



Why Did the Gender Gap in Adolescent Life Satisfaction Grow? Evaluating Methodological and Demographic Explanations

Julia M. Rohrer¹ · Richard McElreath² · Gregor Kachel³

Received: 5 June 2025 / Accepted: 10 June 2026
© The Author(s) 2026

Abstract

The city of Leipzig in Germany conducts large-scale school surveys of adolescents in secondary schools. While in 2010 and 2015 girls were somewhat less satisfied than boys, in 2023 this gender gap had doubled. Why? When asking such a reverse causal question, answers often focus on broad narratives, such as the psychological impact of social media. Here, we illustrate how to probe alternative explanations, such as demographic changes and methodological issues. First, we rule out that the observed pattern is a simple scaling artifact. Second, we find that the widening of the gender gap is much more pronounced among students with a migration background. This could plausibly be explained by a shift in the composition of the underlying population, with a strong increase in the proportion of Syrian students, and a relative decrease of Vietnamese students. Third, part of the increasing gender gap could potentially be attributed to survey mode: In 2023, for the first time, the survey was conducted on tablets—and unexpectedly, girls (but not boys) reported significantly lower satisfaction when surveyed on tablet rather than on paper. Lastly, beyond these patterns, we still find significantly widening gender gaps in satisfaction with leisure time activities and relationships to friends. Thus, there may be a substantive increase in the gender gap in satisfaction in those two domains that is not readily attributable to changes in demographics and survey mode.

Keywords Gender gap · Life satisfaction · Adolescence · Reverse causal questions · Survey mode · Demographic shift

✉ Julia M. Rohrer
julia.rohrer@uni-leipzig.de

¹ Wilhelm Wundt Institute for Psychology, Leipzig University, Leipzig, Germany

² Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

³ Institute for Psychology in Education, Leuphana University, Lüneburg, Germany

There have been widespread—and controversially debated—concerns that mental health may be declining on a large scale among young people in particular. These concerns pre-date the COVID-19 pandemic of the early 2020s but may have been further aggravated by it. Since then, one particular narrative—highlighted in books such as Jonathan Haidt’s “The Anxious Generation” (2024) and Donna Hackson Nakazawa’s “Girls on the Brink” (2022)—has been gaining public attention: Mental health has deteriorated in girls and young women in particular, and social media are often invoked as a major part of the explanation.

Against this backdrop, at the end of 2023, the city of Leipzig published a summary of their 2023 survey among students of all school tracks and types in the German secondary education system (Abel et al., 2023). The press release issued by the city highlighted that life satisfaction in adolescence had strongly declined since the last survey of the same age group. While in 2015, 72% of the youths reported to be satisfied or even very satisfied, that number decreased to 61% in 2023 (Jugendstudie, 2023: *Lebenszufriedenheit der Schüler sinkt deutlich* [Youth Survey 2023: Life satisfaction of students declines markedly], 2023). Furthermore, the decline was reportedly much more pronounced in girls (minus 14% points) than in boys (minus 7% points).

Generally, this local empirical finding fits the idea that well-being has been declining among girls in particular. But before accepting possible explanations based on simple narratives—which may result in biased interpretations (Sewall & Parry, 2024)—in this study we aim to analyze the data from Leipzig more thoroughly for additional hints as to *why* the gender gap might have grown. In particular, we want to highlight factors that may often receive less attention, such as methodological concerns and demographic changes in the underlying population. Such “alternative” explanations should be taken into account when explicitly asking reverse causal questions (sensu Gelman & Imbens, 2013): *Why* does Y occur? In general, studies more routinely aim to explicitly answer forward causal questions (“How does X affect Y?”), and so reverse causal questions are less familiar territory. There is less clarity about how to tackle them systematically, and they may not result in a single well-defined answer. Nonetheless, asking them may be worthwhile to identify gaps in our current (implicit or explicit) model of the world, which in turn allows us to refine our models and ultimately improves our ability to explain the world. With our particular focus on factors that are often neglected, we want to motivate researchers to more explicitly incorporate the specifics of the data at hand—such as how they were collected, from which population—into their models. We illustrate how to translate explanations into suitable statistical analyses and how to transparently discuss limitations of the data which are inevitable in practice. This manuscript is accompanied by a website containing the full analysis code and model results with additional annotations: <https://j-rohrer.github.io/Jugendstudienanalysen/>.

1 Method

1.1 Design of the Leipzig Youth Surveys

The Office for Statistics and Elections of the City of Leipzig, Germany, has conducted annual surveys of adult citizens on behalf of the city’s executive board since 1991. In the years 1993, 2000 and 2006, this survey was supplemented with a survey on leisure targeting children and youths starting from the age of 12 which was collected in schools (Heinemann

et al., 2011). From the 2010 survey onwards, this youth survey was supposed to be repeated every five years; however, after the 2015 survey, the COVID-19 pandemic necessitated a change of plans and so the 2020 survey was postponed to 2023 (Abel et al., 2023). These three most recent surveys—2010, 2015 and 2023—underlie our analyses.

1.2 The 2010 Youth Survey

Students between the ages of 12 and 17 were targeted within their respective schools and surveyed between October 18th and November 5th 2010 (Heinemann et al., 2011).¹ The initially targeted sample size was 3,000, which was distributed according to the actual distribution of students between the ages of 12 to 17 in the four types of schools included and the grade levels included:

- General secondary schools (*Mittelschule*, later renamed *Oberschule*), the lower secondary school track focused on practical education, grades 7 to 10.
- Grammar schools (*Gymnasium*), the upper secondary school track focused on preparation for higher education, grades 7 to 12.
- Vocational secondary schools (*Berufs-/Fachoberschule*), first and second year.
- Special needs schools (*Förderschule*), grades 7 to 10.

Schools were picked across Leipzig, prioritizing schools who had participated in earlier surveys to reduce organizational overhead and to ensure some degree of continuity. Within the selected schools, all students within eligible classes were assessed, even if they fell outside of the focal age range. This resulted in an initial sample of 3,459 students. Parents were informed upfront and had to consent to their children's participation; children could also deny to participate independently. Then, trained staff from the City of Leipzig went into classrooms during either school periods or free periods and handed out paper surveys. A total of 2,411 students (70% of the initial sample) filled out the survey, see Table 1.

Throughout our analyses, we limit ourselves to students attending either general secondary school or grammar school. These school types are considered *allgemeinbildend*, that is, providing a general education, and they are the second step after primary education in the standard course of education. We thus focus on students within the normative educational course in Germany. After secondary school, students may move on to a vocational secondary school; however, these tend to be more heterogeneous – including young people who still live at home, young people who already moved out and may have already started a family, and older people in second chance education. These students were excluded as their lifestyle and daily routine can differ significantly from that of the rest of the sample. Likewise, students at special needs school are a more heterogeneous population; an additional concern here is that the inclusion criteria may have changed between repetitions of the youth

¹ In 2010, the City of Leipzig also conducted a second parallel Youth Survey targeting people between the ages of 18 and 27 who had already left school and whom were contacted based on registry information. This survey was never repeated and the data were not included in our analyses.

Table 1 Overview of the Analyzed Leipzig Youth Surveys

Included in	2010		2015		2023	
	Data set	Analysis	Data set	Analysis	Data set	Analysis
Students	2,411	1,383	2,255	1,578	3,036	1,803
Classrooms	114	73	114	84	168	110
Schools	38	21	31	21	64	33
Schooltype						
General secondary	28%	46%	29%	41%	22%	35%
Grammar school	32%	54%	42%	59%	41%	65%
Vocational	36%	—	25%	—	32%	—
Special Needs	4%	—	3%	—	5%	—
Survey mode						
Pen and paper	100%	100%	100%	100%	34%	31%
Tablet	—	—	—	—	66%	69%
Gender						
Female	49%	51%	54%	55%	50%	50%
Male	50%	49%	46%	45%	46%	50%
Diverse or no response	— ^a	— ^a	— ^a	— ^a	3% ^a	—
No response	<1%	—	1%	—	1%	—
Age						
<12	<1%	—	0%	—	0%	—
12	8%	14%	2%	2%	4%	6%
13	18%	29%	14%	18%	13%	20%
14	14%	21%	18%	24%	15%	22%
15	11%	16%	18%	25%	14%	20%
16	12%	10%	16%	18%	12%	17%
17	12%	6%	15%	11%	10%	10%
18	7%	3%	7%	2%	9%	6%
>18	16%	—	10%	—	21%	—
No response	<1%	—	<1%	—	2%	—
Migration background ^b			Inferred migration background ^b		Migration background ^b	
No	89%	90%	90%	91%	84%	87%
Yes	9%	10%	9%	9%	12%	13%
No response	2%	—	1%	—	4%	—

^a In 2010 and 2015, only two gender response options were available. In 2023, a third response option labeled “diverse or no response” was added

^b In 2010 and 2023, students provided detailed information on where they and their parents were born. Migration background was coded as yes if either the student or both parents were born outside of Germany. In 2015, students were instead asked to indicate whether the language mainly spoken at home was German or not. Migration background was inferred to be yes if they indicated that they did not mainly speak German at home. See also Method section

surveys.² Thus, these students were also excluded. On the companion website, we provide additional information regarding the match between the target population and our sample.

1.3 The 2015 Youth Survey

Students between the ages of 12 and 17 were targeted within their respective schools; again, schools who had participated in previous youth surveys were prioritized. The initially targeted sample size was 3,000 students distributed over school types and classrooms according to the actual distribution in the population. One major change was that in grammar schools, only the grades 7 to 11 were included. This was because data were collected between May 7th and June 1st 2015, when the 12th grade students at grammar schools had already finished their final exit examinations (*Abitur*) and thus left school.³ The initial sample included 3,298 students, of which 2,277 (75%) filled out the paper survey according to the original report (Abel et al., 2015). The final data we received contained 2,255 students (Table 1).

1.4 The 2023 Youth Survey

Again, students between the ages of 12 and 17 were targeted within their respective schools. This time, to capture more variability, classrooms within schools rather than whole schools were selected and students at special needs schools were intentionally oversampled. The initially targeted sample size was 5,000 students which was expected to result in 3,000 usable responses. In fact, 3,050 filled out the survey according to the official report (Abel et al., 2023); the data we received contained 3,036 students (Table 1). In 2023, for the first time, two thirds of the students filled out the survey on tablets provided by the school, leading to a partial change in survey mode. The details of this will be discussed later when it becomes relevant to our analyses.

1.5 Measures of Interest

In all three youth surveys, students were asked eight pages of questions on various topics, such as their priorities in life, satisfaction with various areas of life, leisure time activities, athletic activities, drug use, and plans for the future. While many thematic blocks of questions re-occurred across surveys, many individual items were rephrased or completely exchanged, reducing the number of comparable variables across the three surveys. In the following, we only describe the measures analyzed for the purposes of the present study; any changes in phrasing across the years will be highlighted.

² In 2023, students at special needs school were intentionally oversampled, and the official report states that schools focusing on students with severe cognitive, social/communicative and emotional impairments (*Förderschwerpunkt geistige Entwicklung*) were excluded (Abel et al., 2023). No reference to such exclusion criteria can be found in the 2015 or 2010 reports (Abel et al., 2015; Heinemann et al., 2011).

³ For this reason, one may expect that in the sample, the relative proportion of grammar school students is somewhat lower in 2015 than in the other years. This, however, is not the case because (1) the study was still planned to preserve the proportion of school types of the general student population (rather than the proportion of school types of the student population *after excluding 12th grade of grammar school*) and because of slightly higher response rates among grammar school students, see Abel et al. (2015, p. 7).

1.6 Gender and Age

Gender and age: In 2010 and 2015, students were asked whether they were male or female. In 2023, a third response option labeled “diverse or no response” was added. In all three surveys, students were asked to write down their age.

1.7 Migration Background

In contemporary Germany, the term race is considered extremely fraught due to its association with Nazi ideology. Thus, surveys do not collect this variable but usually instead ask for “migration background”, that is, migration history of a person or their parents. In 2010, students were asked to indicate whether they, their father, and their mother were born in Germany and, if not, to write down in which country they were born instead. In 2023, students again indicated whether they, their father, and their mother were born in Germany; however, they were no longer asked to write down any birth countries. From this information, we generated a dichotomous variable representing migration background which is coded *yes* if either the students themselves or both of their parents were born abroad.

Unfortunately, in 2015, migration background was not directly collected. Instead, a simplified measure was chosen in which students indicated whether they mainly spoke German at home; if they did not, they were asked to write down their main language at home. From this, we inferred a migration background if the main language spoken at home is not German.

However, it is clear that this inferred migration background is not directly comparable to migration background in 2010 and 2023. Deutsche Welle (2019) reported based on micro-census data that in households with only one family member with immigrant background, German was the dominant language in 95% of the cases; these would be mapped consistently according to our coding decisions (no migration background, German main language at home). However, in households in which all members were immigrants, German was reportedly still the dominant language in 44% of homes; these would be mapped inconsistently according to our coding (migration background *yes* if sampled in 2010 or 2023; migration background *no* if sampled in 2015).⁴ Because of these potential inconsistencies, analyses that hinge on comparisons by migration background focus on the contrast between 2010 and 2023; 2015 is included in the figures for full transparency, but always visually de-emphasized.

⁴ This mapping issue should lead to an underestimation of the number of students with a migration background in 2015. We can confirm that this is the case by checking our numbers against official statistics issued by the City of Leipzig which reports the number of students with a migration background by school type for 2010 and 2015 (Bein 2023a, b; Dütthorn and Gugutschkow 2010a, b). According to these numbers, and given the distribution of school types in our final analysis samples, we would expect that the number of students with a migration background increased from 10.1% to 13.5% from 2010 to 2015. Accordingly, we would expect that the number of students with a migration background in our data increased at least somewhat. Yet, in our data, it slightly *decreased* from 10% in 2010 (migration background based on countries of birth) to 9% in 2015 (migration background inferred from language at home).

1.8 Satisfaction Items

In all three surveys, students were asked a block of questions about their satisfaction with various aspects of life. The question always opened with “How satisfied are you currently with...” and was always answered on a five-point response scale (1: very dissatisfied, 2: rather dissatisfied, 3: mixed, 4: rather satisfied, 5: completely satisfied). Some of the aspects of life asked for changed between the surveys, leaving the following eight items that are comparable over time:

- satisfaction with your life in general (*general life satisfaction*).
- satisfaction with your school grades (*school grades satisfaction*).
- satisfaction with your leisure time activities (*leisure satisfaction*).
- satisfaction with your relationships to friends (*friendship satisfaction*).
- satisfaction with your relationship to your father (*satisfaction with father*).
- satisfaction with your relationship to your mother (*satisfaction with mother*).
- satisfaction with the amount of money with which you have to get by (*financial satisfaction*).
- satisfaction with your housing situation (*housing satisfaction*).

The leisure time activities item switched from “how satisfied are you with your leisure time activities?” (2010) to “how satisfied are you with the opportunities for leisure time activities?” (2015, 2023). The school grades item switched the German word for school grades from *Zensuren* in 2010 and 2015 to the nowadays more common *Schulnoten* in 2023.

2 Statistical Models and Analysis Approach

Data analyses were conducted in the programming language R 4.4.2 (R Core Team, 2024) with the help of RStudio (Posit Team, 2023). We fitted various statistical models using, apart from base R, the packages *MASS* (Venables & Ripley, 2002) and *brms* (Bürkner, 2017) and then relied on *marginalEffects* (Arel-Bundock et al., 2024) to query the models. Our general approach followed the logic spelled out by Rohrer and Arel-Bundock (2026) which treats statistical models as counterfactual prediction machines to answer substantive questions.

In the first step, we always fitted a statistical model to predict life satisfaction (or some other satisfaction measure). The precise model specifications are explained and justified in the Results section, since the mapping between substantive question and statistical model is central to the present study. In the next step, we calculated gender gaps with the help of counterfactual comparisons: For every student, we predicted their life satisfaction if they were a girl and their life satisfaction if they were a boy. We then took the difference between these two values and averaged across students, which results in a *ceteris paribus* contrast. Later on, we additionally evaluated gender gaps in counterfactual scenarios in which the distribution of demographic variables is held constant across years; we achieved this by calculating counterfactual comparisons on new (hypothetical) data with constant demographics across the years. Furthermore, we also calculated average gender gaps for certain subgroups of interest, and contrasted these different averages. All confidence intervals, *p* values, and credible intervals were calculated with the help of *marginalEffects*, which can

return these numbers for a large number of quantities (see Rohrer & Arel-Bundock, 2026 for the underlying approach).

All analysis code can be found on the companion website (<https://j-rohrer.github.io/Jugendstudienanalysen/>). Unfortunately, we are not able to share the data from the Leipzig youth surveys; these data are only available upon request from the data holder (City of Leipzig) due to legal requirements.

3 Results

In the following, we first establish the central finding that we seek to explain, a growing gender gap in life satisfaction. Then, we consider methodological and demographic explanations: Could this pattern actually just be a scaling artifact resulting from the life satisfaction measure used? Could it be that the demographic composition of the sample has changed over the years? Could the pattern be attributable to changes in survey mode in the most recent data? These questions lead us to continuously expand and refine our statistical model. Lastly, to explore potential *substantive* explanations for expanding gender gaps, we apply our final model to additional items measuring satisfaction with various aspects of life.

3.1 Establishing the Explanandum: Decreased Life Satisfaction in Girls in 2023

First, we checked whether our analysis supported the central observation of a widening gender gap in life satisfaction in 2023. The press release by the city of Leipzig simply compared the percentage of youths who used the highest two response options for life satisfaction across the years, additionally reporting separate numbers for girls and boys (Jugendstudie, 2023: *Lebenszufriedenheit der Schüler sinkt deutlich [Youth Survey 2023: Life satisfaction of students declines markedly]*, 2023). Instead of simply comparing the percentages, we first fitted a simple linear model predicting life satisfaction from the categorical predictors *gender*, *survey year*, and their interactions. This approach allowed us to increase model complexity later on to probe various explanations of the pattern.

In all three years, girls reported lower satisfaction than boys. The gender gap grew from 0.22 scale points in 2010 (95% confidence interval: [0.13, 0.30]) to 0.28 scale points in 2015 (95% CI: [0.20, 0.36]) to 0.48 scale points in 2023 (95% CI: [0.41, 0.56]). Expressed in standard deviations of life satisfaction across all observations, the gap grew from 0.26 *SD* to 0.33 *SD* to 0.57 *SD*. The widening of the gap from 2010 to 2015 in itself was not statistically significantly different from zero (difference of 0.06 scale points, 95% CI: [0.06, 0.18]), but the widening in the gap from 2015 to 2023 was (0.20 scale points, 95% CI: [0.09, 0.31]). In total, from 2010 to 2023, the gender gap grew by 0.26 scale points (95% CI: [0.15, 0.38]), or by 0.31 *SD* of the total variability in life satisfaction. (Fig 1)

3.2 Could this Pattern Be a Scaling Artifact?

So far, we analyzed life satisfaction as if it was a continuous variable – which it is not. Data were collected on a five-point response scale, resulting in an asymmetric distribution (Fig. 2). In particular, while many students reported a fairly high score of 4, surprisingly few chose the highest response option, 5. This corresponds to observations for life satis-

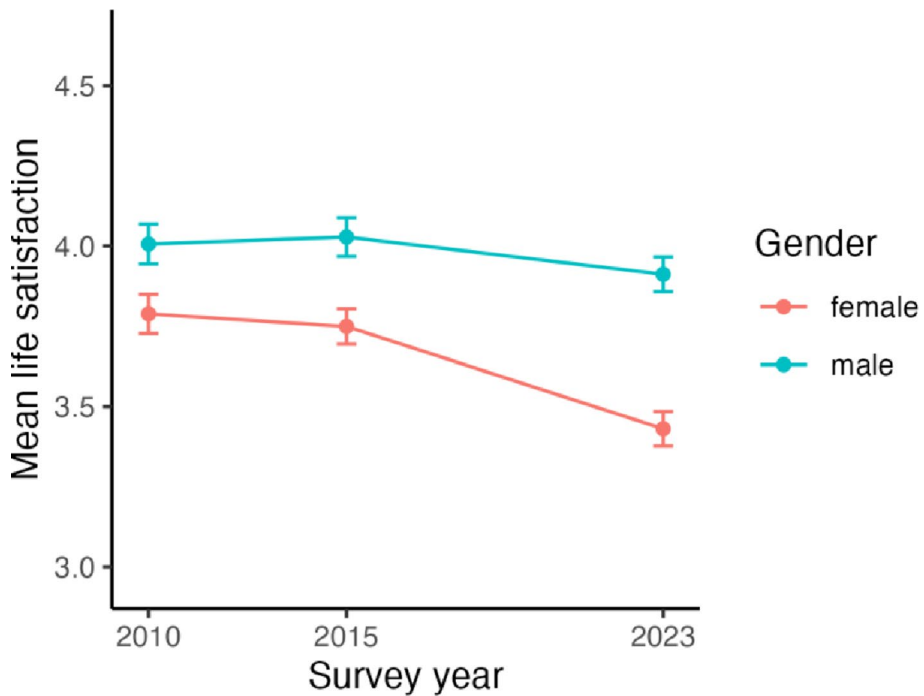


Fig. 1 Mean life satisfaction in boys and girls over time. Scale from 1 to 5, error bars are 95% confidence intervals

faction scales that go all the way up to 10 – despite high average values, usually only few respondents report a 10, and cognitive interviews suggest that for many respondents 8 out of 10 constitutes the “top of the scale”, with higher values being hard to conceive for them (Fabian et al., 2024).

In general, categorical response scales can cause trouble if the psychological differences between response options are not the same – for example, if the difference between a 2 (rather dissatisfied) and a 3 (mixed) is smaller than the difference between a 4 (rather satisfied) and a 5 (completely satisfied). In those scenarios, simply modeling the mean of a continuous (normal) distribution—as we did above—may not accurately capture central tendencies. Most importantly for the present purposes, uneven differences between response options could also dissolve the observation of a widening life satisfaction gap. This pattern is an interaction effect, and interactions are scale-dependent: they can appear, disappear, or even reverse depending on how the data are scaled (Rohrer & Arslan, 2021).

Consider the hypothetical scenario depicted in Fig. 3, Panel A. This shows the density curve of latent life satisfaction as a continuous variable which is normally distributed. When respondents have to report their (continuous) life satisfaction on the five-point response scale presented to them, their response is determined by certain cut-off values (so-called thresholds) that translate the unobserved continuous variable into an observed categorical variable. Two horizontal lines mark the gap in latent life satisfaction between a girl and a boy in 2010 and 2023. In both years, this line has the same length, so the difference in latent life satisfaction remains the same. However, in 2023, the whole line is shifted to the left.

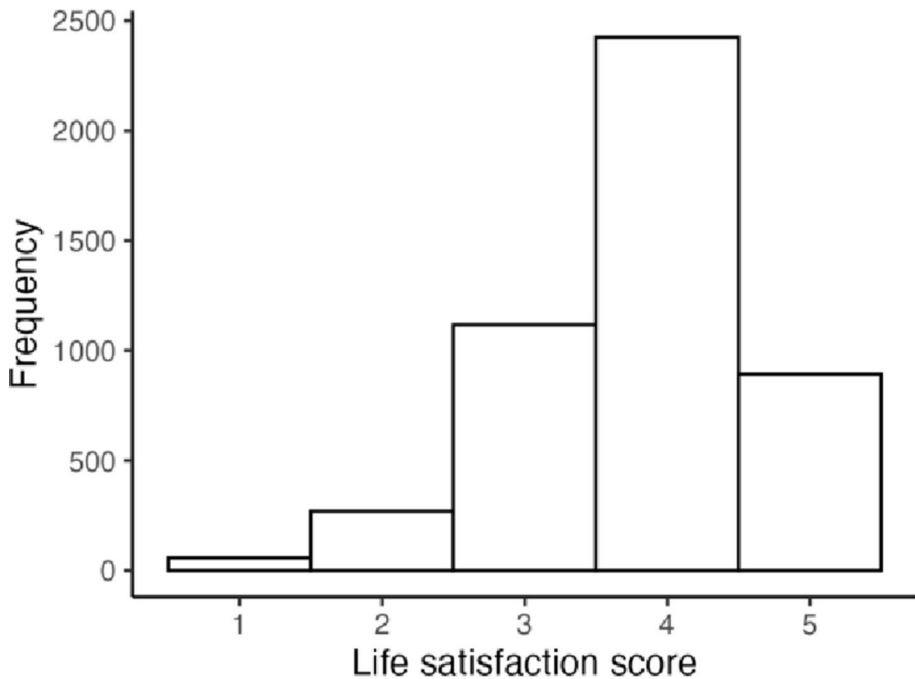


Fig. 2 Distribution of life satisfaction scores across all youth surveys

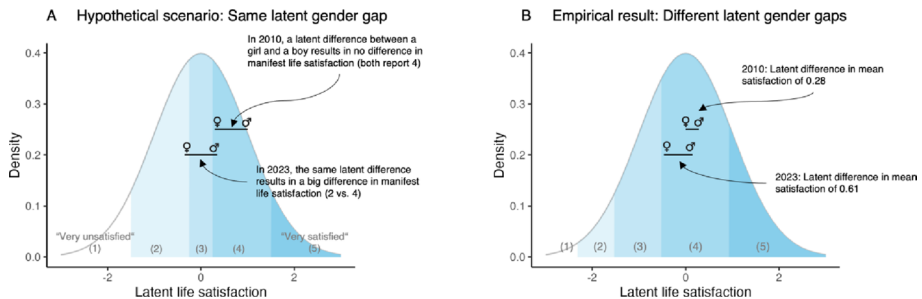


Fig. 3 In principle, the widening gender gap could be a scaling artifact (Panel A), but the data do not support this conclusion (Panel B)

This affects how the latent life satisfaction scores and their differences are translated into observed categorical values. In 2010, the boy and the girl fall into the same category and report a 4; thus, there would be no observed difference in life satisfaction on the five-point scale. In 2023, the boy and the girl fall into different categories – while the boy ends up reporting a 4, the girl ends up reporting a 2. We thus end up with a much larger observed difference in life satisfaction, despite the underlying latent life satisfaction difference remaining the same. Following the same logic, the widening gender gap we observed earlier could actually be a scaling artifact insofar that on the latent level, gender differences in life satisfaction may have remained unchanged.

To test whether this is the case, we again regressed life satisfaction on gender, survey year, and their interaction, this time using an ordered probit model. These models correspond to the mapping process depicted in Fig. 3: Latent life satisfaction is assumed to follow a normal distribution that is translated into an ordinal categorical variable according to thresholds which are estimated empirically. The results, including the estimated thresholds, are depicted in Fig. 3, Panel B. In line with the observed distribution of responses (Fig. 2), the thresholds separating a life satisfaction score of 4 from lower or higher scores were quite far apart – a large chunk of the area under the density curve falls into this category. Importantly, on the estimated latent continuous life satisfaction variable, we still observed a widening gender gap; starting from 2010, the gender gap widened from 0.28 to 0.61 in 2023, or by 0.33 *SD* of the latent variable (95% CI: [0.18, 0.48]). This closely matches the results from initial linear regression model, where we observed a widening of 0.31 *SD* of life satisfaction across all observations. Thus, we can discard the idea that the widening gender gap is a simple scaling artifact.⁵

Ordinal modeling thus resulted in conclusions that are very similar to those resulting from linear regression models. For the sake of simplicity, further analyses thus always started from the linear model, although we additionally conducted a final robustness check with an ordinal model.

3.3 Can Demographic Variables Explain Widening Gender Gaps?

Another explanation for the widening gender gap could be changes in demographic variables. For example, maybe the gender gap in life satisfaction varies by age (e.g., larger gender gap among older students) and maybe the age composition of the sample was different in 2023 (e.g., more older students). Such a change in the composition of the sample may either be due to changes in the composition of the underlying population (e.g., students in Leipzig indeed became older on average) or due to changes in the sampling procedure (e.g., older students became more likely to be included in the study). In either case, such changes could contribute to an explanation of the widening gender gap. We considered three demographic variables: students' age, the type of school they attend, and their migration background.

Age can be included in models in different ways. For example, one may include it as a continuous linear predictor which forces any age trend to be linear; one may include it so that each year of age becomes its own category, which does not impose any functional form but may render some estimates quite imprecise if age groups are small. Here, we chose something in between those two extremes and include age basis splines with 3 degrees of freedom (the default setting). This allowed our estimates to pick up the underlying functional form in a flexible manner but still results in smooth trajectories, so that the individual estimates do not become overly imprecise. Note that the precise modeling of age does not impact conclusions regarding the widening of the gender gap, as demonstrated on the companion website. See Rohrer and Arel-Bundock (2026) for a more extensive discussion of how to best include potential confounders.

⁵ Notice that this analysis assumes that girls and boys use the scale in the same manner. As demonstrated in analyses on the companion website, we can relax this assumption and allow the thresholds to vary by gender in Bayesian ordered probit models. According to these analyses, the gender gap grew by 0.30 *SD* of the latent variable (95% credible interval: [0.15, 0.46]).

More importantly, as the widening that we want to “explain away” is an interaction (gender*year), we included the interactions between all demographic variables and gender and year – probing whether interactions are confounded requires interaction controls (Simonsohn, 2019; Yzerbyt et al., 2004). Beyond this, we also included the three-way interaction between each demographic variable and gender and year (e.g., we include migration background*gender*year). Adding this flexibility allowed us to rule out that the observed pattern can be attributed to changes in how demographic variables related to the outcome in either boys or girls over time.

Our final model in this step is thus a linear regression in which we regressed life satisfaction onto gender, survey year, and, to account for demographic differences, age splines, schooltype (categorical variable with two levels) and migration background (categorical variable with two levels). All variables in the model were interacted with gender and survey year, and also interacted with the gender*survey year interaction. Standard errors were clustered at the level of the classroom to account for nesting in the data, which also applied to all subsequent analyses we conducted.

3.4 Do Demographics Explain Away the Widening Gender Gap?

To find out whether changes in demographic variables explained the widening gender gap, we calculated model-implied gender gaps for the individual survey years *in a world in which the distribution of the demographic variables was exactly the same in every survey year*. If, in such a world, our model does not imply any widening of the gender gap, we have successfully explained it away.

We can, of course, imagine many different worlds in which the distribution of the demographic variables did not change over the years – and since we allowed all predictors to interact, the gender gap and how it changes over time may look somewhat different across worlds. One sensible choice could be a world in which the distribution of demographic variables is always the distribution of our baseline survey year, 2010. For such a world, our model implied a gender gap of 0.22 scale points in 2010, 0.31 scale points in 2015, and 0.50 scale points in 2023. From 2010 to 2023, the gap thus widened by 0.27 scale points (95% CI: [0.15, 0.39]) or by 0.32 *SD*, which is virtually the same pattern that we observed when ignoring demographic variables. Thus, changes in the demographic variables do not explain the widening gender gap. We arrive at the same conclusion when we evaluated the model in a world in which the distribution of demographic variables equals the distribution in 2015 or 2023; in fact, in these worlds, we would conclude that the widening is even slightly more pronounced (widening from 2010 to 2023 at 2015 distribution: 0.33 scale points or 0.39 *SD*; at 2023 distribution: 0.31 scale points or 0.37 *SD*).

3.5 Do Demographics Modify the Widening of the Gap?

Our model also allowed us to investigate whether the widening of the gender gap is *modified* by demographic variables. For example, maybe the gender gap widens in particular among older students, or in particular among grammar school students. Considering age and school type, results descriptively suggested some trends, although these were estimated with a lot of uncertainty and were not statistically significant. This may be unsurprising from the perspective that such patterns would be three-way interactions (gender*survey year*age

or school type). Descriptively, the widening of the gender gap increased with age until it peaked among students aged 14 to 16 and then potentially declined again. The widening of the gender gap was also somewhat more pronounced in general secondary schools than in grammar schools. Both of these patterns may plausibly just reflect random variability.

Results were a lot clearer when considering migration background as a modifying factor. Among students without a migration background, from 2010 to 2023, the gender gap widened by 0.22 scale points (95% CI: [0.10, 0.35]); among students with a migration background, the gender gap widened by 0.72 scale points (95% CI: [0.29, 1.14]) – amounting to a *difference* in the widenings of the gap of 0.49 scale points (95% CI: [0.05, 0.94]). This difference is descriptively large, which is why in subsequent analyses, we always considered the widening of the gap separately for students with and without migration background. Figure 4 summarizes the preliminary results, contrasting the model-implied means over time by migration background while holding the remaining demographic variables (age and schooltype) constant at their 2010 distribution. Results for 2015 are greyed out because migration background in 2015 had to be inferred from the language spoken at home, making it hard to compare this year to the other years.

3.6 Could this Pattern Result from Changes in Survey Mode?

In 2023, an important change happened that we have ignored so far: For the first time, tablets were introduced as an assessment mode. This change in survey mode may have affected how students reported their life satisfaction. Importantly, not all students used tablets in 2023 with 31% of participants still filling out the survey on paper. First, in some schools, leadership decided against tablets, so that whole schools filled out the survey on paper. Second, sometimes not enough tablets were available because multiple classrooms were surveyed at the same time, or because tablets were needed for other classroom activities elsewhere, so that whole classrooms filled out the survey on paper. Third, sometimes there were not enough tablets for all students in the classroom, in which case some students in a classroom filled out the survey on paper; or a battery failed, in which case individual

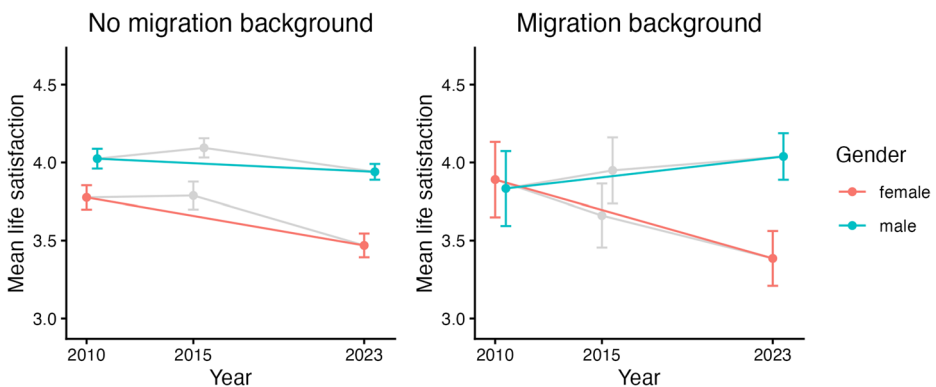


Fig. 4 Model-implied life satisfaction in boys and girls over time, separately for students with and without a migration background. Scale from 1 to 5, error bars are 95% confidence intervals, demographic variables held constant at their 2010 distributions. Note that in 2015, migration background was inferred from whether or not German was the main language students spoke at home

students switched to paper. The distribution of survey mode was virtually the same across genders (girls: 32% paper vs. 68% tablet; boys: 30% paper vs. 70% tablet, $p = .484$).

The variability in survey mode outlined above allowed us to examine its association with life satisfaction and whether this relationship differed by gender. For this purpose, in the 2023 data, we regressed life satisfaction onto survey mode, gender, and their interaction. This model predicted that girls reported higher satisfaction on paper ($M = 3.55$) than on tablet ($M = 3.38$), with a difference of 0.17 scale points (95% CI: [0.06, 0.29]). In contrast, boys reported the same average satisfaction on paper ($M = 3.91$) as on tablet ($M = 3.91$). The offset in the survey mode differences between the genders (i.e., the interaction between gender and survey mode) was 0.17 scale points [0.02, 0.32].

3.7 Two Explanations for the Survey Mode Differences

So, girls appeared unhappier on tablet than on paper, boys did not. *If* these differences reflect causal effects of the assessment mode, they could partially explain the widening gender gaps – reporting on tablet led to lower responses in girls but not in boys and thus to a widening of the gender gap, and tablets were only introduced in 2023. In that case, it would be appropriate to call the induced pattern an artifact, as it was caused by arbitrary changes in survey design rather than by substantive changes over time. To check whether there is a widening of the gender gap beyond such mode artifacts, one would need to limit the analysis over years to students who used the same survey mode – here, this would necessarily be paper, since tablets were only used in the last survey wave.

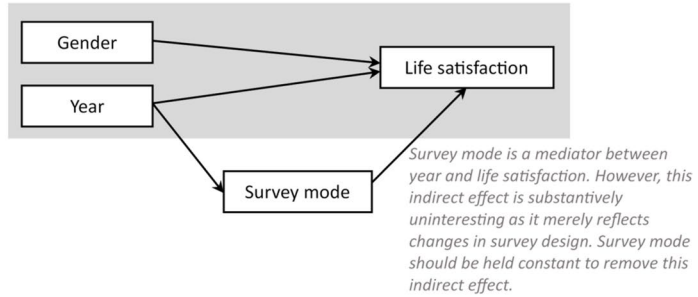
But it is not a given that the observed survey mode differences are causal effects of the assessment, considering that survey mode was not randomized. In particular, schools or classrooms that used paper rather than tablets may have been systematically different with respects to third variables (e.g., socio-economic status of the neighborhood), and these third variables in turn could result in different gender gaps, without survey mode actually having any causal effect on top. In this case, we could simply ignore survey mode as it does not induce any artifacts. In fact, in this scenario, limiting the analysis over years to students who used the same survey mode – paper – could *induce* bias. The intuition behind this is that we would compare the *full* 2010 and 2015 samples to a subsample of the 2023 data which systematically differs in the third variables (e.g., socio-economic status of the neighborhood) that led to paper being used despite the availability of tablets. Figure 5 presents causal graphs visualizing the two different scenarios (see Elwert & Winship, 2014 for more on collider bias; Rohrer, 2018 for a general introduction to such graphs).⁶ See Tomova et al. (2026) for a more general discussion of when (not) to condition on survey mode.

The central questions to determine how to analyze our data—in particular, whether we *should* limit the analysis to responses on paper—are (1) whether or not students who filled out the 2023 survey on paper are systematically different, either on observed or unobserved variables, and (2) whether there would still be an association between survey mode and the gender gap in life satisfaction after accounting for *all* these possible differences. In a hypothetical scenario in which we accounted for all possible confounders and still observed

⁶ In practice, survey mode may also *both* affect life satisfaction and be confounded with life satisfaction. In such a scenario, including only one response mode in the analysis may simultaneously remove one type of bias and induce another one. From that angle, the following analysis can be taken as an attempt to figure out which type of bias would be worse and thus which analysis is to be preferred.

Scenario 1: Survey mode causally affects the gender gap in life satisfaction

Of central interest is how the gender gap in life satisfaction changes over time, that is, how do gender and year jointly affect life satisfaction?

**Scenario 2: Survey mode is confounded with the gender gap in life satisfaction**

In this scenario, students who filled out the survey on paper systematically vary on third variables which in turn affect life satisfaction. Holding survey mode constant induces a non-causal association between year and third variables which in turn biases conclusions regarding the joint effects of gender and year.

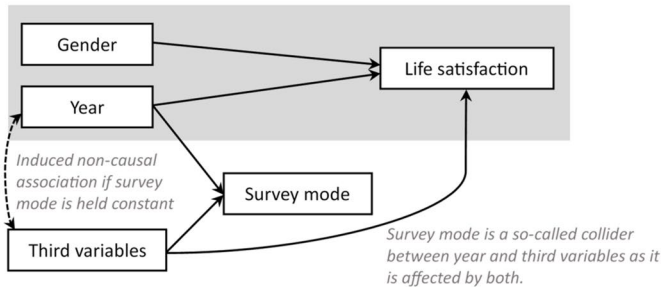


Fig. 5 Two different explanations for the observation that in 2023, the gender gap in life satisfaction varied between the two survey modes (paper and tablet)

an association, we would conclude that survey mode causally affects responses (Fig. 5, Scenario 1), and it would be appropriate to limit analyses to a single survey mode.

3.8 Accounting for Observed Differences in Demographic Variables

In our 2023 data, we indeed find that students who filled out the survey on paper (rather than on tablet, which would have been the default) systematically differed. In particular, only 11% of students at general secondary schools used paper, as opposed to 41% of students at grammar schools (χ^2 test: $p < .001$). Furthermore, students who filled out the survey on paper were slightly younger ($M = 14.55$ years) than students who filled out the survey on tablets ($M = 14.83$ years, Welch Two-Sample t-test for the difference: $p < .001$). We observed no significant association between survey mode and whether or not students had a migration background ($p = .32$).

To check whether these differences in demographics can explain why girls (but not boys) were more satisfied on paper than on tablets, we re-ran the previous regression comparing the two survey modes, this time additionally including the predictors age (included again with the help of splines), schooltype, and migration background, as well as their interactions with survey mode and gender. Using this model, we calculated the counterfactual contrasts by survey mode (predicted life satisfaction if everybody filled out the survey on paper vs. predicted life satisfaction if everybody filled out the survey on tablet) separately by gender. Again, our model implied that girls were happier on paper than on tablet (difference of 0.16

scale points, 95% CI: [0.03, 0.29]) but boys were not (difference of -0.01 scale points, 95% CI: $[-0.11, 0.10]$), with a difference between the differences of 0.17 scale points (95% CI: $[0.01, 0.32]$). These unchanged conclusions support a scenario in which survey mode indeed causally affects life satisfaction (Fig. 5, Scenario 1)—unless there are additional possibly unobserved confounders affecting both survey mode and life satisfaction (Fig. 5, Scenario 2).

3.9 Accounting for School-Level Confounding with School-Fixed Effects

Given that survey mode was not randomized, it is of course still possible that there are unobserved confounders – classrooms that opted out of tablet usage may have been different in subtle ways which in turn explain why girls in them were more satisfied. We cannot rule out this possibility in general due to the study design. However, it turns out that in 12 schools, both survey modes were used. This, in turn, allows us to conduct additional analyses that can rule out unobserved confounders on the level of the school.

More specifically, we modified our previous linear regression model (analyzing associations with survey mode in the 2023 data) with the focal predictors gender and survey mode (as well as their interactions with schooltype, age, and migration background) and included (1) dummy variables representing the 12 different schools to estimate school-specific intercepts and (2) the interaction between these dummies and gender to allow for unique gender gaps in each of the schools. This fixed-effects model uses only within-school information to evaluate differences in life satisfaction by survey mode and gender. If this model still implies a gender-specific effect of survey mode, we can rule out that this is due to unobserved confounding *on the school level*. We again found that girls are happier on paper than on tablet, this time by 0.23 scale points (95% CI: $[0.07, 0.39]$). Again, this does not seem to be the case for boys, who report descriptively slightly lower well-being on paper (-0.05 scale points, 95% CI: $[-0.17, 0.06]$). The difference between these differences amounts to 0.29 scale points (95% CI: $[0.06, 0.51]$).

It follows that unobserved school-level variables—such as neighborhood SES or other neighborhood characteristics—cannot explain away the particular association between survey mode and life satisfaction that we observe in girls. There could still be classroom-level confounders—for example, maybe teachers who particularly disliked digital devices also happened to behave in certain ways that positively affected girls’ well-being. As we do not consider such an explanation particularly plausible, here, we are going to instead assume that survey mode is exogenous with respects to the explanandum, and that the association between survey mode and life satisfaction actually reflects a causal effect: filling out the survey on the tablet made girls (but not boys) report lower values than on paper. In this scenario, survey mode should be held constant when comparing the years to prevent it from introducing differences between the years that are not substantively relevant (Fig. 5, Scenario 1). Holding it constant, in turn, required us to limit the analyses to students who used paper, since nobody filled out the survey on tablet in 2010 and 2015.

3.10 Re-assessing the Explanandum Only for Responses on Paper

We now return to the model which we used to evaluate the widening of the gender gap while holding constant the distribution of age, schooltype, and migration background across the

years (allowing each of these variables to interact with the survey year and gender). This time, we estimated it on data from 3519 students who filled out the survey on paper (all students from the 2010 and 2015 surveys, as well as the 558 students from the 2023 survey who used paper), and we held all covariates at their 2010 distributions.

In this model specification, the gender gap widened from 0.22 scale points in 2010 to 0.31 scale points in 2015 to 0.41 scale points in 2023. This constitutes a widening by 0.19 scale points (95% CI: $[-0.02, 0.40]$), or 0.22 *SD* of life satisfaction across all observations. Note that this widening is not as stark as the widening implied by the same model fitted on the whole sample (including both survey modes), which was 0.31 scale points. Thus, assuming that the association with survey mode actually does reflect a causal effect of survey mode on life satisfaction in girls (Fig. 5, Scenario 1), changes in survey mode explained away about 30% of the widening.⁷

Before, we had already seen that widening of the gender gap varies by migration background, with a much more pronounced widening for students with a migration background. As can be seen in Fig. 6, this is still the case when we limited the sample to students who filled out the survey on paper. In fact, in this model, the widening of the gender gap for students without migration background from 2010 to 2023 has shrunk to a mere 0.11 scale points (95% CI: $[-0.11, 0.33]$) or 0.13 *SD* of life satisfaction across all observations; the two lines in the left panel of Fig. 6 are close to parallel. Restricting the survey mode to paper halved the gap that resulted from the same model fitted to the whole sample. In contrast, the widening of the gender gap among students with a migration background was still very strong at 0.88 scale points (95% CI: $[0.27, 1.48]$); in fact, it was descriptively slightly larger than without restricting survey mode. This corresponds to a widening of the gender gap by 1.03 *SD* of life satisfaction across all observations. Directly contrasting the widening of the gap between the two groups, it was stronger in students with a migration background by

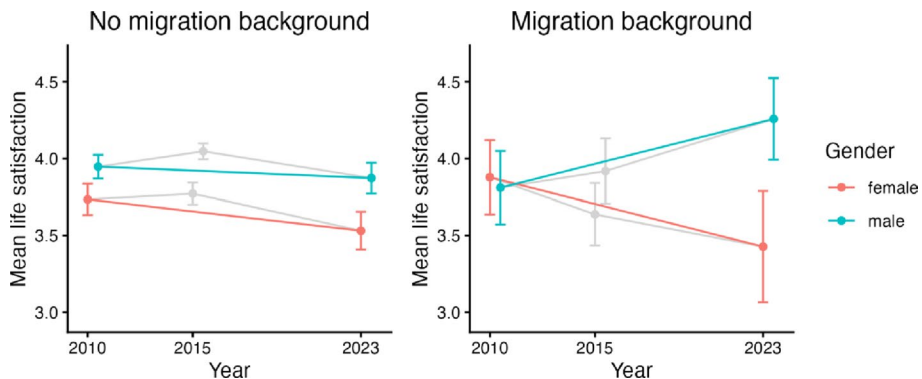


Fig. 6 Model-implied life satisfaction in boys and girls over time, separately for students with and without a migration background, only including responses on paper. Scale from 1 to 5, error bars are 95% confidence intervals, demographic variables held constant at their 2010 distributions. Note that in 2015, migration background was indirectly inferred from whether or not German was the main language spoken at home

⁷ Assuming instead that the association with survey mode reflects differences due to unobserved third variables (Fig. 5, scenario 1), the less strong widening of the gender gap on paper could be attributed to the fact that the 2023 sample, conditional on survey mode, differs on those unobserved third variables.

0.77 scale points (95% CI: [0.14, 1.41]). We checked the robustness of this pattern in an ordinal model (reported on the companion website), which supported the conclusion that the widening of the gender gap was stronger in students with a migration background (by approximately 1 *SD* of the underlying latent variable; 95% credible Interval: [0.29, 1.69]).

3.11 What about Alternative Satisfaction Measures?

So far, we only tried to understand the widening gender gap in overall life satisfaction. But the surveys also contained a number of more specific satisfaction items, which may point to potential *substantive* explanations. We additionally analyzed these items, relying on our final model specification that we developed for overall life satisfaction (holding constant the distributions of age, schooltype, and migration background; limiting analyses to respondents who filled out on paper; breaking results down by migration background). The results are depicted in Fig. 7, see companion website for detailed numerical results.

For grade satisfaction, we actually observed a gender gap among students with a migration background *in favor of girls* in 2010 that has closed in 2023. Considering both satisfaction with leisure time activities and satisfaction with the relationship to one's friends, we observed widening gender gaps in favor of boys, both among students with and without a migration background. Considering satisfaction with one's relationship to both parents, we observed widening gender gaps in favor of boys exclusively among students with a migration background. Lastly, considering both financial satisfaction and housing satisfaction, we observed widening gender gaps exclusively among students with a migration background. To sum it up, for some domains—satisfaction with school grades, satisfaction with one's relationship to one's parents, satisfaction with one's financial situation and one's housing—we observed a widening of gender gaps in favor of boys or a corresponding closing of a gender gap in favor of girls exclusively among students with a migration background. At the same time, there were at least two domains—satisfaction with leisure time activities and satisfaction with one's relationships to one's friends—for which we observed widening gender gaps in favor of boys among both students with and without a migration background.

4 Discussion

Why did the gender gap in life satisfaction among adolescents in Leipzig increase? Our analyses suggest that the pattern cannot easily be explained away as a scaling artifact, nor can it be explained by changes in underlying demographic variables (age, school type, migration background). However, we did observe a much more pronounced widening of the gender gap among students with a migration background. In the last wave of data, we also observed a wider gender gap when looking at responses on tablets (as opposed to paper). We suggested that this reflects effects of survey mode which should not be interpreted as part of any underlying substantively meaningful trend over time. Once we take into account survey mode effects, the life satisfaction gender gap among students without a migration background may actually not open at all. However, even when taking into account all of this, we still find widening gender gaps in satisfaction with leisure time activities and in satisfaction with relationships to friends, among both students with and without migration background.

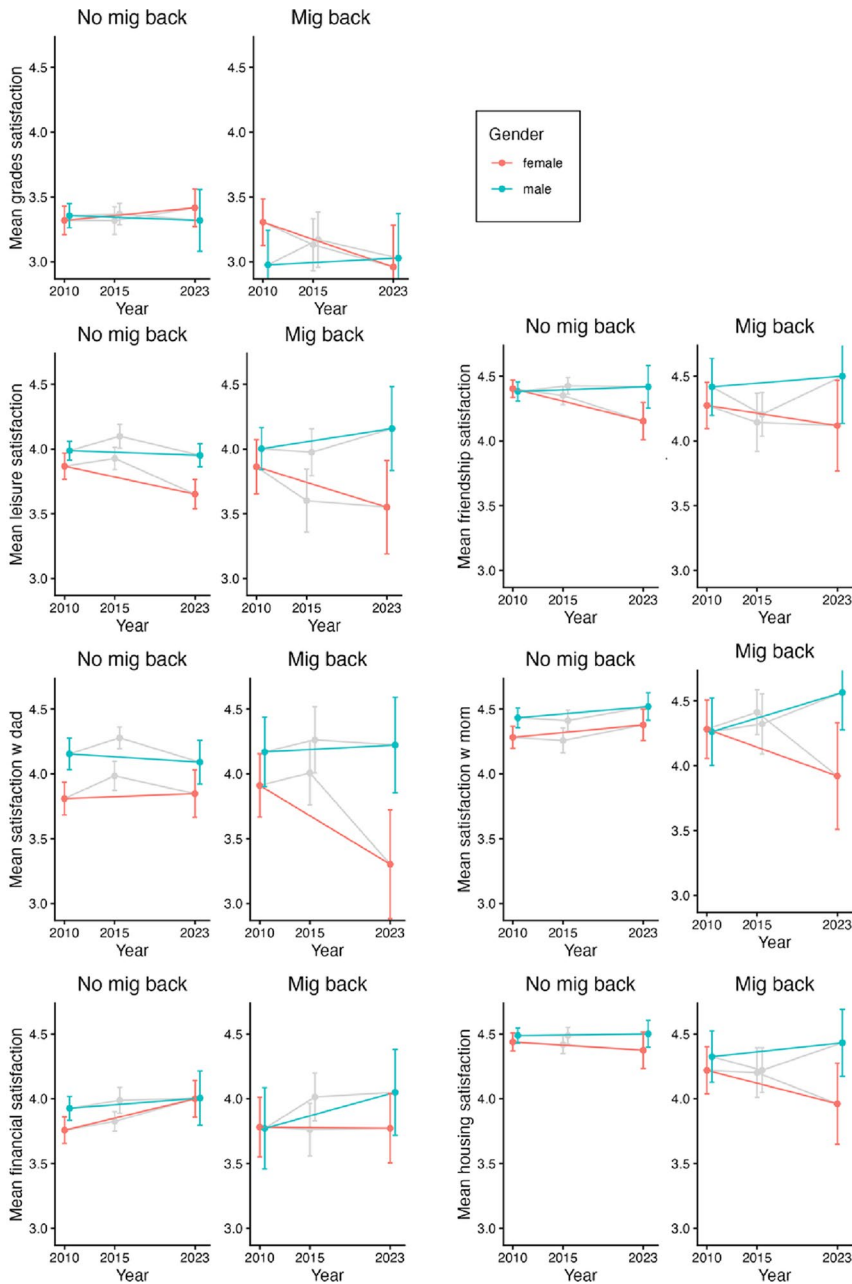


Fig. 7 Model-implied satisfaction with various aspects of life in boys and girls over time, separately for students with and without a migration background (mig back), only including responses on paper. Scale from 1 to 5, error bars are 95% confidence intervals, demographic variables held constant at their 2010 distributions. Note that in 2015, migration background was indirectly inferred from whether or not German was the main language spoken at home

We first discuss our specific findings for Leipzig in more detail, before zooming out to discuss which generalizable insights—if any—result from our investigation.

4.1 Migration Background: Pronounced Changes in the Underlying Population

Why would the gender gap widen in such a pronounced manner among students with a migration background? One explanation to consider is that something happened in Leipzig that made the city a worse place for girls but not for boys with a migration background. However, the explanation that we consider both more parsimonious and more plausible here is that the widening gender gap mainly reflects changes in the composition of the underlying population.

At schools providing a general education (such as the schools considered in this study), in Leipzig in 2010, a fairly large share of students with a migration background had Vietnamese parents (Düttborn and Gugutschkow 2010a, b). This is due to specific historic circumstances: In the German Democratic Republic (now East Germany, including Leipzig), Vietnamese “contract workers” migrated in the context of treaties between socialist countries to tackle shortages of skilled labor. Until 1988, female workers who became pregnant faced deportation unless they aborted their pregnancies („*Vertragsarbeiter*“ in der DDR, n.d.). Despite legal uncertainties following the fall of the Berlin wall in 1989 which negated the previous bilateral treaties („*Vertragsarbeiter*“ in der DDR, n.d.), many Vietnamese contract workers went on to have children after deportation was no longer a guaranteed consequence. The very oldest cohort of these children would have been around 20 years old in 2010 and thus past school age during the first data collection. However, those born slightly later are represented in our 2010 data: When a student reported that a parent was not born in Germany, Vietnam was the most frequently named country of birth.

The relative composition of the population with migration background in Leipzig then started to shift with strongly increasing numbers of refugees from Syria, Iraq, and Afghanistan in 2015, and refugees from Ukraine starting in 2022 (Bein 2023a, b; Düttborn and Gugutschkow 2010a, b, 2016a, b). Comparing the number of migrants in Leipzig based on their origin, the number of migrants with a Vietnamese background did increase somewhat (2009: 2 758, 2011–2015: 2 742, 2022: 3 882), but this is nothing in comparison to the more than 20-fold increase in the number of migrants with a Syrian background (2009: 447, 2011–2015: 516, 2022: 12 104). The number of migrants with a Ukrainian background also strongly increased (2009: 3 038, 2011–2015: 3 046, 2022: 12 970), although it is unclear how many adolescents would have already attended regular German secondary schools in 2023 due to the language barrier. Ideally, we would know the precise migration backgrounds of the students in our 2023 data. These data unfortunately were not collected, but based on the shifts in the underlying population, it is plausible that the relative share of students with a Vietnamese background declined, whereas the number of students with a background from most noticeably Syria and potentially also other Muslim majority countries increased.

Thus, the population of Leipzig students with a migration background in 2010 likely looked quite different from the population of Leipzig students with a migration background in 2023, with very different representation of various countries of (parental or own) origin. We consider it plausible that this is the main factor driving the widening gender gap among students with a migration background. Groups with larger gender gaps—for example, students who recently fled from Muslim majority countries—may simply be more strongly

represented in the later sample. We considered testing this hypothesis with the data available to us but discarded the idea due to small sample sizes, in particular when taking into consideration that detailed migration background information was only available in 2010 (with only a handful of students who plausibly recently fled from a Muslim majority country back then).

That being said, we can still provide a back-of-the-envelope calculation to gauge what our explanation would imply about differences between students with different migration histories. This calculation is necessarily loaded with assumptions. Let us assume that in 2010, 10% of the students with a migration background recently fled from a Muslim majority country, whereas 90% have some different migration history. With a tenfold relative increase of the number of the former group, in 2023 we would end up with 53% who recently fled a Muslim country vs. 47% who did not. Assuming that the gender gap in both groups remained the same over time, and assuming that the gender gap across all students with a migration background was 0 in 2010 and 0.7 scale points in favor of boys in 2023, we end up with two equations that we can solve for the gender gaps in the two groups ($0.1a + 0.9b = 0$; $0.52a + 0.47b = 0.7$, with a being the gender gap among students who recently fled a Muslim country and b being the gender gap among students with some different migration background). We end up with an estimate of a gender gap of almost 1.5 scale points in favor of boys among students who recently fled a Muslim country versus a smaller gap of 0.16 scale points in favor of girls among the other students with a migration background. This is a very stark contrast that may be considered unrealistically large, in which case one may want to adjust assumptions. For example, maybe our analyses happen to overestimate the growth of the gender gap among students with a migration background, or maybe there are other factors that lead to a genuine widening of the gap among students with a migration background specifically, even when holding country of origin constant.

We could not explicitly test whether demographic change tells the full story here, and we also did not explicitly test any hypotheses for *why* the gender gap may vary between different groups of students with a migration background. However, the very strong widening of the gender gap in satisfaction with the relationship to one's parents from 2015 to 2023 may suggest that the relationship between adolescents and their parents is part of the story—maybe adolescent girls with parents from more gender-conservative countries such as Syria experience more friction with their families than their male counterparts.

The literature provides various lines of research pointing toward such potential frictions. For example, Nilles et al. (2023) compared gender role attitudes among German adolescents without a migration background, and adolescents with migration backgrounds from Syria, Afghanistan and Iraq. The authors found that the adolescents with migration background held more traditional attitudes, which may cause more friction, especially for girls, when seeking greater autonomy as a young adult. Further friction may occur from a higher expectation to take on caretaker responsibilities (i.e., instrumental and emotional *parentification*) – if only because the children acculturate faster and outperform their parents socio-culturally (Titzmann, 2012). In combination with traditional views of women as caretakers or homemakers this may burden boys and girls differently. Earlier findings comparing father-daughter and father-son dyads in Turkish-German families suggest that second-generation daughters showed a significant shift towards more egalitarian values, while sons remained as conservative as their fathers (Idema & Phalet, 2007 – which may produce very different experiences with regard to the bond they have with their parents, and the organization of

family live on a day-to-day basis. While none of these findings could explain our data in particular and no two cultural backgrounds or even families are alike, they may serve to indicate some tendencies that need to be considered when discussing the life satisfaction particularly of young women in families with migration background.

4.2 Survey Mode: Why Would Girls Report Lower Satisfaction on Tablets?

It is less easy to explain why girls reported lower life satisfaction on tablets than on paper. While conducting our analyses, we did consider explanations such as differences in presentation on the tablets, or the possibility that on the tablets, answering all questions may have been mandatory (which may have led to differences in selectivity between tablets and paper). However, we did not find any satisfactory explanation; presentation of the survey was very similar and students could skip questions on either medium. We are thus left with our best guess that survey mode did indeed causally affect the gender gap in reported life satisfaction, without knowing anything specific about the underlying mechanisms. One may speculate that there are gender differences in familiarity with tablets, or maybe—in line with the idea that social media hurts girls in particular—girls have less pleasant associations with the mode than boys, who may be more likely to use tablets for gaming rather than browsing. To our knowledge, no such gendered mode effects have been reported before; we thus cannot confidently rule out that this pattern is just a chance fluctuation.

We also do not know whether life satisfaction assessments are more accurate on paper or on tablet, which would determine—regardless of statements about trends in gender gaps over time—whether the smaller paper gender gap or the larger tablet gender gap more appropriately captures gender differences. Studies tend to emphasize that assessments on paper and electronic assessments are generally comparable (e.g., Gwaltney et al., 2008; Howell et al., 2010; Květon et al., 2007; Muehlhausen et al., 2015) as long as differences in presentation are minor (Muehlhausen et al., 2015). However, for online assessments at least, respondents may feel like their anonymity is better protected (Ward et al., 2014). In our particular setting, this may also extend to tablets—while responses on tablets were simply entered and submitted by the students, paper questionnaires had to be collected and in a classroom context students may have reasonably expected that their handwriting could be identified (e.g., by teachers). Assuming this is the case, girls might have felt freer to report their true (lower) life satisfaction on tablet than on paper.

4.3 Other Outcome Measures: Widening Gender Gaps in Leisure Satisfaction and Friendship Satisfaction

Both the migration background and the survey mode pieces of the puzzle may suggest that any widening gender gap is just an artifact arising from changes in population composition and from changes in survey design. However, these factors cannot explain away widening gender gaps in students' satisfaction with their relationships to friends, and students' satisfaction with their leisure time activities. These observations are still compatible with accounts in which social media impairs well-being specifically in girls more so than in boys (Twenge & Martin, 2020), although in our data this pattern would be limited to two domains that are plausibly quite directly affected by social interactions online in students' leisure time. Of course, alternative or additional explanations exist. The opening of the gender gap

may also reflect the differential impact of the COVID-19 pandemic on the social lives and leisure time activities of girls in comparison to boys.

5 General Discussion

Our investigation started as the search for an explanation of a highly local observation: a widening gender gap in life satisfaction from 2010 to 2023 among adolescents in Leipzig. What, if any, generalizable insights arise from this? After all, we wouldn't necessarily expect widening gender gaps in life satisfaction in other places, and we certainly wouldn't assume that the specifics of the context of our data—such as strong compositional changes in the population of people with a migration background, and partial changes in survey mode—would occur elsewhere.

However, we believe that there are generalizable insights with respects to how such data and how such a reverse causal question over time should be approached. First, the composition of the population of interest played a major part in our explanation. As far as we can tell, such changes are rarely considered in psychological well-being research. At best, researchers may “control for” third variables such as migration background. However, notice how in our case, this would not have been sufficient to understand what is happening. We needed to additionally take into account how the associations vary between groups (i.e., we need to explicitly model an interaction with migration background), and then we additionally needed to consider the broader historical context to make sense of the observed patterns.

The change in survey mode from 2015 to 2023 is an unfortunate variation from the perspective of us as researchers who are interested in change over time. This issue can of course be prevented in the design phase of a study by ensuring no changes in survey mode; however, researchers may often rely on pre-existing longitudinal or repeated cross-sectional survey studies when investigating well-being over time and thus have little control over these matters. In fact, obstacles in implementation and the resulting variations in the sample are part of the day-to-day reality of large-scale survey studies, and may even be considered typical, especially in the context of educational research.

Our analyses are our best attempt at dealing with such variations in a transparent manner; we hope that they can provide one example with how to deal with unintended and unplanned stratifications, and how to work through their overall influence on a research question. For example, our final analyses rest on the assumption that survey mode causally affects the gender gap in life satisfaction; if, as we spelled out, the association is instead mainly to be attributed to selection effects, then the results underlying Fig. 4—with a larger life satisfaction gap among students without migration background in 2023—would provide a more accurate picture.

Of course, other “measurement artifacts” may also matter; for example, when investigating the incidence of mental health issues over time, changes in screening procedures and diagnoses (Guthrie et al., 2013; Wallace-Wells, 2024) may bias results beyond changes in the actual underlying phenomena. A further limitation of our investigation is that we cannot rule out that changes in the selectivity of the sample relative to the population may bias our results. We are fairly confident that the youth surveys provide good coverage of the underlying population—in each survey, schools from all planning districts were included, and our analyses include between 42 and 53% of all general secondary and grammar schools

in Leipzig—but this does not fully rule out the potential for bias. Furthermore, we have no additional information about non-respondents. In 2010, 70% of questionnaires were returned; in 2015, 75% of questionnaires were returned; and in 2023, 61% of questionnaires were returned. The missing respondents were likely a mix of students for whom parents declined participation, who declined participation themselves, or who happened to not attend school when data were collected. The absence of these students could bias our estimates of average life satisfaction, and in case the mechanisms underlying non-response changed over time, they may also bias the trends over time that are of central interest in our analysis. To contextualize the somewhat lower response rate in 2023, there is ample evidence in the survey literature that response rates have declined over time, across countries and different types and modes of data collection (e.g., Brick & Williams, 2013; Jabkowski & Cichocki, 2025; Luiten et al., 2020). Somewhat reassuringly, rising non-response rates do not *necessarily* imply rising non-response bias (Jabkowski et al., 2025), though of course they could still partially contribute to our findings.

Our analyses highlight the value of considering multiple well-being measures in parallel, which resulted in a more nuanced picture potentially supporting some explanations. Of course, multiple outcomes need to be disclosed transparently—including multiple dependent variables but only reporting those that return the “right” results increases the risk of false-positive findings (Simmons et al., 2011). A good template here may be personality psychology, a field in which it is very common to report results for all Big Five personality traits (extraversion, emotional stability, agreeableness, conscientiousness, openness to experience) by default, which reduces the risk of null results ending up in the file drawer (Atherton et al., 2021).

Lastly, we hope we have convinced readers that taking both methodological and demographic explanations seriously is worthwhile when addressing reverse causal questions. The introduction sections of psychological research articles often invoke such questions (why Y?); however, this often seamlessly transitions into a narrow forward causal question (how does X affect Y?) just before the Method section starts, which results in the standard format of many empirical journal articles. In contrast, entertaining multiple explanations, including some that may be deemed “uninteresting” as they point towards artifacts of the specifics of the study at hand, and which may hold true at the same time, results in a more complex article structure and less straightforward answers. This is only compounded by the fact that causal inference usually involves many uncertainties. However, we believe that these complexities and uncertainties, communicated transparently, are worth sharing if we want to tackle reverse causal questions more seriously and improve our understanding of the world.

Acknowledgements We would like to thank Taym Alsalti, Saloni Dattani, Jens Lange, and Jan-Paul Ries who provided helpful input on this project in a colloquium; Stefan C. Schmukle provided extensive feedback on a draft of this manuscript. At the city of Leipzig, we would like to thank Karen Hemming, Andrea Schultz and Martin Waschpik for their advice and feedback.

Author contributions All authors contributed to the study conception and design of the analyses. Material preparation was performed by Gregor Kachel. Analysis was performed by Julia M. Rohrer with additional input from Richard McElreath. The first draft of the manuscript was written by Julia M. Rohrer and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received to assist with the preparation of this manuscript.

Data Availability Due to legal and data protection requirements, the underlying data are only available upon request from the data holder (City of Leipzig). All analysis code can be found on the companion website (<https://j-rohrer.github.io/Jugendstudienanalysen/>).

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Informed consent Informed consent of the original data collection was handled by the City of Leipzig; parents were informed upfront and consented to their children's participation.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abel, F., Bittner, J., Ehlert, T., Greunke, P., Hemming, K., Kachel, G., Köbler, T., König, R., Lagrange, M., Netwall, N., Schultz, A., & Waschipy, M. (2023). *Jugend in Leipzig 2023 [Youth in Leipzig 2023]*. Stadt Leipzig. https://static.leipzig.de/fileadmin/mediendatenbank/leipzig-de/Stadt/02.1_Dez1_Allgemeine_Verwaltung/12_Statistik_und_Wahlen/Stadtforschung/Jugend-in-Leipzig-2023.pdf
- Abel, F., Heinemann, J., Lehmann, K., Schultz, A., Lamperti, K., & Bischof, M. (2015). *Jugend in Leipzig 2015 [Youth in Leipzig 2015]*. Stadt Leipzig. https://static.leipzig.de/fileadmin/mediendatenbank/leipzig-de/Stadt/02.1_Dez1_Allgemeine_Verwaltung/12_Statistik_und_Wahlen/Stadtforschung/Jugend_in_Leipzig_2015.pdf
- Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to Interpret Statistical Models Using marginaeffects for R and Python. *Journal of Statistical Software*, 111(9), 1–32. <https://doi.org/10.18637/jss.v111.i09>
- Atherton, O. E., Chung, J. M., Harris, K., Rohrer, J. M., Condon, D. M., Cheung, F., Vazire, S., Lucas, R. E., Donnellan, M. B., Mroczek, D. K., Soto, C. J., Antonoplis, S., Damian, R. I., Funder, D. C., Srivastava, S., Fraley, R. C., Jach, H., Roberts, B. W., Smillie, L. D., & Corker, K. S. (2021). Why has personality psychology played an outsized role in the credibility revolution? *Personality Science*. <https://doi.org/10.5964/ps.6001>
- Bein, C. (2023a). *Migrantinnen und Migranten in Leipzig 2022 [Migrants in Leipzig 2022]*. Stadt Leipzig, Amt für Statistik und Wahlen. https://static.leipzig.de/fileadmin/mediendatenbank/leipzig-de/Stadt/02.1_Dez1_Allgemeine_Verwaltung/12_Statistik_und_Wahlen/Analysen_zur_Stadtgesellschaft/001-Migrantinnen-und-Migranten-2022.pdf
- Bein, C. (2023b). *Migrantinnen und Migranten in Leipzig 2022*. Stadt Leipzig, Amt für Statistik und Wahlen. https://static.leipzig.de/fileadmin/mediendatenbank/leipzig-de/Stadt/02.1_Dez1_Allgemeine_Verwaltung/12_Statistik_und_Wahlen/Analysen_zur_Stadtgesellschaft/001-Migrantinnen-und-Migranten-2022.pdf
- Brick, J. M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American Academy of Political and Social Science*, 645(1), 36–59.
- Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Deutsche Welle (2019, September 10). *Most migrant families speak German at home*. Deutsche Welle. <https://www.dw.com/en/german-is-the-most-spoken-language-in-immigrant-households/a-50374819>
- Dütthorn, P., & Gugutskow, S. (2010). *Migranten in der Stadt Leipzig 2010 [Migrants in the City of Leipzig 2010]*. Stadt Leipzig, Amt für Statistik und Wahlen in Kooperation mit dem Referat für Migration und Integration.
- Dütthorn, P., & Gugutskow, S. (2010). *Migranten in der Stadt Leipzig 2010*. Stadt Leipzig, Amt für Statistik und Wahlen in Kooperation mit dem Referat für Migration und Integration.

- Dütthorn, P., & Gugutskow, S. (2016a). *Migrantinnen und Migranten in Leipzig 2015 [Migrants in Leipzig 2015]*. Stadt Leipzig, Amt für Statistik und Wahlen in Kooperation und Referat für Migration und Integration.
- Dütthorn, P., & Gugutskow, S. (2016b). *Migrantinnen und Migranten in Leipzig 2015*. Stadt Leipzig, Amt für Statistik und Wahlen in Kooperation und Referat für Migration und Integration.
- Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40, 31–53.
- Fabian, M., Kaiser, C., Panasiuk, S., Funk, S., Pountney, L., & Brett, C. (2024). Evidence Against the Simple Validity of Life Satisfaction Scales from Long Cognitive Interviews. <https://iariw.org/wp-content/uploads/2024/08/FINAL-Fabian-et-al.-Cognitive-Interviewing-IARIW.pdf>
- Gelman, A., & Imbens, G. (2013). Why ask Why? Forward Causal Inference and Reverse Causal Questions (No. w19614). National Bureau of Economic Research. <https://doi.org/10.3386/w19614>
- Guthrie, W., Swineford, L. B., Nottke, C., & Wetherby, A. M. (2013). Early diagnosis of autism spectrum disorder: Stability and change in clinical diagnosis and symptom presentation: Stability of early diagnosis and symptoms in ASD. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 54(5), 582–590.
- Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 11(2), 322–333.
- Haidt, J. (2024). The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness. <https://strategylab.ca/wp-content/uploads/2024/07/The-Anxious-Generation-Supplemental-Resources.pdf>
- Heinemann, J., Lehmann, K., Rosk, G., & Schultz, A. (2011). *Jugend in Leipzig – Ergebnisse einer Befragung 2010 [Youth in Leipzig – Results from a 2010 Survey]*. Stadt Leipzig.
- Howell, R. T., Rodzon, K. S., Kurai, M., & Sanchez, A. H. (2010). A validation of well-being and happiness surveys for administration via the Internet. *Behavior Research Methods*, 42(3), 775–784.
- Idema, H., & Phalet, K. (2007). Transmission of gender-role values in Turkish-German migrant families: The role of gender, intergenerational and intercultural relations. *Zeitschrift Für Familienforschung*, 19(1), 71–105. <https://www.ssoar.info/ssoar/handle/document/5806>
- Jabkowski, P., & Cichocki, P. (2025). Survey response rates in European comparative surveys: A 20-year decline irrespective of sampling frames or survey modes. *Quality & Quantity*, 59(S1), 635–655.
- Jabkowski, P., Kohler, U., & Kołczyńska, M. (2025). Against all odds: On the robustness of probability samples against decreases in response rates. *Survey Research Methods*, 19(1), 83–104.
- Jugendstudie (2023). : Lebenszufriedenheit der Schüler sinkt deutlich [Youth Survey 2023: Life satisfaction of students declines markedly]. (2023, December 7). <https://www.leipzig.de/newsarchiv/news/jugendstudie-2023-lebenszufriedenheit-der-schueler-sinkt-deutlich>
- Květon, P., Jelínek, M., Voboril, D., & Klimusová, H. (2007). Data collection on the internet: Evaluation of web-based questionnaires. *Studia Psychologica*, 49(1), 81–88.
- Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, 36(3), 469–487.
- Muehlhausen, W., Doll, H., Quadri, N., Fordham, B., O'Donohoe, P., Dogar, N., & Wild, D. J. (2015). Equivalence of electronic and paper administration of patient-reported outcome measures: A systematic review and meta-analysis of studies conducted between 2007 and 2013. *Health and Quality of Life Outcomes*, 13(1), 167.
- Nakazawa, D. J. (2022). Girls on the Brink: Helping our daughters thrive in an era of increased anxiety, depression, and social media. https://books.google.de/books?hl=en&lr=&id=xKZUEAAQBAJ&oi=fnd&pg=PT10&dq=Girls+on+the+Brink+Donna+Jackson+Nakazawa&ots=_2ARmf8TIT&sig=vTgqrNeMKkPTEKJ8sUiVi211z5
- Nilles, H., El-Awad, U., Kerkhoff, D., Braig, J., Schmees, P., Kilinc, Y., Rueth, J. E., Eschenbeck, H., & Lohaus, A. (2023). Gender role attitudes and well-being of German and refugee adolescents-same or different? *Bmc Psychiatry*, 23(1), 663.
- Posit Team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC. <http://www.posit.co/>
- R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42.
- Rohrer, J. M., & Arel-Bundock, V. (2026). Models as prediction machines: How to convert confusing coefficients into clear quantities. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/25152459261424825>

- Rohrer, J. M., & Arslan, R. C. (2021). Precise Answers to Vague Questions: Issues With Interactions. *Advances in Methods and Practices in Psychological Science*, 4(2), 25152459211007370.
- Sewall, C. J. R., & Parry, D. A. (2024). Social media and youth mental health: Simple narratives produce biased interpretations. *Journal of Psychopathology and Clinical Science*, 133(7), 507–514.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Simonsohn, U. (2019). [80] Interaction Effects Need Interaction Controls. <http://datacolada.org/80>
- Titzmann, P. F. (2012). Growing up too soon? Parentification among immigrant and native adolescents in Germany. *Journal of Youth and Adolescence*, 41(7), 880–893.
- Tomova, G. D., Silverwood, R. J., Tennant, P. W. G., & Wright, L. (2026). How can the use of different modes of survey data collection introduce bias? An introduction to mode effects using directed acyclic graphs (DAGs). *American Journal of Epidemiology*. <https://doi.org/10.1093/aje/kwag017>
- Twenge, J. M., & Martin, G. N. (2020). Gender differences in associations between digital media use and psychological well-being: Evidence from three large datasets. *Journal of Adolescence*, 79(1), 91–102.
- Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. <https://www.stats.ox.ac.uk/pub/MASS4/>, *Vertragsarbeiter in der DDR*. (n.d.). DOMiD | Dokumentationszentrum und Museum über die Migration in Deutschland. Retrieved May 13, 2025, from <https://domid.org/news/vertragsarbeit-in-der-ddr/>
- Wallace-Wells, D. (2024). May 1. Opinion. The New York Times. <https://www.nytimes.com/2024/05/01/opinion/smartphones-social-media-mental-health-teens.html>
- Ward, P., Clark, T., Zabriskie, R., & Morris, T. (2014). Paper/pencil versus online data collection: An exploratory study. *Journal of Leisure Research*, 46(1), 84–105.
- Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, 40(3), 424–431.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.