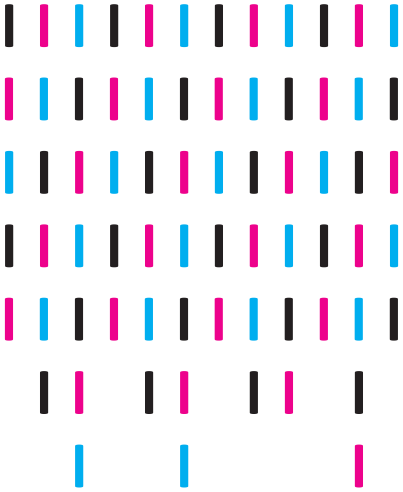

NEPS SURVEY PAPERS

Sebastian E. Wenz, Melanie Olczyk, and Georg Lorenz

# MEASURING TEACHERS' STEREOTYPES IN THE NEPS

NEPS Survey Paper No. 3
Bamberg, May 2016

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

**Survey Papers of the German National Educational Panel Study (NEPS)**
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

**The NEPS Survey Papers are available at** https://www.neps-data.de (see section “Publications“).

**Editor-in-Chief**: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

**Contact**: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

# Measuring teachers' stereotypes in the NEPS[1]

*Sebastian E. Wenz, GESIS – Leibniz Institute for the Social Sciences, Cologne*

*Melanie Olczyk, Bamberg Graduate School of Social Sciences (BAGSS), Bamberg*

*Georg Lorenz, Berlin Institute for Integration and Migration Research (BIM) at the Humboldt University of Berlin*

**E-mail address of lead author:**

sebastian.wenz@gesis.org

**Bibliographic data:**

Wenz, Sebastian E., Olczyk, M., & Lorenz, G. (2016). *Measuring teachers' stereotypes in the NEPS* (NEPS Survey Paper No. 3). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

[1] We would like to thank Corinna Kleinert, Jasper Dag Tjaden, Marion Fischer-Neumann, Steffen Schindler, and the participants of the BAGSS weekly colloquium for valuable feedback on earlier versions of this paper.

# Measuring teachers' stereotypes in the NEPS

**Abstract**

German data on the processes underlying discrimination is still sparse. Against this backdrop, this working paper discusses the measurement of a major source of discrimination: Stereotypes. Conceptualizing stereotypes as beliefs or sets of beliefs about the characteristics, attributes, or behaviors of a particular group of people, we introduce an explicit measure of teachers' stereotypes at the National Educational Panel Study (NEPS). Teachers are asked to estimate the average competencies of girls, boys, students of lower, middle, and upper class background, students of Turkish and Russian origin as well as immigrant students and ethnic majority students. Quantitative analyses based on a sample of 52 second-grade teachers show both a large variation in the expressed stereotypes between groups of students and between teachers. Furthermore, the analyses suggest that teachers' stereotypes are quite accurate overall in that they reflect group differences in achievement as reported in the recent literature. However, we also find biases to the disadvantage of boys, immigrants in general, as well as immigrants of Turkish and Russian origin in particular. We argue that these results and evidence from cognitive interviews speak for the validity of the instrument.

**Keywords**

teacher beliefs, prejudice, stereotypes, discrimination, cognitive interviews, mixed methods

## Introduction

International comparative studies have repeatedly shown that Germany features relatively high levels of inequality with regard to various dimensions of educational success—e.g., competencies, grades, degrees completed—along socioeconomic status or social class as well as ethnicity (Shavit & Blossfeld, 1993; Marks, 2005a; Marks, 2005b). Additionally, students' gender significantly predicts these outcomes (Bos et al., 2012a, 2012b; Breen et al., 2010; Prenzel et al., 2013). One explanation for these achievement gaps that has received only little attention by quantitative researchers is discrimination by teachers and its underlying mechanisms. In this paper we focus on one of the most prominent mechanisms leading to discriminatory judgments and behavior, namely stereotypes. We introduce an explicit measure of teachers' stereotypes about average competencies of different groups of students developed at the National Educational Panel Study (NEPS), the largest panel study in German education (Blossfeld et al., 2011).

We proceed as follows: In section 1 we give a brief overview of the role of stereotypes in human cognition in general and in education in particular. We also summarize previous research on discrimination in German education and stereotypes held by German teachers. In section 2 we elaborate on how to conceptualize stereotypes. In section 3 we discuss different ways of measuring stereotypes. We describe the development of an item battery on teachers' stereotypes about group-specific performances in assessments in section 4, where we present qualitative evidence for the validity of the measure. In section 5 we briefly describe the quantitative data from the NEPS pilot study as well as our analytical strategy to assess various desirable properties of the measure. Empirical results will be presented in section 6. We conclude by briefly summarizing and discussing both the development process of the item battery and first results.

## 1. Why Study Teachers' Stereotypes?

### 1.1 Stereotypes in human cognition

In the more recent social psychological literature stereotypes are usually conceptualized as "beliefs about the characteristics, attributes, and behaviors of members of certain groups" (Hilton & von Hippel, 1996: 240). As such, stereotypes have been shown to serve particular functions in perceiving, storing, and retrieving information in numerous studies. It has been shown that stereotypes and their use in interpersonal interactions are connected to a largely inevitable and automatic process of categorizing people on the basis of biological and social cues (Fiske et al., 1999). People tend to seek (Darley & Gross, 1983; Snyder & Swann, 1978), encode (Bodenhausen & Lichtenstein, 1987), recall (Bodenhausen & Lichtenstein, 1987), and interpret (Darley & Gross, 1983) information in a stereotype-consistent way. Eventually, stereotypes may influence the way people judge and treat other people and, therefore, lead to discrimination to the disadvantage of individuals and—under certain conditions (Aigner & Cain, 1977; England & Lewin 1989)—to the disadvantage of a group of people as a whole. In this way, stereotypes may lead to inequality between different social and ethnic groups.

## 1.2 Teachers' stereotypes as mechanism of discrimination in German education

Most of the few studies that explicitly investigate discrimination by teachers at German schools follow a residual approach, where teachers' judgments such as grades, expectations, or recommendations are regressed on variables that identify different societal groups and covariates that are intended to account for group differences in merit. Conditional—or residual—group differences are then interpreted as estimates of discrimination (Oaxaca, 1973). Using this residual approach to discrimination, some studies provide evidence for discrimination in Germany's education system against ethnic minorities (Arnold et al., 2007; Kiss, 2013), while others do not find any residual group differences (Kristen, 2006; Schneider, 2011). Also mixed is the evidence for whether or not teachers discriminate on the basis of gender—some studies find significant but usually rather small disadvantages for boys (Arnold et al., 2007; Ditton et al., 2005; Kiss, 2013), while others find no differences (Stubbe et al., 2012; Schneider, 2011). The strongest evidence exists for discrimination against children from families with low socioeconomic status (Arnold et al., 2007; Ditton et al., 2005; Schneider, 2011). Yet, none of these studies directly or indirectly tested for different mechanisms of discrimination. As a consequence, it remains unclear whether discriminatory behavior is the result of the—conscious or unconscious—activation and application of stereotypes or the consequence of other mechanisms.

At the same time, stereotypes—whether biased or unbiased—may lead to different forms of discriminatory treatment of which rather subtle forms such as self-fulfilling prophecies cannot be identified without measures of stereotypes (Schofield, 2006; Jussim et al., 2009). Studies using experimental setups find only weak evidence for discriminatory grading practices, but stronger evidence for negatively biased expectations towards Turkish students (Sprietsma, 2013; Wenz & Hoenig, 2013; Lorenz et al., 2016), children from families of lower social class background (Wenz & Hoenig, 2013; Lorenz et al., 2016) and partly also towards male students (Lorenz et al., 2016). These results point to—possibly biased—stereotypes as the source of discriminatory judgments and, eventually, treatment. Discriminatory treatment might lead to worse track recommendations (Sprietsma, 2013; Wenz & Hoenig, 2013) but also to worse competence development (Lorenz et al., 2016).

However, to date, no quantitative study on German education has measured teachers' stereotypes about different groups of students and, hence, teachers' stereotypes have not been related to teachers' evaluations of students' performance or achievement.

In this paper we report first steps undertaken within the framework of the German National Educational Panel Study (NEPS) to fill this gap (Blossfeld et al., 2011). We present the development process of an item battery that measures teachers' stereotypes about the average competencies in math and reading of different social and ethnic groups.

## 2. Conceptualizing stereotypes

Research on stereotypes has a long history: From Lippmann's (1922) „pictures in our heads" metaphor until today's multifaceted perspectives on the term, definitions of stereotypes abound. In line with many contributions in the social psychological literature (see, e.g., Ashmore & Del Boca, 1981: 16; Ehrlich, 1973: 20; Hilton & von Hippel, 1996: 240; Schneider 2004: 24), we define a stereotype as *a belief or a set of beliefs about the characteristics,*

*attributes, or behaviors of a particular group or category of people*. Put differently, a stereotype is a cognitive structure that links knowledge to a category of people (Bless et al., 2004: 53; Macrae & Bodenhausen, 2000). Below we clarify our conceptualization further by saying a little more about what our definition implies and—maybe even more importantly—what it does *not* imply.

**True or false?**

One of the oldest debates around the term has been concerned with the question of whether stereotypes should be conceptualized as incorrect per se. Allport (1954), for example, suggests that a stereotype is "an exaggerated belief associated with a category" (191) and, hence, rules out by definition that a stereotype can be "a valid generalization" (192). In contrast, some 20 years later, Ehrlich (1973: 20) was much less restrictive and allowed stereotypes to also be correct, in referring to stereotypes as "a set of beliefs and disbeliefs about any group of people". Similarly, Schneider (2004: 24) refers to stereotypes as "qualities perceived to be associated with particular groups or categories of people". Over time it has become the "standard viewpoint" (Hilton & von Hippel, 1996: 240) to allow stereotypes to contain more or less accurate beliefs—exactly right or completely wrong or anything in between. Our conceptualization is consistent with this standard viewpoint.

**Individual or cultural?**

Furthermore, different forms of stereotypes have been discussed. One important distinction separates *personal* or *individual* from *cultural stereotypes* (Ashmore & Del Boca, 1979, 1981; Gardner, 1973). Ashmore & Del Boca, for instance, have suggested "that the term 'stereotype' should be reserved for the set of beliefs held by an individual regarding a social group and that the term 'cultural stereotype' should be used to describe shared or community-wide patterns of beliefs" (Ashmore & Del Boca, 1981: 19). Some scholars, especially early ones, have argued that a cultural consensus about particular attributes of a group of people is a necessary condition for a belief to be called a stereotype (Gardner, 1973). Others disagree and simply acknowledge that individual beliefs—and, thus, stereotypes—can but do not have to be shared by others and that widely shared beliefs are also known by those who do not endorse them (Devine, 1989: 5). However, knowing about cultural stereotypes might be enough to build implicit associations that are different from explicit beliefs in that they are hard to control, automatic constructs.

Even though we are not interested in stereotypes held by single individuals—i.e. single teachers—we follow the logic of methodological individualism and aim at measuring individual stereotypes. We leave it to the researcher whether and how to aggregate them—be it in a statistical model or by defining a criterion for a cultural consensus among teachers.

**Explicit or implicit?**

In recent years the distinction between explicit beliefs and implicit associations has been the most important and most debated in research on stereotypes and attitudes (see, e.g., Fazio & Olson, 2003; Greenwald & Banaji, 1995). While some scholars suggest a distinction between implicit and explicit stereotypes (e.g., Greenwald & Banaji, 1995), others distinguish between explicit and implicit measures of stereotypes, attitudes, and the like (Fazio & Olson,

2003: 302-303). We take the latter position and will elaborate on this distinction below in section 3.

**Stereotypes and related constructs**

For one thing, a stereotype is not the same as stereotyp*ing*, by which we—in line with the literature—mean *the process of applying a stereotype in any judgment or treatment of a particular group of people or of an individual* (Hilton & von Hippel, 1996; Macrae & Bodenhausen, 2000).

Also, stereotypes are distinct from prejudice and discrimination. Of course, numerous social psychological, sociological, and economic theories (see, e.g., Aigner & Cain, 1977; Becker, 1971; Fiske at al., 1999, and Fiske et al., 2002 for particular theories, and see, e.g., England & Lewin, 1989; Pager & Shepherd, 2008 and Reskin, 2003 for reviews) suggest what empirical studies have shown: Stereotypes, prejudice, and discrimination are empirically related (see, e.g., Schütz & Six, 1996 and Talaska et al., 2008 for meta-analytic evidence). In fact, this is the major reason for us to measure teachers' stereotypes. Eventually, these measures are intended to predict students' outcomes. However, stereotypes, prejudice, and discrimination have been conceptualized as different constructs.

By the term prejudice we refer to *an attitude toward a particular group or category of people* (see, e.g., Correll et al., 2010; Ehrlich, 1973; Schneider, 2004). As an attitude, prejudice contains "general evaluations of people, objects, and issues" (Fazio & Petty, 2008: 1). In contrast, remember that stereotypes are typically defined as beliefs and, hence, lack any evaluative component. Thus, other than stereotypes, prejudices are neither false nor true, or inaccurate nor accurate, respectively. However, stereotypes—sometimes referred to as the cognitive component of prejudice (cf. Dovidio et al., 2010: 5; Fiske, 1998: 357)—may serve as justifications for prejudice (Crandall et al., 2011).

We define discrimination as an *unequal treatment of individuals because and only because they are member of a particular group or category of people.*[2] This definition highlights the most important difference between discrimination and stereotypes or prejudice: Discrimination is about *behavior*, whereas stereotypes are about perception and prejudice is about evaluation. This distinction is also known as tripartite conceptualization of category based reactions with stereotypes as cognitive, prejudice as affective, and discrimination as behavioral component, respectively (Fiske, 1998: 357). Note, however, that discrimination is not simply the behavioral manifestation of prejudice or stereotypes and, hence, neither equals stereotyping nor applied prejudice (cf. Jones, 1997). People with negative prejudices or negative stereotypes do not necessarily engage in discriminatory behavior toward members of the target group—for example on rational grounds or because they follow norms (LaPiere, 1934; Merton, 1949). On the other hand, these mechanisms may also cause people to discriminate against members of a particular group even though they do *not* hold negative stereotypes about or have negative prejudices against the same group (see Merton, 1949, for a classic typology on this).

[2] Note that this excludes some forms of discrimination that have been described in the literature. A deeper discussion of these different forms is clearly beyond the scope of this paper.

## 3. How to measure stereotypes[3]

Probably the most important distinction between various measures of stereotypes is the one between explicit measures, or measures of explicit *beliefs,* and implicit measures, or measures of implicit *associations*. Implicit measures of stereotypes and attitudes are relatively recent tools that have created a lot of attention among social scientists. These measures—e.g., priming methods or the implicit association test (IAT) and related tasks—"rely on processes that are uncontrolled, unintentional, autonomous, goal-independent, purely-stimulus-driven, unconscious, efficient, or fast" (De Houwer & Moors, 2007: 192)—or at least "less controllable by respondents" than explicit measures (Fazio & Olson, 2003: 636; Gawronski, 2009). Therefore, it has been suggested that implicit measures have two major advantages over explicit measures: Firstly, implicit measures should be less sensitive to social desirability bias (Fazio et al., 1995: 1022; Greenwald et al., 1998: 1465). In an "era of contested prejudice" (Lucas, 2008), respondents might shy away from honestly reporting their stereotypes to not violate personal or societal norms. Secondly, implicit measures could allow for a more accurate measure of stereotypes, since respondents might lack introspective access to implicitly stored associations and, hence, would simply be unable to accurately report all aspects of a stereotype explicitly (Hofmann & Wilson, 2010).

Not only have these supposed advantages been called into question (Gawronski et al., 2007, 2009), there are also at least two major advantages of explicit measures that make them our method of choice to assess teachers' stereotypes in NEPS: Firstly, they are very easy to implement, as the researcher only has to ask one or more questions and the respondent answers in more or less closed form. Secondly, explicit measures usually can be implemented in a paper-pencil survey questionnaire like they are used in NEPS and filled in by the respondents without assistance[4]. Hence, they do not require additional data collection and are, thus, more cost effective in a large scale survey such as the NEPS.

Explicit measures of stereotypes have a long history in social science research. Katz & Braly (1933) measured stereotypes by using an *adjective checklist*. Such a method asks the respondents to select those adjectives they consider to be typical of a particular group of people. The adjective checklist yields estimates of socially shared stereotype contents in the aggregate presumably due to both prevalence and extremity of individual stereotypes. However, on the individual level these measures are less useful, as differences between groups on a particular dimension—e.g., intelligence—cannot be quantified beyond the dichotomy 'mentioned-not mentioned'.

*Percentage estimates* or *scale ratings* are usually used to assess the prevalence of a stereotype. Percentage estimates ask the respondent to estimate the proportion of people from a social group that is characterized by an attribute or engages in a particular behavior (see, e.g., Park & Rothbart, 1982). In scale ratings respondents either rate the likelihood or how typical it is that a member of a social group features a particular attribute or engages in a particular behavior. Brigham (1971) used percentage ratings to assess how prevalent respondents believe a particular characteristic or behavior is among a particular group of people and to quantify the deviation of individuals from the average respondent in the

[3] This section closely follows the overviews in Correll et al. (2010) and Schneider (2004).

[4] While there are paper pencil versions of the IAT and other implicit measures (see, e.g., Vargas et al., 2007 for a review), they are all still much more complex than explicit closed-ended questions, where one item can be enough to assess the stereotype dimension of interest.

sample. This way Brigham (1971) seeks to identify unjustified generalizations, precisely what he defines as a stereotype.

The *stereotype differential technique* (Gardner, 1973) builds on the methodology of the semantic differential (Osgood et al., 1957) to assess respondents' stereotypes. Respondents rate social groups on a bipolar scale – usually a 7-point scale – with endpoints labeled with opposing adjectives or traits. Socially shared or cultural stereotypes are defined through a significant deviation of the sample mean from the scale's midpoint and through the standard deviation in the sample, where a smaller variation means more consensus. An individual's stereotype score could be obtained by summing up an individual's ratings on those dimensions identified as being part of the cultural stereotype (Gardner, 1973: 141).

Yet another way of measuring stereotypes is the *diagnostic ratio*, suggested by McCauley and Stitt (1978). Applying a Bayesian logic, the authors argue that former methods ignore baseline probabilities and suggest that a valid measure of stereotypes has to relate group specific estimates of the prevalence of a particular characteristics or a particular behavior to estimates how prevalent the same characteristic of behavior is among all people.

Methods that focus on the distribution of a particular characteristic or behavior among members of a group of people are the so called histogram or distribution task (Wyer et al., 2002; Park & Judd, 1990) and range task (Park & Judd, 1990). While the former – drawing a histogram or distribution of a characteristic within a social group – seems to be too much of a burden for some respondents (Park & Judd, 1990: 175), the range task is considered a fairly easy to understand measure that yields reliable estimates of both stereotypicality and dispersion (Corell et al., 2010: 53).

Since none of these traditional methods yields informative individual level data that is easy to collect through a concise instrument in paper-pencil self-administered questionnaires, we chose a simple and straightforward way of asking for the stereotype in question. How we developed and improved our measure is described in the next section.

## 4. Measuring Teachers' Stereotypes in the German National Educational Panel Study

Since NEPS uses paper-pencil self-administered questionnaires for educators and teachers at all stages, implicit measures were unfeasible to implement and we turned to explicit measures instead. The first assessment of teachers' stereotypes about the performance of different groups of students takes place in the fourth wave of Starting Cohort 2 ("Kindergarten and Elementary School"). This wave focusses on 2nd grade students and features interviews with their teachers and parents. We implemented measures of stereotypes in this cohort and at this early stage of the academic career since effects of stereotypes on academic performance are reported to be strongest among the youngest pupils (Jussim et al., 2009: 360). Measures of child competencies undertaken at later times may, thus, be influenced by teachers' stereotypes.

Because of the limited scope of the questionnaires and our interest in several groups of students, we had to restrict our measure to one key dimension. Theory suggests that the single most important belief for teachers' judgments in grading, ability grouping, and track recommendations should be the performance of a student or, for that matter, the average

performance of the group the student is categorized in by the teacher. This is backed up by empirical studies that find individual test scores to be the best predictor of grades and track recommendations at the end of elementary school in Germany (e.g., Bos et al., 2012a).

Surprisingly, we neither found an explicit measure of teachers' stereotypes about average group competencies readily available, nor did we find a measure that could have served as a starting point. Hence, we developed our own explicit measure of teachers' beliefs about the average competence of students from various social and ethnic groups. On this way, we had to answer the following questions:

*Which groups?* We ask teachers to report their stereotypes about those groups on which researchers in German education have—for various reasons—recently focused (for recent reviews see Kristen et al., 2011 and Stocké et al., 2011). These groups are: Girls, boys, students with lower, middle, and upper social class background, students of Turkish and Russian origin, as well as immigrants and ethnic majority students in general.

*Which stereotypes exactly should we assess?* Since math and German are the two major subjects in German elementary school and math and reading competencies are the two key competencies related to these subjects and therefore assessed in this NEPS wave of starting cohort 2, we decided to assess teachers' stereotypes regarding mathematical and reading competencies.

*How to ask for stereotypes?* The introduction serves the purpose of a cover story and is supposed to reduce social desirability bias by turning teachers' attention to the NEPS competence tests—instead of just asking for general or innate abilities or competencies of groups. Therefore, the introduction for the item battery reads as follows (see also Figure 1):

"In the NEPS study 'Educational trajectories in Germany' the competencies of children are assessed in different domains. What do you think how 2nd graders from various groups will perform in mathematics [reading]?"

Through this introduction we intend to give a good reason for asking such a question to keep teachers from ruminating about and, potentially, correctly guessing the question's true aim, namely to assess their stereotypes about different groups. At the same time, we wanted to ask in a general way that would allow teachers to report whatever they think of first when thinking of the competencies of different groups.

*Which response scale should we use?* We chose a response scale that allows teachers to express the belief that a particular group's competencies are average and, therefore, chose a scale with a midpoint. Also, we wanted to avoid confusion with the German grading scale that ranges from 1 for "very good" to 6 for "insufficient" or "fail". Therefore, we decided to use an 11-point scale—instead of 9-, 7-, or 5-point scale—that allows teachers to express finely nuanced beliefs.

Based on these considerations we intended to develop an item battery featuring two items for each of the nine groups, asking about the average competence level in math and reading, respectively. Figure 1 shows the first version of our instrument (see Appendix 1 for the original version in German language).

**In the NEPS study "Educational trajectories in Germany" the competencies of children are assessed in different subjects. What do you think how 2nd graders from various groups will perform in mathematics [reading]?**

**Compared to the mathematics [reading] performance of 2nd graders in general…**

*The further left you tick, the worse the group will perform according to your estimate, the further right you tick, the better. Please tick one square each line.*

| | | perform very poorly | | | | | | | | | | perform very well |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| a) | … girls | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| b) | … boys | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| c) | … children from lower social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| d) | … children from middle social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| e) | … children from higher social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| f) | … children of immigrant origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| g) | … children of Turkish origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| h) | … children of Russian origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| i) | … majority | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

*Figure 1*: Measuring teachers' stereotypes: first version (own translation)

**Developing the instrument and assessing its validity through cognitive interviews**

This first version was modified after feedback from both colleagues and teachers with whom we conducted cognitive pretests (Desimone & Floch, 2004). We evaluated our instrument through structured interviews with five teachers, recruited in the region of Bamberg, Germany[5]. Results from the cognitive interviews[6] lead to three major modifications of the first version (see Figure 2 and Appendix 2):

*Introduction*: While in the first version (see Figure 1 and Appendix 1) it is asked how children attending the second grade perform compared to second graders, in the second version—which is also implemented in a pilot study—we added a concrete reference and asked teachers to report their beliefs "[…] compared to the average". We did this because through cognitive interviews we learned that teachers almost exclusively referred to students in their current or previous classes. We wanted the question to allow for a broader understanding of it.

[5] The interviewed teachers teach at least mathematics or German. Two teachers are working in elementary schools, three in secondary schools. The recruitment of these teachers was realized through social contacts within the NEPS project.

[6] We probed participants retrospectively—that is we decided against the think aloud technique—to not disturb the thought process that respondents go through when answering our items.

*Repetition of the question wording*: In the revised version we introduced a question that repeats the key question and separates the different groups to remind the teachers of the task at hand. This was done to assure that teachers use the same anchor of reference for all judgments, and, thus, to avoid unwanted assimilation and contrast effects (Schwarz et al., 1991).

**In the NEPS study "Educational trajectories in Germany" the competencies of children attending the 2nd grade are assessed in different domains.**

**What do you think how 2nd graders from the following groups will perform compared to the average in the domain <u>mathematics</u> [<u>reading</u>]?**

*The further left you tick, the worse the group will perform according to your estimate, the further right you tick, the better. Please tick one square each line.*

| | | very poorly 0 | | | | | 5 | | | | | very well 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a) | Girls | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| b) | Boys | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**And how will the following groups perform compared to the average?**

| | | very poorly 0 | | | | | 5 | | | | | very well 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c) | Children from lower social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| d) | Children from middle social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| e) | Children from higher social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**And how will the following groups perform compared to the average?**

| | | very poorly 0 | | | | | 5 | | | | | very well 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| f) | Children of immigrant origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| g) | Children of Turkish origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| h) | Children of Russian origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| i) | Majority | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

*Figure 2.* Measuring teachers' stereotypes: second version (own translation)

*Labels of the response scale*: In addition, results of cognitive interviews suggested that the initial labeling of all numerical values from 0 to 10 on the response scale might have been misleading to some of the teachers who had in mind the German grading scale, which ranges from 1 to 6. Apparently, they ticked the same value—e.g., 2—they had in mind as grade—2

for "good"—ignoring the other values and the endpoint labels. By restricting the labels to values 0, 5, and 10 we hope to decrease the likelihood of such misunderstandings but still allow teachers to successfully navigate the scale.

**In the NEPS study "Educational trajectories in Germany" the competencies of children attending the 2nd grade are assessed in different domains.**

**What do you think how 2nd graders from the following groups will perform compared to all 2nd graders in Germany in mathematics [reading]?**

*The further left you tick, the worse the group will perform according to your estimate, the further right you tick, the better. Please tick one square each line.*

| | | far below average 0 | | | | | 5 | | | | | far above average 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a) | Children from lower social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| b) | Children from middle social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| c) | Children from higher social strata | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**And how will 2nd graders from the following groups perform compared to all 2nd graders in Germany?**

| | | far below average 0 | | | | | 5 | | | | | far above average 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| d) | Girls | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| e) | Boys | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**And how will 2nd graders from the following groups perform compared to all 2nd graders in Germany?**

| | | far below average 0 | | | | | 5 | | | | | far above average 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| f) | Children of immigrant origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| g) | Children of Turkish origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| h) | Children of Russian origin | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| i) | Majority | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

*Figure 3.* Measuring teachers' stereotypes: final version (own translation)

In the final version of the instrument—which was implemented in the main study and, hence, through which the data for the scientific use files of the NEPS is collected—three further modifications were undertaken based on discussions with colleagues (see Figure 3 and Appendix 3):

*Lead-in*: The lead-in is modified to what might be translated to "What do you think how 2nd graders from the following groups will perform compared to all 2nd graders in Germany in the domain mathematics [reading]?" Hence, while in the second version "the average" is the reference, in the third and *final version* a more precise description of the reference, namely "all 2nd graders in Germany" is used.

*Labels of the response scale*: For the final version, we changed the labels of the response scale from perform „very poorly" and perform „very well" into perform „far below average" and „far above average". The aim is to make the scale more relative, stress the reference group („all 2nd graders in Germany"), and, in consequence, to achieve a less skewed distribution as well as more variance.

*Order of social groups*: Finally, we swapped questions referring to the performance in dependency to sex and social origin. In consequence, the query now starts with children from lower social strata. The aim of this approach is to avoid a pattern where respondents contrast their responses within the social groups, e.g., referring to girls when estimating the performance of boys – rather than referring to all children attending the second grade.

To test the first modification of the lead-in in this version, we recruited new teachers for cognitive pretests and conducted in total four further interviews.[7] The results indicate that this modification yields the desired result.[8]

Finally, all nine cognitive pretests show that most teachers understand that the questions aim at their personal assessment. Two teachers explicitly reckon that the questions aim at their stereotypes about certain groups and their influence on the academic success of those groups. Furthermore, most teachers share a common understanding of key terms used in the instrument.

Almost all interviewed teachers define social strata mainly through income and/or education of the parents. In addition, some refer to the occupational status of the parents as well as to the learning environment and support at home. All in all, the teachers tend to have a similar understanding of the various social strata. Only one teacher had problems classifying different social strata. According to the interviewed teachers, lower social strata are characterized by living on welfare and/or a relatively low household income and/or a less beneficial learning environment at home. The middle social strata are associated with higher income. The higher social strata are associated with high education—e.g., a high rate of tertiary education—which also command higher financial resources.

Seven teachers described how they define persons of immigrant origin. Again, the results show that teachers largely agree: Almost all teachers refer to individuals who were born in a foreign country themselves or have at least one parent born abroad. Only one teacher

[7] All of them are working in elementary schools and teach mathematics and German.

[8] A broader database is needed to evaluate the second modification, thus, we cannot make any statements at this point.

mentions solely first generation immigrants. Six of the seven teachers also mention language competence and language use in the home environment as criterion. Furthermore, when estimating the performance of children of Russian origin, all interviewed teachers consider children from today's Russia as well as children stemming from the Former Soviet Union and the successor states.

There is no indication that teachers are confused or overburdened by the fact that we ask for performances in reading and mathematics over all, and not in specific fields within these domains; they rather show quite a similar understanding of these terms. Finally, there is no evidence that teachers misunderstood the fact that the value labels for the endpoints of the scale range over more than one box.

## 5. Data and Analytical Strategy

In this section we describe the data and the analytical strategy to assess the quality of our measure.

### 5.1 Data

Our analysis is based on data of the fourth wave of the pilot study in the NEPS Kindergarten cohort. At the time the survey was conducted, the children attended the second grade. The main aim of the NEPS pilot studies is to guarantee smooth main studies[9], e.g., by testing instruments and fieldwork. Just like the main studies, the corresponding pilot studies are conceptualized as panel studies. The sampling procedures of main and pilot studies are essentially equivalent. However, the pilot studies feature smaller samples and are conducted only in selected federal states of Germany. The sample for the pilot study in the Kindergarten cohort was drawn on four states: Bavaria, Thuringia, North Rhine-Westphalia, and Hamburg. In total, we can draw on 52 teacher interviews. Note that data from NEPS pilot studies is not released to the scientific community.

All quantitative results reported in this section refer to the second version of the instrument (see section 5). This version was implemented only in the pilot. However, all our findings should carry over to the final version that differs just slightly from the second version (see section 4).

### 5.2 Analytical strategy

We check the instrument for the following desirable properties: (i) variation between groups as a consequence of variation *within teachers*, (ii) variation *between teachers*, (iii) validity of the measure, and for the rather undesirable property of (iv) missing values. While examining missing values as well as variation within and between teachers is straightforward, validating our measure is less so. With regard to different forms of validity—content, criterion, and construct validity—we argue that content validity is satisfied by the question wording that rather explicitly asks for what has been defined as stereotype above. The cognitive interviews provide evidence that teachers understand our questions as intended. To assess criterion validity and construct validity, we perform various quantitative tests. We suggest that if the instrument is a valid measure of teachers' stereotypes, mean differences between

[9] The corresponding data from the main study are scheduled to be released as scientific use files in 2016.

groups, corresponding effect sizes[10], and correlations should be in line with theory and previous research.

Past research has shown that teachers' stereotypes—measured by *other* and *different* instruments—about the average academic performance of different groups of students are quite accurate (for a review see Jussim & Harber, 2005). Since teachers are experts in teaching students from different social and ethnic groups we expect them to be pretty accurate in their beliefs, too (Judd & Park, 1993). Therefore, one way of validating our measure, is to assess the *accuracy of teachers' beliefs* as measured by our instrument.

We assess this accuracy by comparing our results to findings from the most recent published studies we found. We check whether groups are ranked correctly on average, how many teachers rank groups correctly and how many don't, and by looking at effect sizes of group differences.

Whenever teachers' stereotypes are not accurate—that is biased—the bias should show patterns of ingroup favoritism or outgroup derogation (Tajfel & Turner, 1986), respectively, if the measure is a valid measure. That is, teachers' assessments should show a bias in favor of groups they belong to, and a negative bias towards those they do not belong to. Since teachers in German elementary schools are overwhelmingly female, belong to the middle or upper (middle) class, and are of German ethnicity without immigration background, we expect biases—if any—to the disadvantage of boys, students from lower class families, as well as immigrants in general, and different groups of immigrants in particular.

Another well replicated phenomenon in intergroup perception is outgroup homogeneity, which means that members of outgroups tend to be perceived as more similar to one another than members of ingroups (e.g., Judd & Park, 1988). We suggest that this mechanism should also hold on the group level: outgroups that can be categorized into one superordinate outgroup should be perceived as more similar than they actually are. In particular, we expect the different groups of immigrants, Turks and Russians, to be perceived more similar than they actually are, as they are easily categorized into a superordinate group of immigrants. Teachers' beliefs about groups that can be categorized into one superordinate group or groups that are otherwise perceived to be similar should also correlate positively. Therefore, we expect positive correlations between teachers' stereotypes for: boys and girls, immigrants and Turks, immigrants and Russians, Turks and Russians. In contrast, we expect low and insignificant correlations between unrelated groups such as girls or boys on the one hand and different groups of immigrants on the other hand.

[10] We calculate Cohen's *d* as $d = \frac{(\bar{x}_1 - \bar{x}_2)}{s^*}$, where $s^* = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ is the pooled standard deviation and where $s_1^2$ and $s_2^2$ are variances of $x_1$ and $x_2$, the beliefs of teachers about average performances of groups 1 and 2, respectively. Thus, $\bar{x}_1$ and $\bar{x}_2$ are the means of the beliefs about the performances of groups 1 and 2, respectively, over all teachers. Note that this strategy assumes that $s_1^2$ and $s_2^2$ are valid proxies for the average dispersion of groups 1 and 2, respectively, as perceived by the teachers.

# 6. Quantitative Results

## 6.1 Within Teacher Variation

Stereotypes may only help to explain discriminatory treatment and (conditional) inequality in education if they vary between target groups or if they are biased. Figure 4 summarizes between group variations as mean differences between teachers' stereotypes of the competencies of different groups. Teachers' stereotypes vary considerably between groups for both math (left panel) and reading (right panel). As math and reading competencies are empirically correlated (e.g., Rindermann, 2007), it is not surprising that the overall patterns look very similar. However, there are systematic differences with regard to gender.

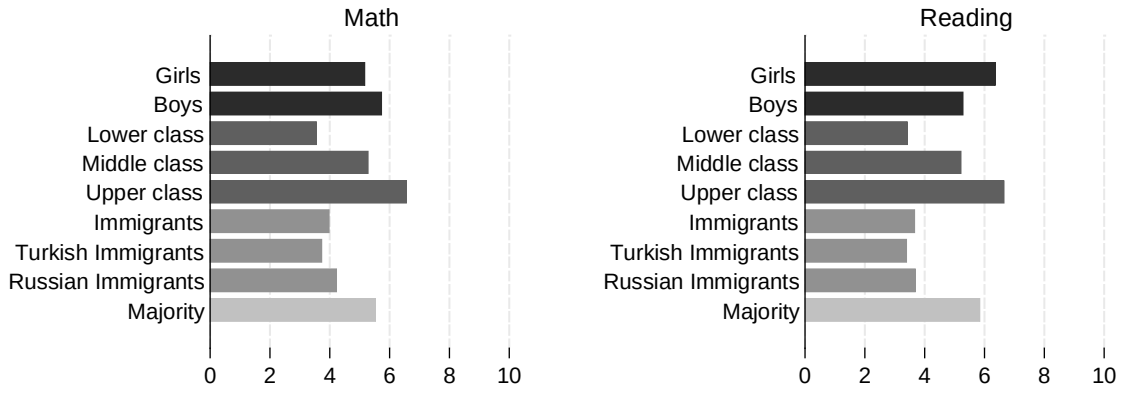


*Figure 4.* Means of teachers' estimation of students' results in NEPS competence tests for math (left panel) and reading (right panel). Groups (number of observations by stereotype in parentheses) from top to bottom: girls (math: 49/reading: 50), boys (49/50), lower class (45/46), middle class (45/48), upper class (45/48), immigrants (40/42), Turkish immigrants (35/37), Russian immigrants (37/39), majority (40/42).

### Gender

While teachers believe that boys outperform girls in math (-.55, $p<.05$), the opposite is true for reading (1.08, $p<.001$). Empirical studies provide strong evidence for this pattern (Bos et al., 2012a; Bos et al., 2012b; Mücke & Schründer-Lenzen, 2008). However, the same studies suggest that the teachers in our sample are mistaken in estimating the advantage of girls in reading (mean difference: 1.08, $p<.001$; Cohen's d: .78) to be about twice as large as boys' advantage over girls in mathematics (mean difference: -.55, $p<.05$; Cohen's d: -.41). Nationwide evidence for fourth-graders suggests that the gender gaps in math and reading competences are of similar size (Bos et al., 2012a; Bos et al., 2012b). Using data from a longitudinal study of 26 schools in Berlin, Mücke & Schründer-Lenzen (2008) moreover find that boys' advantage in math is even larger than their disadvantage in reading. Interestingly, the teachers' stereotypes are consistent with the results of the PISA study (e.g., Prenzel et al., 2013)—a study German media has covered extensively.

**Students from different social classes**

Teachers perceive large competence differences between students from different social classes: According to figure 4 these differences are very similar for math and reading and are all statistically significant (p<.001). Despite the promising results from the cognitive interviews, we don't know precisely what teachers had in mind when reporting their expectations about the competencies of different social classes. Thus, we cannot assess teachers' accuracy as precisely as for their stereotypes on gender differences. However, teachers correctly rank the three groups: All available studies show that students from lower class families perform worse than those from middle class families who, in turn, perform worse than upper class families (for recent evidence see, e.g., Bos et al., 2012a; Bos et al., 2012b; Prenzel et al., 2013).

**Immigrant students**

The performance of immigrants in general is estimated to be only marginally better than the performance of children from the lower social classes (see figure 4). The difference is larger in math (.58, p<.05) than in reading (.46), where it does not reach significance (p=.12).

The teachers' stereotype about Turkish immigrants does not differ significantly from the one about children from the lower social classes in both math (p=.19) and reading (p=.35). In contrast, teachers expect students of Turkish origin to perform worse than those of Russian origin in both math (-.45, p<.05) and reading (-.34, p=.07). This ranking is correct for both subjects and at different ages (Walter, 2009; Stanat et al., 2012).

However, judged by the numbers reported for fourth graders in Stanat et al. (2012), it seems as if teachers perceive the two groups of immigrants as more similar than they actually are: The *actual achievement gap* between Germans and Turks (Math: 515-433=82; Reading: 514-429=85) is about 2.6 times larger than the one between Germans and Russians (Math: 515-483=32; Reading: 514-481=33). In contrast, the *perceived achievement gap* towards Germans differs much less between Turks and Russians: In math the Turkish disadvantage is perceived to be about 1.4 times larger (-1.89/-1.35=1.4) than the Russian disadvantage, in reading the Turkish disadvantage is perceived to be only 1.15 times (-2.54/-2.21=1.15) larger.

With regard to subjects, teachers perceive larger achievement gaps in reading than in math—1.34 times larger for Turks, 1.64 times larger for Russians. In reality, these differences are substantially smaller—1.04 times for Turks and only 1.03 times for Russians (Stanat et al., 2012)—and, as simple t-tests of the published results reveal, not significant for both Turks and Russians. In fact, the differences between math and reading within groups are perceived to be larger than the between-group differences. This is clearly wrong.

Effect sizes allow for a more direct comparison of the magnitude of the actual and perceived achievement gaps between the two groups of immigrants and students of the ethnic majority. By and large, effect sizes yield the same conclusions as means: Teachers perceive *less pronounced differences* between the two groups of immigrants in both math (Cohen's *d*: -.36) and reading (Cohen's *d*: -.18) than actually exist (math: -.57, reading: -.56, Stanat et al., 2012). Furthermore, results for effects sizes confirm that teachers *perceive larger differences* between Germans and the two groups of immigrants (Turks: -1.72 for Math, -1.62 for

Reading; Russians: -1.17 for Math, -1.44 for Reading) than actually exist (Turks: -.93 for math, -.88 for reading; Russians: -.35 for math, -.34 for reading). Therefore—judged by effect sizes—teachers *underestimate both groups*, Turks and Russians, relative to their German peers. For Russians, the perceived gap is 3.3 times larger than the true gap for Math and even 4.2 times larger for Reading. While the Turkish disadvantage is also perceived to be larger than it actually is, the teachers' bias is less extreme: in both math and reading the Turkish disadvantage is overestimated by a factor of about 1.9.

In contrast to what mean differences show, effect sizes yield slightly larger achievement gaps in math compared to reading for both groups (Turks: -.93 vs -.88, factor of 1.06; Russians: -.35 vs -.34, factor of 1.03). Effect sizes for perceived differences suggest that teachers correctly believe that the disadvantage for Turkish students in Math is larger than the one in reading. However, teachers believe that Russians will perform worse in Reading compared to math. Judged by the results of Stanat et al. (2012), the stereotype of a comparatively high math proficiency of students with a Russian background is incorrect.

## 6.2 Between teacher variation

Since the same between group variation may stem from few teachers perceiving large differences or many perceiving small differences, figures 5 and 6 show the variation within teachers as a difference between two selected groups as rated by the same teacher. For both math and reading the plots show that teachers differ to some degree in their estimates of group differences: Not only are different groups of students estimated to have different competencies, some teachers perceive much larger differences between the groups than others, some see no differences at all.

Very stable, however, is the perception of which group of two, if any, is in front. Actually, for all comparisons shown in figures 5 and 6—except the one between girls and boys—teachers agree on which group, if any, they expect to be ahead. A notable exception is the difference between girls and boys in math, where some teachers (18.4%) see girls ahead of boys, while about a third of teachers perceives no differences (32.7%) or a small advantage for boys (49%). Thus, teachers tend to hold biased views to the disadvantage of boys. In reading, teachers either see no difference between the sexes (46%) or—in line with empirical results—girls ahead of boys (54%).

Most teachers correctly rank students of different social class background: 80% expect students from middle class families to perform better than their lower class peers. In the same vein, 86.7% expect students from upper class families to outperform students with lower class background. 73.3% also see an advantage for upper class students compared to middle class students. These numbers support the notion of correct stereotypes with regard to social class differeces held by teachers.

The vast majority of teachers correctly expects majority students to outperform immigrants in both reading (85%) and math (76.3%) assessments. However, teachers' ranking of students of Turkish and Russian origin are not as accurate. For math, only 39.4% of teachers expect Russians to perform better than Turks, while a majority of 51.2% mistakenly believes that these groups of students will perform equally well and 9.1% even suggest that students of Turkish origin will perform better than those of Russian origin. For reading the numbers are similar and, thus, similarily wrong: only 31.4% expect what actually is the case, namely

that Russians outperform Turks in reading. Instead, 62.9% expect the two groups to perform equally well, 5.7% expect Turks to perform better. These results suggest a bias disadvantaging students of Russian origin.

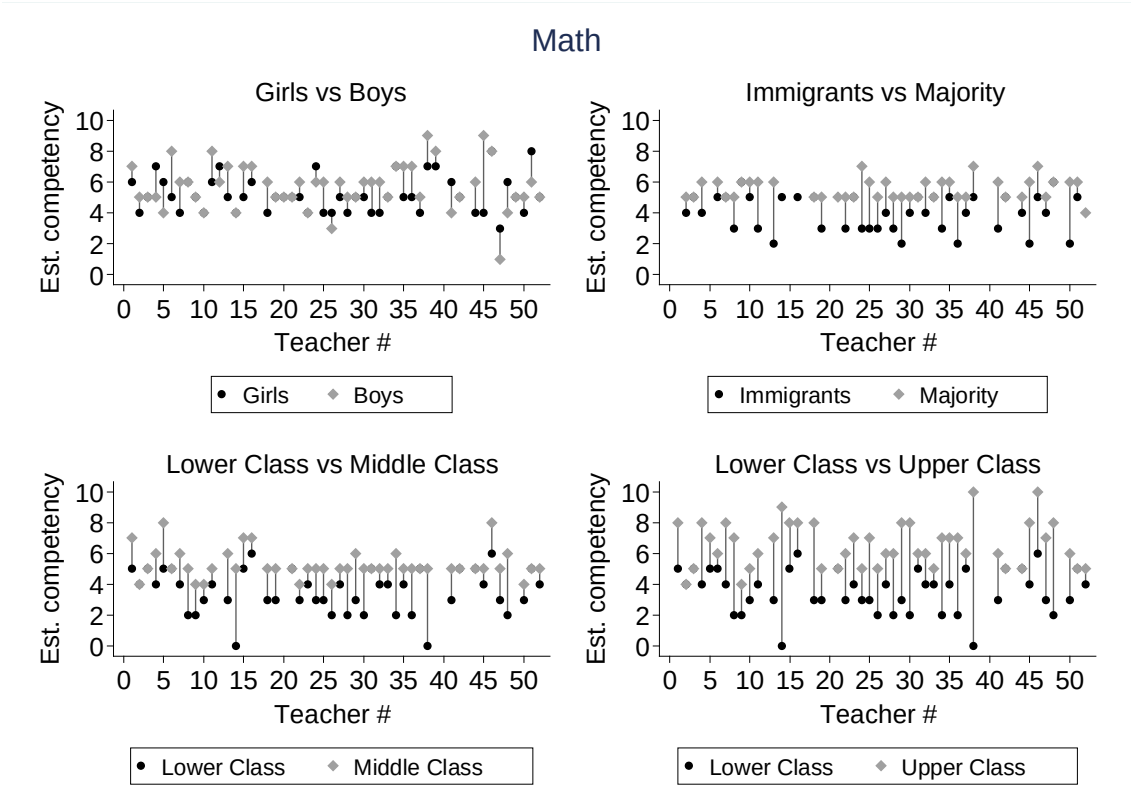


*Figure 5:* Range plots of the differences between teachers‘ stereotypes of group specific competencies in math by teacher ID.

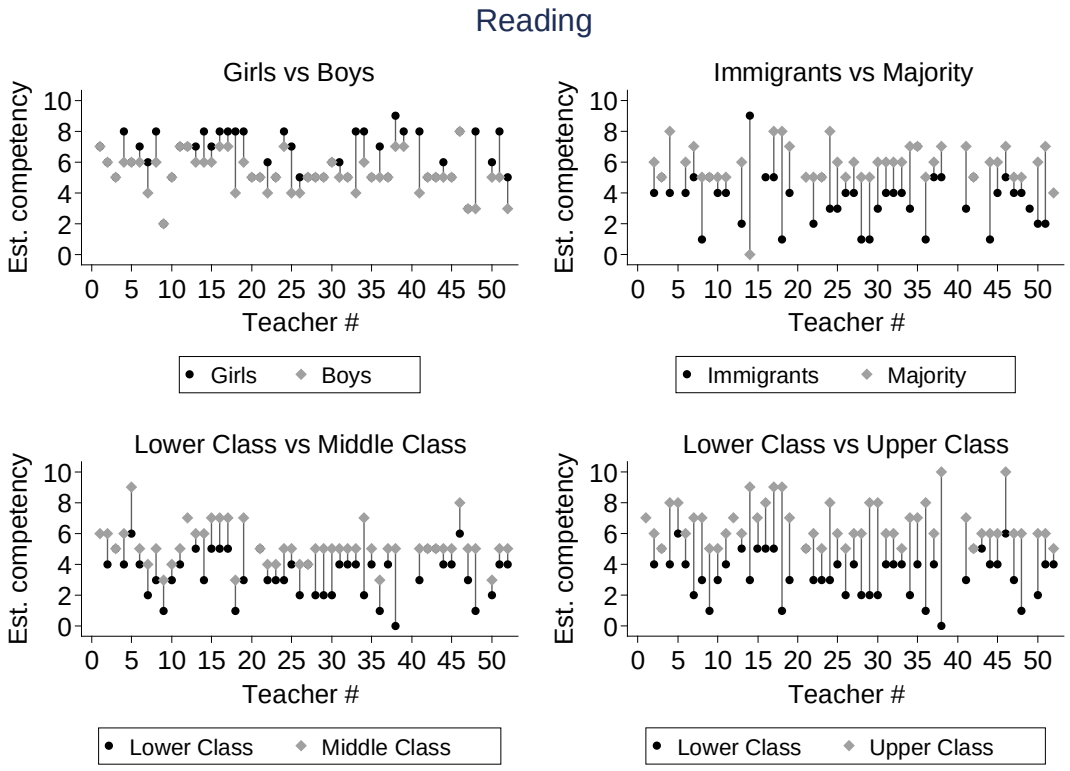


*Figure 6:* Range plots of the differences between teachers' stereotypes of group specific competencies in reading by teacher ID.

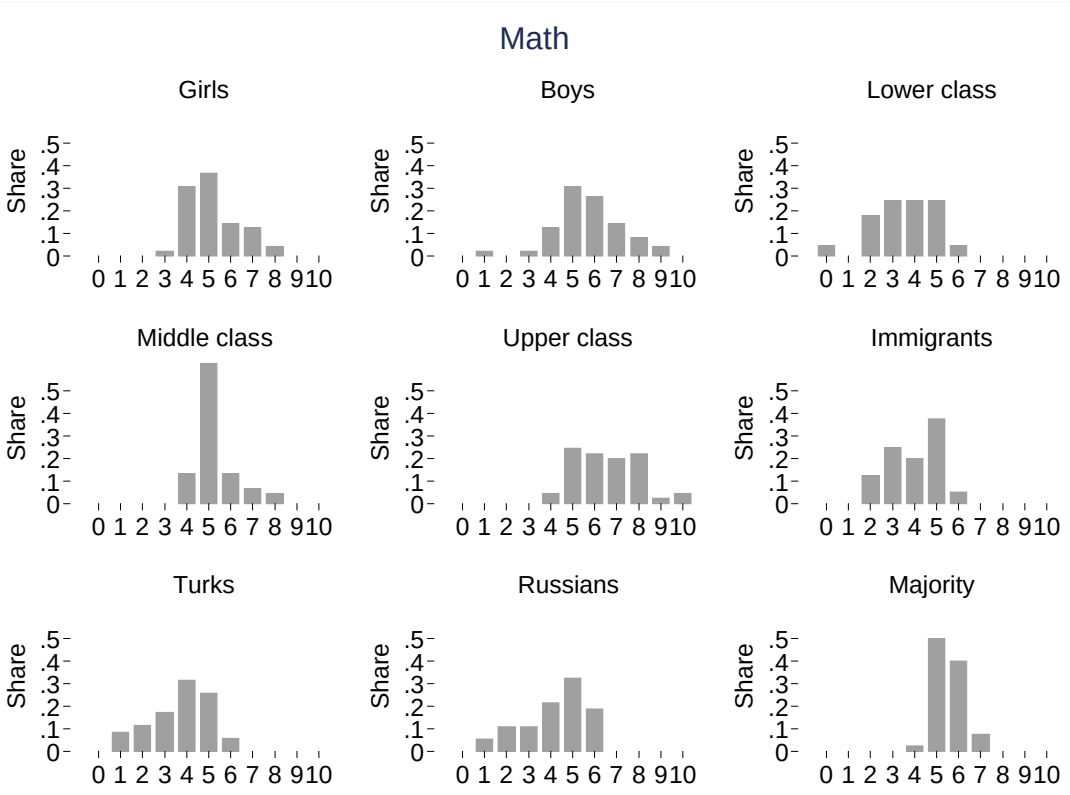


*Figure 7.* Histograms of teachers' stereotypes about group specific competencies in math.

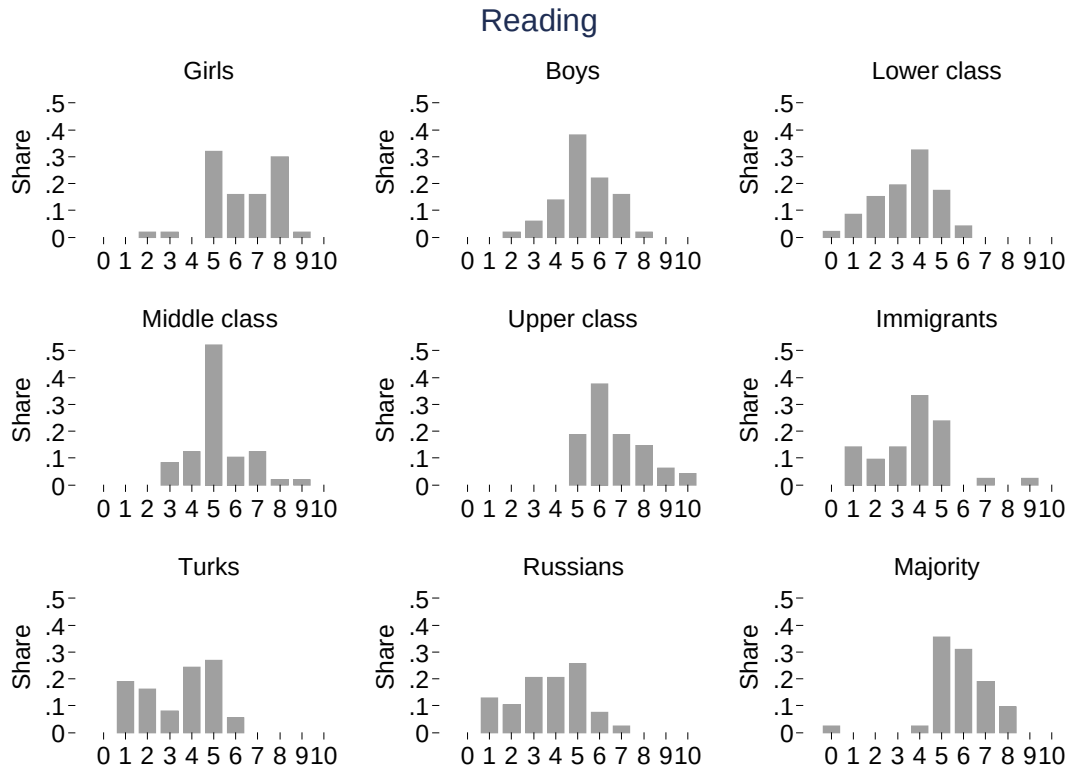


*Figure 8.* Histograms of teachers' stereotypes about group specific competencies in reading.

Figures 7 and 8 show histograms of all single items. The histograms highlight the variation between teachers. If there were none, each histogram would only show one bar. It is quite clear from figures 7 and 8 that for both math and reading there are large differences between teachers in the assessment of average competencies of one and the same group of students. In addition to the variation between teachers, the histograms also show that there are differences in the degree to which teachers vary in their assessment of one and the same group. Take immigrants' and majority students' math competencies (figure 7), for instance: While teachers' stereotypes of majority students' competencies are virtually limited to 5, the midpoint of the scale, and 6, immigrants' competencies are estimated to vary considerably.

**Item intercorrelations**

Tables 1 and 2 present item intercorrelations for both domains and all groups. Girls and boys are very likely to serve as standards of comparison for each other in both domains and, hence, should correlate positively. Actually, we observe positive and significant correlations of .42 for math and .56 for reading. The same logic should apply to the different social classes. Thus, we also expect positive and significant correlations among teachers' stereotypes for the three groups. Interestingly, we observe positive and significant correlations in both math and reading between lower class and middle class (math: .41, reading: .60) and between middle class and upper class (math: .60, reading: .40) but not between lower class and upper class (math: -.19, reading: -.14). Furthermore, we expect strong correlations among the estimates for immigrants in general and immigrants of Turkish and Russian origin in particular. In fact, these correlations are all statistically significant and range from .56 to .79. These strong correlations contrast with low and

insignificant correlations between unrelated groups such as girls and boys on the one hand and the different groups of immigrants on the other hand—these correlations are virtually zero.

*Table 1: Item intercorrelations for math*

| | Girls | Boys | Lower Class | Middle Class | Upper Class | Immi-grants | Turks | Russians | Majority |
|---|---|---|---|---|---|---|---|---|---|
| Girls | 1 | | | | | | | | |
| Boys | .42** | 1 | | | | | | | |
| Lower Class | .10 | .16 | 1 | | | | | | |
| Middle Class | .48*** | .25 | .41** | 1 | | | | | |
| Upper Class | .25 | .36* | -.19 | .60*** | 1 | | | | |
| Immigrants | .09 | -.12 | .21 | .12 | .04 | 1 | | | |
| Turks | .04 | .06 | .34* | .06 | .03 | .59*** | 1 | | |
| Russians | .04 | .09 | .18 | .06 | .07 | .69*** | .75*** | 1 | |
| Majority | .57*** | .46** | -.12 | .28 | .34* | .06 | .00 | .04 | 1 |

* $p < .05$, ** $p < .01$, *** $p < .001$

*Table 2: Item intercorrelations for reading*

| | Girls | Boys | Lower Class | Middle Class | Upper Class | Immi-grants | Turks | Russians | Majority |
|---|---|---|---|---|---|---|---|---|---|
| Girls | 1 | | | | | | | | |
| Boys | .56*** | 1 | | | | | | | |
| Lower Class | -.03 | .36* | 1 | | | | | | |
| Middle Class | .39** | .62*** | .60*** | 1 | | | | | |
| Upper Class | .61*** | .58*** | -.14 | .40** | 1 | | | | |
| Immigrants | -.08 | .13 | .26 | .29 | .07 | 1 | | | |
| Turks | -.17 | .11 | .34* | .26 | .02 | .78*** | 1 | | |
| Russians | -.06 | .00 | .29 | .19 | -.11 | .56*** | .79*** | 1 | |
| Majority | .33* | .26 | .05 | .16 | .29 | -.28 | -.11 | .15 | 1 |

* $p < .05$, ** $p < .01$, *** $p < .001$

## 6.3 Missing values

Theoretically, all teachers should have at least some stereotypical knowledge about how competent the different groups are. However, teachers might fail to introspectively access relevant information, to compute a judgment, as well as format and edit their response (see Sudman et al., 1996). Of course, teachers might also simply be unwilling to report their stereotypes or save time by skipping the question.

Table 3 reports both the absolute number and the relative share of missing values per item. While almost all teachers give a valid answer in case of girls and boys, the share of missing values increases to 7.7% to 13% for the items on social class. For immigrants and majority students in general these numbers jump up to 19% for reading and 23% for math. Between

25% and 33% of teachers do not estimate the competence of students of Turkish and Russian origin.

These numbers aggregate up to patterns of missing values. By far the most common pattern is a valid response on all items for both math (63%) and reading (67%). For math, the second most common pattern (8%) is valid answers only for girls and boys, followed by missing values on all items (6%), missing values on the last four items (6%)—i.e. immigrants, Turks, Russians, and majority students—and only the value for majority students missing (6%). In contrast, for reading, the second most common patterns are the one with missing values on the last four items (6%) and only majority students missing (6%). That teachers give valid estimates for girls and boys only happens less often than for math (4%).

*Table 3: Absolute number and relative share of missing values per item*

| | Math | | Reading | |
|---|---|---|---|---|
| | # of missings | Share | # of missings | Share |
| girls | 3 | 0.06 | 2 | 0.04 |
| boys | 3 | 0.06 | 2 | 0.04 |
| lower class | 7 | 0.13 | 6 | 0.12 |
| middle class | 7 | 0.13 | 4 | 0.08 |
| upper class | 7 | 0.13 | 4 | 0.08 |
| immigrants | 12 | 0.23 | 10 | 0.19 |
| Turkish | 17 | 0.33 | 15 | 0.29 |
| Russian | 15 | 0.29 | 13 | 0.25 |
| majority | 12 | 0.23 | 10 | 0.19 |

## 7. Summary and Conclusion

In this paper we have introduced an item battery to measure teachers' stereotypes about the average competencies in math and reading of different social and ethnic groups, namely girls, boys, students with lower, middle, and upper class background, students of Turkish and Russian origin, as well as students of immigrant origin and majority students.

Understood as a belief or a set of beliefs about the characteristics, attributes, or behaviors of a particular group or category of people, a stereotype contains more or less accurate beliefs, is held by individuals, and may be measured using implicit or explicit methods. Like many other large-scale assessments in education, the NEPS makes use of paper-pencil self-administered questionnaires, where implicit measures are unfeasible to implement. Therefore, we developed an explicit measure of teachers' stereotypes.

By means of cognitive interviews we identified a few minor problems respondents might have had with the first version and developed an improved second version of the item battery. This second version was tested in a pilot study with a sample of 52 second-grade teachers from four German federal states.

Quantitative analyses show both a large variation between groups—as a consequence of variation *within* teachers—and a large variation *between* teachers. Both are desirable properties if the instrument is to be used to answer substantive research questions by

means of quantitative analyses. Furthermore, the analyses suggest that teachers' stereotypes are *quite accurate overall* in that they reflect group differences in achievement as reported in the recent empirical literature. Since teachers are experts with regard to scholastic achievement of different groups of students, we take this as indicative of the validity of the instrument. However, we also find biases to the disadvantage of boys, immigrants in general, as well as immigrants of Turkish and Russian origin in particular. As expected, these results speak to the general phenomenon of bias in favor of one's—here: teachers'—ingroups, and, hence, to the validity of the instrument. We also find evidence for an outgroup homogeneity effect on the group level: Students of Turkish origin and those of Russian origin are perceived to be more similar than they are according to published studies. What is especially harmful for students of Russian origin—they receive relatively poor assessment in comparison to students of Turkish descent—is yet another piece of evidence for the validity of our measure of teachers' stereotypes. The fact that estimates for similar or related groups correlate positively, while estimates for unrelated groups do not also speaks to the validity of the instrument.

Quite obviously, both instrument and paper have shortcomings and limitations: With regard to the instrument, we cannot rule out that teachers adjust their responses towards what they believe to be socially desirable responses. If they do, chances are that they report *smaller group differences* in general and *less negative* stereotypes than they truly hold towards outgroups in particular. However, as shown by cognitive interviews, this should only affect the responses of a minority of teachers. Our question wording seems to successfully hide the true purpose of the instrument and motivate teachers to truthfully report their beliefs. Therefore, the problem of social desirability bias should not be severe. Another shortcoming is the relatively large share of missing values for immigrants, Turks, Russians, and majority students. However, it might be that teachers who did not answer these items have less or no experience with students of such origin. If so, the larger share of missing values for these groups would be less problematic, since teachers' stereotypes should affect only the outcomes of students they actually teach. Whether teachers who did not report a particular stereotype actually have less or no experience with students from the group in question, could and should be tested in future research.

The major limitation of this paper is that we cannot directly assess the validity of the instrument. All analyses reported above provide rather indirect evidence that the instrument is indeed a valid measure of teachers' stereotypes. Unfortunately, we could not implement alternative measures of the assessed stereotypes to more directly test the instrument's validity. In this regard we expect future research to provide further insights.

Finally, we like to stress that with this item battery, the NEPS is the first panel study that offers explicit measures of teachers' stereotypes about the average competencies in math and reading of different social and ethnic groups available to the scientific community through Scientific Use Files. We look forward to many different applications using the items described in this paper.

## References

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. Industrial and Labor Relations Review, 30(2), 175–187.

Allport, G. W. (1954). The nature of prejudice. Malden, MA: Addison-Wesley.

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (Eds.), IGLU 2006: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich (pp. 271–298). Münster: Waxmann.

Ashmore, R. D., & Boca, F. K. D. (1979). Sex stereotypes and implicit personality theory: Toward a cognitive - social psychological conceptualization. Sex Roles, 5(2), 219–248. http://doi.org/10.1007/BF00287932

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), Cognitive Processes in Stereotyping and Intergroup Behavior (pp. 1–35). Hillsdale, NJ: Erlbaum.

Bless, H., Fiedler, K., & Strack, F. (2004). Social cognition: How individuals construct social reality. Hove, UK: Psychology Press.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, (Special Issue 14). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. Journal of Personality, 52(5), 871–880.

Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (Eds.). (2012a). IGLU 2011: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Münster: Waxmann.

Bos, W., Wendt, H., Köller, O., & Selter, C. (Eds.). (2012b). TIMSS 2011: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Münster: Waxmann.

Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2010). Long-term Trends in Educational Inequality in Europe: Class Inequalities and Gender Differences. European Sociological Review, 26(1), 31–48. http://doi.org/10.1093/esr/jcp001

Brigham, J. C. (1971). Ethnic stereotypes. Psychological Bulletin, 76(1), 15–38. http://doi.org/10.1037/h0031446

Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes, and discrimination. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), The SAGE handbook of prejudice, stereotyping, and discrimination (pp. 45–62). Thousand Oaks: Sage.

Crandall, C. S., Bahns, A. J., Warner, R., & Schaller, M. (2011). Stereotypes as justifications of prejudice. Personality and social psychology bulletin, 37(11), 1488–1498. http://doi.org/10.1177/0146167211411723

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. Journal of Personality and Social Psychology; Journal of Personality and Social Psychology, 44(1), 20-33.

De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwarz (Eds.), Implicit measures of attitudes (pp. 59–102). New York, NY: The Guilford Press.

Desimone, L. M., & Floch, K. C. L. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. Educational Evaluation and Policy Analysis, 26(1), 1–22. http://doi.org/10.3102/01623737026001001

Ditton, H., Krüsken, J., & Schauenberg, D.-S. M. (2005). Bildungsungleichheit—der Beitrag von Familie und Schule. Zeitschrift für Erziehungswissenschaft, 8(2), 285–304.

Dovidio, J. F., Hewstone, M., Glick, P., Esses, V. M. (2010). Prejudice, stereotyping and discrimination: theoretical and empirical overview. In Dovidio, J. F., Hewstone, M., Glick, P., Esses, V. M. The SAGE handbook of prejudice, stereotyping and discrimination (pp. 3-29). Thousand Oaks: Sage.

Ehrlich, H. J. (1973). The Social psychology of prejudice: A systematic theoretical review and propositional inventory of the American Social Psychological Study of Prejudice. New York: John Wiley & Sons.

England, P., & Lewin, P. (1989). Economic and sociological views of discrimination in labor markets: Persistence or demise? Sociological Spectrum, 9(3), 239–257.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? Journal of Personality and Social Psychology, 69(6), 1013–1027. http://doi.org/10.1037/0022-3514.69.6.1013

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. Annual Review of Psychology, 54(1), 297–327. http://doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., & Petty, R. E. (2008). Attitudes: Their structure, function, and consequences. New York: Psychology Press.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The Handbook of Social Psychology (p. 357–411), New York, NY: McGraw-Hill.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model. Ten years later. In S. Chaiken & Y. Trope (Eds.), Dual process theories in social psychology (p. 231-254). New York, NY: Guilford.

Fiske, S. T., C, J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology, 82(6), 878–902. http://doi.org/10.1037/0022-3514.82.6.878

Gardner, R. (1973). Ethnic stereotypes: The traditional approach, a new look. Canadian Psychologist, 14(2), 133–148. http://doi.org/10.1037/h0082215

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. Canadian Psychology, 50(3), 141–150. http://doi.org/10.1037/a0013848

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. Perspectives on Psychological Science, 2(2), 181–193. http://doi.org/10.1111/j.1745-6916.2007.00036.x

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4–27. http://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology, 74(6), 1464–1480. http://doi.org/10.1037/0022-3514.74.6.1464

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. Annual Review of Psychology, 47(1), 237–271. http://doi.org/10.1146/annurev.psych.47.1.237

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. Personality and Social Psychology Bulletin, 31(10), 1369–1385. http://doi.org/10.1177/0146167205275613

Jones, J. M. (1997). Prejudice and racism (2nd ed.). New York: McGraw-Hill.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. Personality and Social Psychology Review, 9(2), 131–155. http://doi.org/10.1207/s15327957pspr0902_3

Jussim, L., Robustelli, S. L., & Cain, T. R. (2009). Teacher expectations and self-fulfilling prophecies. In K. R. Wentzel & A. Wigfield (Eds.), Handbook of motivation at school (pp. 349–380). New York & London: Routledge.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. Journal of Personality and Social Psychology, 54(5), 778–788. http://doi.org/10.1037/0022-3514.54.5.778

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. Psychological Review, 100(1), 109–128. http://doi.org/10.1037/0033-295X.100.1.109

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. The Journal of Abnormal and Social Psychology, 28(3), 280–290. http://doi.org/10.1037/h0074049

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. Education Economics, 21(5), 447–463. http://doi.org/10.1080/09645292.2011.585019

Kristen, C. (2006). Ethnische Diskriminierung in der Grundschule? Die Vergabe von Noten und Bildungsempfehlungen. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58(1), http://doi.org/79–97.10.1007/s11575-006-0004-y

Kristen, P. D. C., Edele, D.-P. A., Kalter, P. D. F., Kogan, P. D. I., Schulz, B., Stanat, P. P., & Will, D.-P. G. (2011). 8 The education of migrants and their children across the life course.

Zeitschrift für Erziehungswissenschaft, 14(2), 121–137. http://doi.org/10.1007/s11618-011-0194-3

LaPiere, R. T. (1934). Attitudes vs. actions. Social Forces, 13(2), 230–237. http://doi.org/10.2307/2570339

Lippmann, W. (1922). Public opinion. Oxford: Harcourt Brace.

Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68(1), 89–111. http://doi.org/10.1007/s11577-015-0352-3

Lucas, S. R. (2008). Theorizing discrimination in an era of contested prejudice discrimination in the United States. Philadelphia: Temple University Press.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. Annual Review of Psychology, 51(1), 93–120. http://doi.org/10.1146/annurev.psych.51.1.93

Marks, G. N. (2005). Cross-national differences and accounting for social class inequalities in education. International Sociology, 20(4), 483–505. http://doi.org/10.1177/0268580905058328

Marks, G. N. (2005). Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries. Ethnic and Racial Studies, 28(5), 925–946. http://doi.org/10.1080/01419870500158943

McCauley, C., & Stitt, C. (1978). An individual and quantitative measure of stereotypes. Journal of Personality, 36(9), 929–940. http://doi.org/10.1037/0022-3514.36.9.929

Merton, R. K. (1949). Discrimination and the American Creed. In R. M. McIver (Ed.), Discrimination and National Welfare (pp. 99–126). New York & London: Harper & Brothers.

Mücke, S., & Schründer-Lenzen, A. (2008). Zur Parallelität der Schulleistungsentwicklung von Jungen und Mädchen im Verlauf der Grundschule. In B. Rendtorff & A. Prengel (Eds.), Kinder und ihr Geschlecht (pp. 135–146). Frankfurt am Main: Verlag Barbara Budrich, Opladen.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. International Economic Review, 14(3), 693–709. http://doi.org/10.2307/2525981

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Champaign, Illinois: University of Illinois Press.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology, 34, 181-209. http://doi.org/10.1146/annurev.soc.33.040406.131740

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. Journal of Personality, 42(6), 1051–1068. http://doi.org/10.1037/0022-3514.42.6.1051

Petty, R. E., Fazio, R. H., & Briñol, P. (2009). Attitudes: Insights from the New Implicit Measures. New York: Psychology Press.

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Eds.). (2013). PISA 2012: Fortschritte und Herausforderungen in Deutschland. Münster: Waxmann.

Reskin, B. F. (2003). Including mechanisms in our models of ascriptive inequality: 2002 presidential address. American Sociological Review, 68(1), 1–21. http://doi.org/10.2307/3088900

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. European Journal of Personality, 21(5), 667–706. http://doi.org/10.1002/per.634

Schneider, D. J. (2004). The Psychology of Stereotyping (1st ed.). New York & London: The Guilford Press.

Schneider, T. (2011). Die Bedeutung der sozialen Herkunft und des Migrationshintergrundes für Lehrerurteile am Beispiel der Grundschulempfehlung. Zeitschrift Für Erziehungswissenschaft, 14(3), 371–396. http://doi.org/10.1007/s11618-011-0221-4

Schütz, H., & Six, B. (1996). How strong is the relationship between prejudice and discrimination? A meta-analytic answer. International Journal of Intercultural Relations, 20(3), 441–462. http://doi.org10.1016/0147-1767(96)00028-4

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. Public Opinion Quarterly, 55(1), 3–23. http://doi.org/10.1086/269239

Shavit, Y., & Blossfeld, H.-P. (1993). Persistent inequality: Changing educational attainment in thirteen countries. Boulder, Colo.: Westview Press.

Snyder, M., & Swann, W. (1978). Hypothesis-testing processes in social interaction. Journal of Personality, 36(11), 1202–1212. http://doi.org/10.1037/0022-3514.36.11.1202

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. Empirical Economics, 45(1), 523–538. http://doi.org/10.1007/s00181-012-0609-x

Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. Zeitschrift Für Erziehungswissenschaft, 14(2), 103–119. http://doi.org/10.1007/s11618-011-0193-4

Stubbe, T. C., Bos, W., & Euen, B. (2012). Kapitel VIII. Der Übergang von der Primar- in die Sekundarstufe. In W. Bos, I. Tarelli, A. Bremerich-Vos, & K. Schwippert (Eds.), IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. (pp. 209–226). Münster: Waxmann.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), Psychology of intergroup relations (pp. 7-24). Chicago, IL: Nelson-Hall Publishers.

Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. Social Justice Research, 21(3), 263–296. http://doi.org/10.1007/s11211-008-0071-2

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011. Münster: Waxmann.

Vargas, P. T., D. Sekaquaptewa, & W. von Hippel (2007) Armed only with paper and pencil: "Low-tech" implicit measures of attitudes, prejudice, and self-esteem. In B. Wittenbrink & N. (Eds.), Implicit measures of attitudes (pp. 103-124) New York, NY: Guilford Press.

Walter, O. (2009). Herkunftsassoziierte Disparitäten im Lesen, der Mathematik und den Naturwissenschaften: Ein Vergleich zwischen PISA 2000, PISA 2003 und PISA 2006. In M. Prenzel & J. Baumert (Eds.), Vertiefende Analysen zu PISA 2006 (pp. 149–168). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wenz, S. E., & Hoenig, K. (2013). Ethnic and social class discrimination in education. Experimental evidence from Germany. Paper presented at the ASA 2013 Annual Meeting, New York, NY.

Wyer, N. A., Sadler, M. S., & Judd, C. M. (2002). Contrast effects in stereotype formation and change: The role of comparative context. Journal of Experimental Social Psychology, 38(5), 443–458. http://doi.org/10.1016/S0022-1031(02)00010-0

# Appendix

## A 1: First version (original)

**Die NEPS-Studie „Bildungsverläufe in Deutschland“ erfasst die Kompetenzen der Kinder in unterschiedlichen Bereichen. Was denken Sie, wie Schülerinnen und Schüler der zweiten Klassen aus verschiedenen Gruppen im Kompetenzbereich Mathematik abschneiden werden?**

**Im Vergleich zu Zweitklässlern insgesamt schneiden im Kompetenzbereich <u>Mathematik [Lesen]</u> …**

*Je weiter links Sie Ihr Kreuz machen, desto schlechter schneidet die Gruppe Ihrer Einschätzung nach ab, je weiter rechts Sie Ihr Kreuz machen, desto besser schneidet die Gruppe ab. Bitte in jeder Zeile ein Kästchen ankreuzen.*

| | | sehr schlecht ab | | | | | | | | | | sehr gut ab |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| a) | … Mädchen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| b) | ... Jungen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| c) | ... Kinder aus niedrigen sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| d) | ... Kinder aus mittleren sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| e) | ... Kinder aus hohen sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| f) | … Kinder mit Migrations-hintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| g) | ... Kinder mit türkischem Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| h) | … Kinder mit russischem Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| i) | ... Kinder ohne Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

## A 2: Second version (original)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **In der NEPS-Studie „Bildungsverläufe in Deutschland“ werden die Kompetenzen von Kindern in der zweiten Klasse in unterschiedlichen Bereichen erfasst.**<br>**Was denken Sie, wie Zweitklässler aus den folgenden Gruppen im Kompetenzbereich <u>Mathematik [Lesen]</u> im Vergleich zum Durchschnitt abschneiden werden?** | | | | | | | | | | | | |
| *Je weiter links Sie Ihr Kreuz machen, desto schlechter schneidet die Gruppe Ihrer Einschätzung nach ab, je weiter rechts Sie Ihr Kreuz machen, desto besser schneidet die Gruppe ab. Bitte in jeder Zeile ein Kästchen ankreuzen.* | | | | | | | | | | | | |
| | sehr schlecht | | | | | | | | | | sehr gut | |
| | 0 | | | | | 5 | | | | | 10 | |
| a) Mädchen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| b) Jungen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| **Und wie werden die folgenden Gruppen im Vergleich zum Durchschnitt abschneiden?** | | | | | | | | | | | | |
| | sehr schlecht | | | | | | | | | | sehr gut | |
| | 0 | | | | | 5 | | | | | 10 | |
| c) Kinder aus niedrigen sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| d) Kinder aus mittleren sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| e) Kinder aus hohen sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| **Und wie werden die folgenden Gruppen im Vergleich zum Durchschnitt abschneiden?** | | | | | | | | | | | | |
| | sehr schlecht | | | | | | | | | | sehr gut | |
| | 0 | | | | | 5 | | | | | 10 | |
| f) Kinder mit Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| g) Kinder mit türkischem Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| h) Kinder mit russischem Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |
| i) Kinder ohne Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | |

## A 3: Final version (original)

**In der NEPS-Studie „Bildungsverläufe in Deutschland“ werden die Kompetenzen von Kindern in der zweiten Klasse in unterschiedlichen Bereichen erfasst.**

**Was denken Sie, welche Ergebnisse Zweitklässler aus folgenden Gruppen im Kompetenzbereich Mathematik [Lesen] im Vergleich zu Zweitklässlern in Deutschland insgesamt erzielen?**

*Je weiter links Sie Ihr Kreuz machen, desto schlechter werden die Ergebnisse der Gruppe Ihrer Einschätzung nach ausfallen, je weiter rechts Sie Ihr Kreuz machen, desto besser werden sie ausfallen. Bitte in jeder Zeile ein Kästchen ankreuzen.*

| | | weit unter-durchschnittliche 0 | | | | | 5 | | | | | weit über-durchschnittliche 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a) | Kinder aus niedrigen sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| b) | Kinder aus mittleren sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| c) | Kinder aus hohen sozialen Schichten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Und welche Ergebnisse erzielen Zweitklässler aus folgenden Gruppen im Vergleich zu Zweitklässlern in Deutschland insgesamt?**

| | | weit unter-durchschnittliche 0 | | | | | 5 | | | | | weit über-durchschnittliche 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| d) | Mädchen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| e) | Jungen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Und welche Ergebnisse erzielen Zweitklässler aus folgenden Gruppen im Vergleich zu Zweitklässlern in Deutschland insgesamt?**

| | | weit unter-durchschnittliche 0 | | | | | 5 | | | | | weit über-durchschnittliche 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| f) | Kinder mit Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| g) | Kinder mit türkischem Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| h) | Kinder mit russischem Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| i) | Kinder ohne Migrationshintergrund | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |