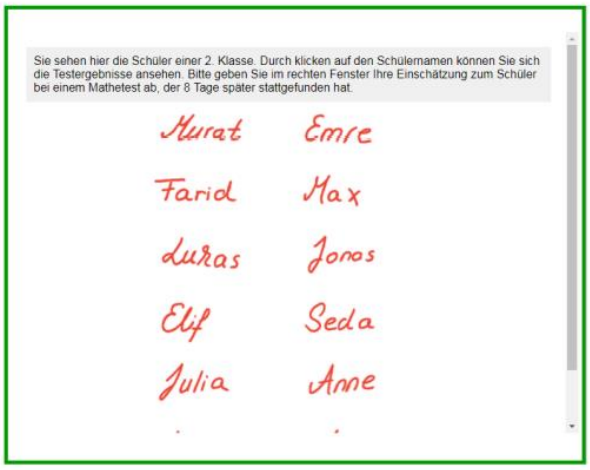

Figure 2: Classroom

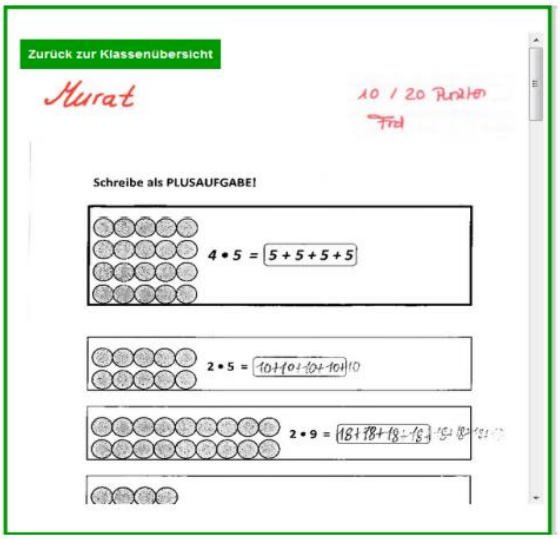

Figure 3: Example of the test booklet

Translation of figure 2:
You see here the students of a 2nd-grade class. By clicking on the student's name, you can see the test results. In the window on the right, please provide your assessment of the student on a math test that took place 8 days later.

Translation of figure 3:
Back to the class overview (headline)
Write as plus task (task)
10/20 points (grading)

The students vary in their performance level. Five performance levels are distinguished (very good, good, average, less good, poor) (see table 1). These performance levels are not explicitly stated to the test subjects but can be deduced from the performance in the test booklets. The empirical determination of the performance levels was based on quintile analyses across the entire sample.

Name	Geschlecht	Migrations-hintergrund	Leistungsniveau[1]
Murat	♂	🇹🇷	●●○○○
Emre	♂	🇹🇷	●●●○○
Farid	♂	🇹🇷	●●●●○
Max	♂	🇩🇪	●●●●●
Lukas	♂	🇩🇪	●●○○○
Jonas	♂	🇩🇪	●●○○○
Elif	♀	🇹🇷	●●●●●
Seda	♀	🇹🇷	●○○○○
Julia	♀	🇩🇪	●●●●○
Anne	♀	🇩🇪	●○○○○
Leonie	♀	🇩🇪	●●●●●
Anna	♀	🇩🇪	●●●○○

[1] ●○○○○=niedriges Leistungsniveau ●●●●●=hohes Leistungsniveau

Table 1: Student variations

Translation of table 1:
Name, gender, migration background, performance level (grids)
Footnote: Low performance level (1/5 circles green), high performance level (5/5 circles green)

The judgments

The judgments to be made (global/specific/overall) are located in the right window of the subjects' screen.

Both the specific judgments and the global judgments represent predictions. Here, subjects are asked to estimate the students' test performance on a test of multiplication 8 days later.

The overall judgment refers to an assessment of the students' current multiplication skills based on the present test booklet in the left window.

Specific judgment:

In the specific judgment format, the subjects see 11 multiplication tasks from a) to k) in the right window (e.g., 2x4= __; cf. Fig. 4) with the request to indicate for each student which of these 11 multiplication tasks the student can solve. Here, the subjects can select for the students individually and for each task separately in a drop-down menu whether the student will indicate the correct or incorrect solution (dichotomous response format) (cf. Fig. 5).

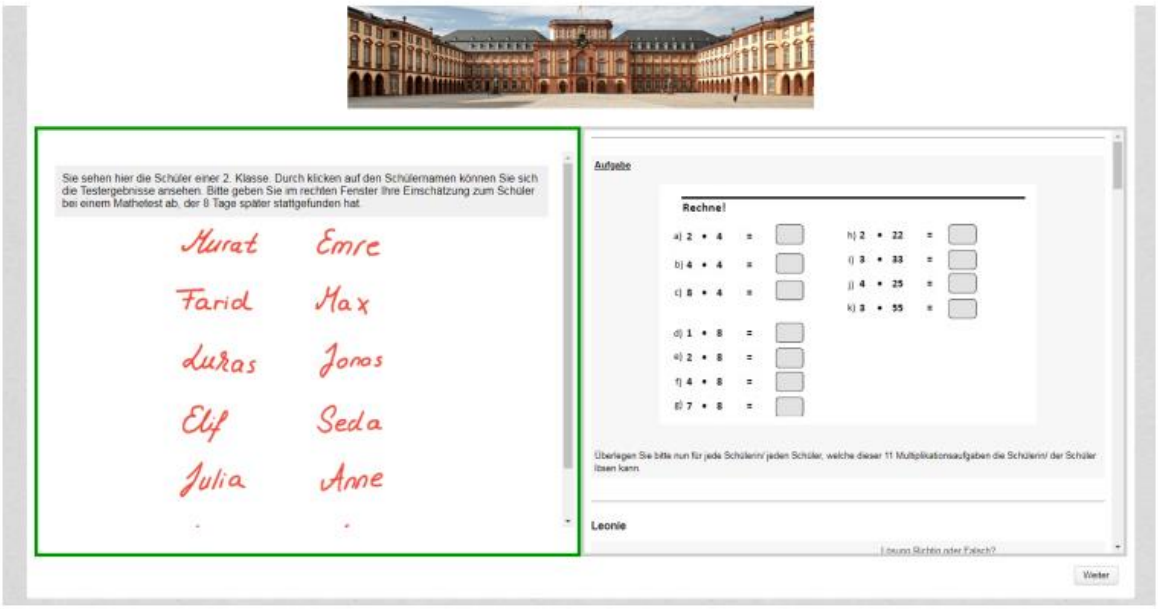

Figure 4: Specific judgment – task

Translation of figure 4:

You see here the students of a 2nd-grade class. You can see the test results by clicking on the student's name. In the window on the right, please provide your assessment of the student on a math test that took place eight days later

Concerning the picture on the right:

Exercise (title)

Now, for each student, please consider which of these 11 multiplication problems the student can solve. (task)

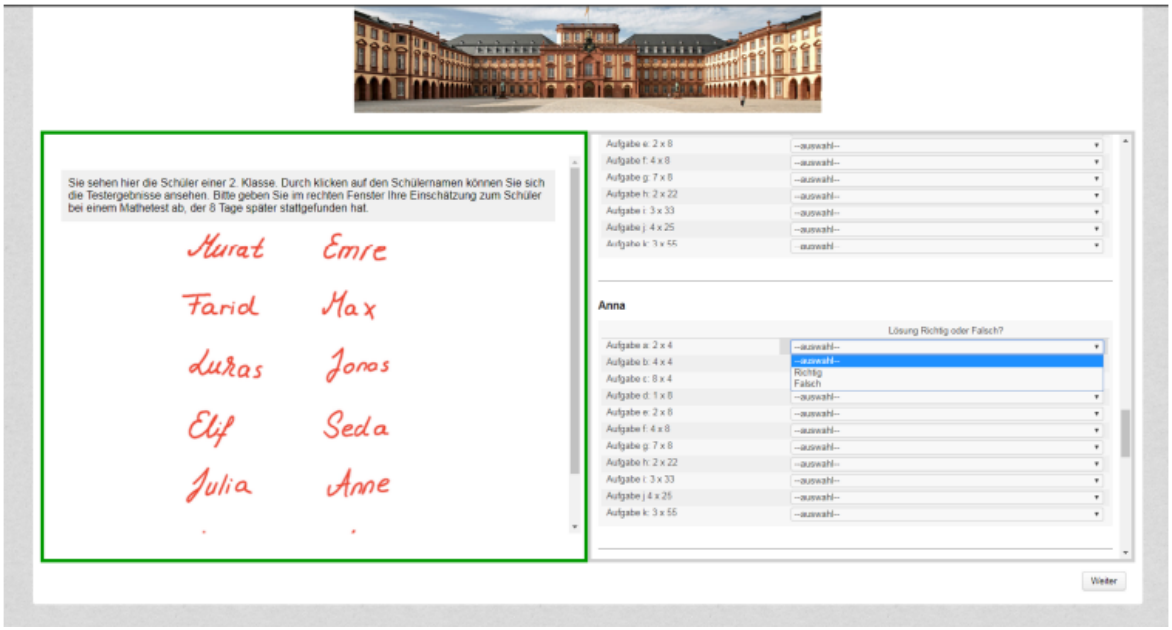

Figure 5: Specific judgment – judgment delivery

Translation of figure 5:

You see here the students of a 2nd-grade class. You can see the test results by clicking on the student's name.In the window on the right, please provide your assessment of the student on a math test that took place 8 days later. (text on the left)

Drop-down-menu (on the right): Right or wrong?

Global Judgment:

The global judgment format is divided into three different judgments, which in turn refer to two different task types. For the first two global judgments (g1 and g2), the subject sees two identical columns, each with the result (24 or 40) of a multiplication task in the roof and a prompt for the students to form a painting task based on the result in the roof. The task also includes a sample solution for each task. The subject is asked to estimate for each student in an open response format how many of the 7 possible correct tasks for the result 24 or 40 are named by the student (cf. figures 6 and 7).

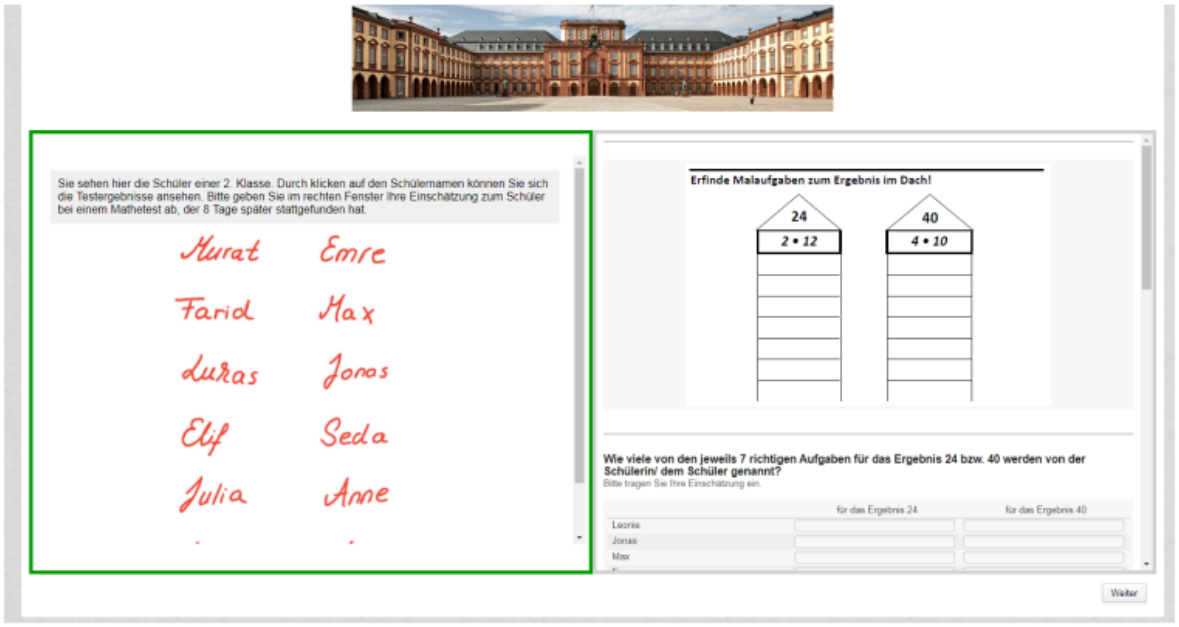

Figure 6: Global judgment (g1, g2) – task

Translation of figure 6:

Text on the left is the same as in figure 5

Task on the right: Invent painting tasks to result in the roof. How many of each of the seven correct tasks for the result 24 or 40 are named by the student?

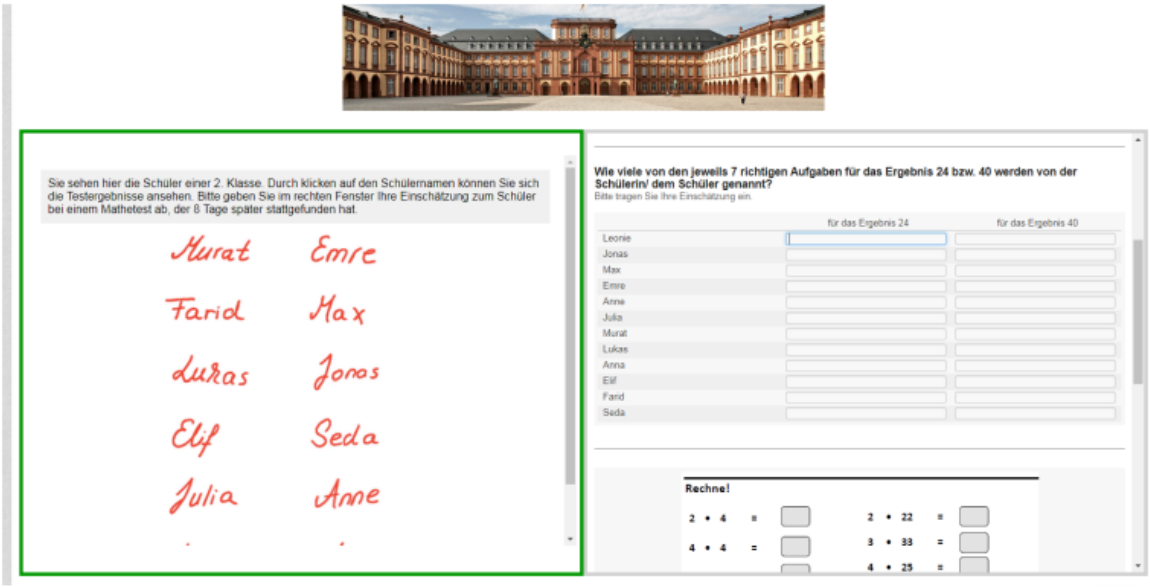

Figure 7: Global judgment (g1, g2) – judgment delivery

Translation of figure 7:

Text in the left is the same as in figure 5.

Task on the right: How many of the 7 correct exercises for the results 24 or 40 are mentioned by the student?

Table grids: names, for the result 24, for the result 40
Exercise on the bottom of the page: Calculate! (instruction)

In addition, the subjects are asked to make a further global judgment (g3). Here, the subjects see 11 multiplication tasks with the request to, again, estimate in an open field how many of these 11 multiplication tasks the student will solve (cf. Fig. 8 and Fig. 9).

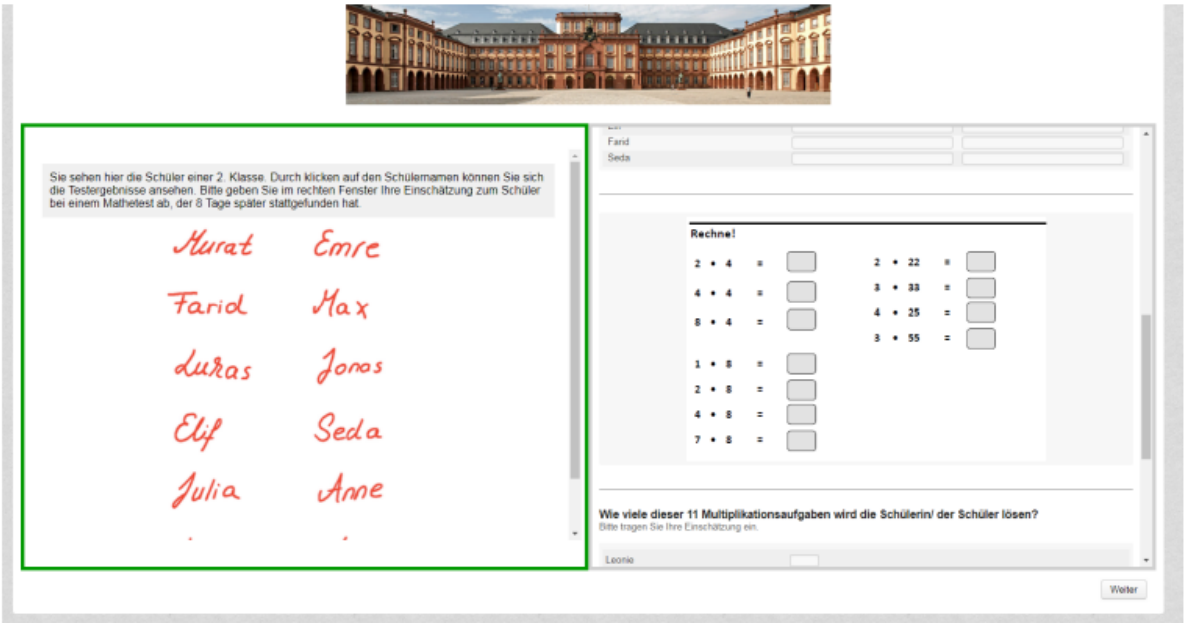

Figure 8: Global judgment (g3) – task

Translation of figure 8:
Text on the left is the same as in figure 5.
Tasks on the right: Calculate! (instruction)
How many of the 11 multiplication exercises will be solved by the student? (exercise)

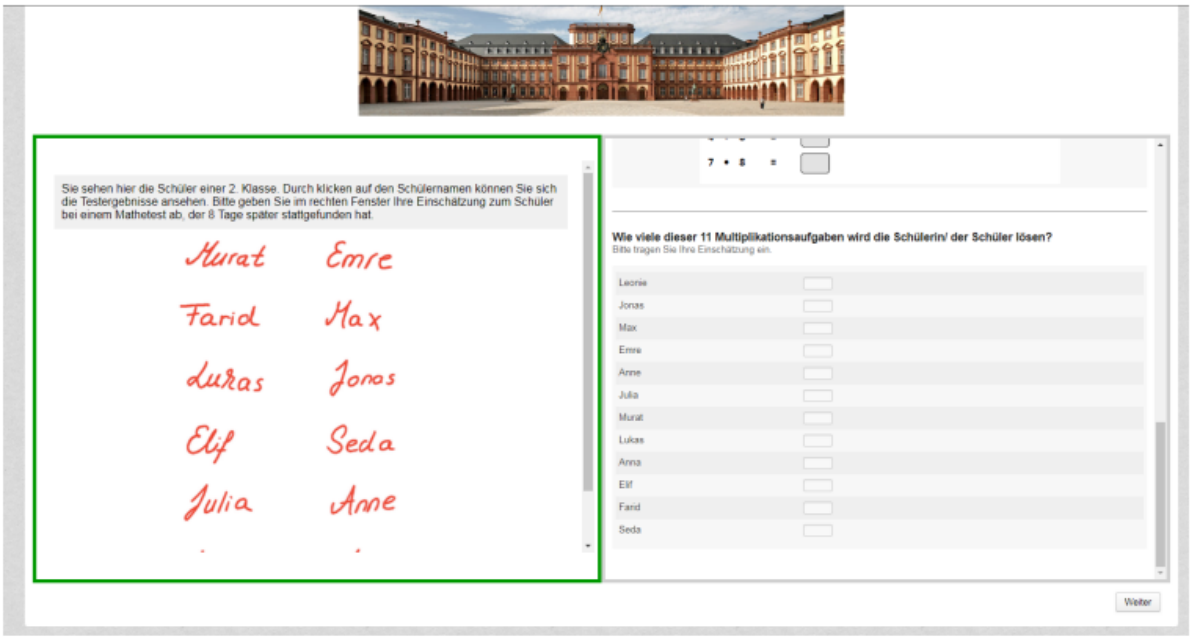

Figure 9: Global judgment (g3) – judgment delivery

Translation of figure 9:
Text on the left is the same as in figure 5.
Tasks on the right: How many of the 11 multiplication exercises will be solved by the student? (exercise) and names of the students is displayed.

Overall judgment:

This judgment format is an assessment of the current performance level of the students. On the right side of the two-pane view in this case is a five-point scale with the scale endpoints "poor" and "very good". This is used to assess the performance level of all students with the instruction: "Overall, how would you rate the student's performance level in multiplication?" (Cf. Fig. 10). Once again, the test subjects have the opportunity to give their assessment separately for each student.

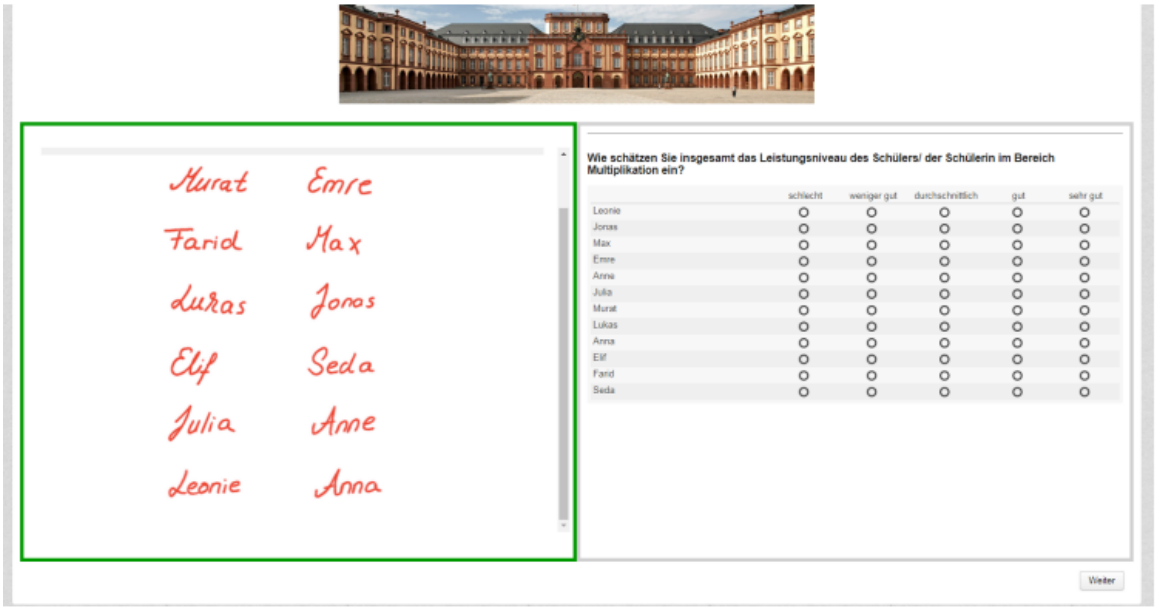

Figure 10: Overall judgment

Translation of figure 10:
Question on the right: How do you estimate the performance level of each student concerning their multiplication skills?

Criteria values

The overall judgment refers to the assessment of the current student performance. Here, the actual test performance of the students at t1 (test booklet) serves as the criterion value.

The specific judgment and the global judgment refer to the performance of the students 8 days later. These judgments are therefore predictions of the test subjects. Here there are two different possibilities, which form of criterion values are used. First, as expected, the actual student performance/test score can be used as criterion value (possibility 1). In addition, however, there is also the possibility (2) to estimate the expected test performances based on the total sample (n = 579) as unstandardized expected values within the framework of a linear regression (with the total test performance at t1 as uV and the performance at t2 as aV). Because the present data are student performances of actual students who continued to attend school during the 8 days between the two measurement time points, this estimation of expected values is useful for assessing subjects' prediction independent of events that cannot be anticipated by the subjects.

Generated data

With the help of the virtual classroom, the answers of the test subjects can be saved (with the help of a data collection software, e.g. Unipark).

Furthermore, it is possible to collect process data with the virtual classroom (duration, click frequency on students). These process data are saved in the so-called "pupillog":

For the realization of the virtual classroom an individual software named "pupillog" was developed and operated on the technical infrastructure of the University of Mannheim. This software can be integrated into any other web-based survey software. For the previous surveys Unipark was used as survey software.

Each navigation of the test person in the virtual classroom is stored in the form of a data point. At the beginning, the test person sees the names of all students. The test person is able to navigate by clicking on the name of a student. This action shows the subject detailed information about the selected student. From this detailed view, the subject can only navigate back to the overview with all student names. A direct change between students is not possible. A data point now consists of three values: on the one hand, it is stored which subject triggered the navigation. This is the only way to later assign the data collected in another system (e.g. Unipark) to the same subject. Furthermore, a number that can be uniquely assigned to a student is stored to indicate which student the respondent is viewing in the detailed view. Navigating to the overview page with all names is also coded with a fixed number. As a third value, the exact time is recorded for each navigation. If all the data points for a test person are viewed after the end of the study, it is also possible to determine how long a test person stayed on the individual pages by forming differences with the data collected.

In conclusion, the virtual classroom makes it possible to record the time spent by the test subjects on the available information and the click behavior of test subjects and students.

Publications based on MITU:

Bonefeld, M., Dickhäuser, O. & Karst, K. (2019). Do Preservice Teachers' Judgments and Judgment Accuracy depend on Students' Characteristics? The Effect of Gender and Immigration Background. Social Psychology of Education. Online First. doi:10.1007/s11218-019-09533-2.

Karst, K., Bonefeld, M. & Siebert, J. (2018). Prozessdaten der Urteilsbildung und die Akkuratheit von Lehrkrafturteilen. Vortrag präsentiert auf dem 51. Kongress der Deutschen Gesellschaft für Psychologie (DGPS), Frankfurt am Main.

Karst K., Bonefeld, M. & Dickhäuser, O. (2018). Judgment accuracy of pre-service teachers. Explained by pupil characteristics and attributes of the judge. Vortrag präsentiert auf der Conference: Cultural diversity, migration and education (CDME), Potsdam.

Bonefeld, M., Karst, K. & Dickhäuser, O. (2018). Urteile über Schüler_innen mit Migrationshintergrund. Zur Bedeutung von Diagnose und Prognose. Vortrag präsentiert bei der 6. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Basel.

Karst, K., Bonefeld, M., Dickhäuser, O. (2017). Need for Cognition und die Akkuratheit von Urteilen angehender Lehrkräfte. Symposiumsbeitrag präsentiert auf der Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Tübingen.

Bonefeld, M., Karst, K. & Dickhäuser, O. (2017). Leistungserwartungen von Lehrenden: Einfluss von Geschlecht und Migrationshintergrund der Lernenden. Symposiumsbeitrag präsentiert auf der Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Tübingen.

Bonefeld, M., Karst, K. & Dickhäuser, O. (2017). Migrationshintergrund von Lernenden und Leistungserwartungen von Lehrkräften: Zur Bedeutung des Leistungsniveaus. Vortrag auf der gemeinsamen Tagung der Fachgruppen Entwicklungspsychologie und Pädagogische Psychologie (PAEPSY), Münster.

Literature

Lipowsky, F., Faust, G. & Kastens, C. (2013) (Hrsg.). Persönlichkeits- und Lernentwicklung an staatlichen und privaten Grundschulen. Ergebnisse der PERLE-Studie zu den ersten beiden Schuljahren. Münster: Waxmann.