European Journal of Psychology of Education (2025) 40:97
https://doi.org/10.1007/s10212-025-00994-y

Ina picks flowers, while Klaus rides a motorcycle: a quantitative text analysis of gender stereotypes in German standardized spelling tests

Nancy Tandler[1] · Felix Peter[2] · Johanna Rimpf[1] · Teresa Wessels[1] · René T. Proyer[1]

Received: 31 December 2023 / Revised: 9 July 2025 / Accepted: 10 August 2025 /
Published online: 3 September 2025

Abstract

While designed to assess spelling, spelling tests often embed specific content within their items. This content could also contain depictions of gender roles or cultural norms. The aim of this study was to investigate whether female and male characters in German-language spelling tests are portrayed in a gender-stereotypical manner. We expected that female characters would be more likely to be portrayed in stereotypically feminine roles, while male characters would be more likely to be depicted in stereotypically masculine roles and conducted two studies with different methodological approaches to test our hypotheses. Study 1 comprised two consecutive quantitative text analyses: First, seven raters assessed the textual content of the five German-language spelling tests recommended by the German Society for Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy (Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V.; 2015). They categorized all subjects presented in the material as male, female, gender neutral, generically masculine, or non-human. Second, the subjects characterized as female or male were further analyzed to determine the roles and activities they were depicted in. Male subjects were more frequently presented in the test material than female subjects, and female characters were more likely to be depicted in parental or caregiving roles or in other caregiving professions. Male subjects were more likely to be depicted in professional and leadership positions, engaging in heavy (physical) housework or pursuing careers in STEM (science, technology, engineering, and mathematics) fields. Study 2 was a quantitative analysis of a subset of the spelling test material from Study 1, with the gender of the subjects masked. Then, 143 participants rated the subjects in the text on a scale of typical masculinity or femininity. The results of Study 2 suggest that the spelling test material contained gender stereotypes. We conclude that standardized spelling tests may contain and elicit gender stereotypes among students and recommend the implementation of our findings in educational materials.

Keywords Gender stereotypes · Spelling tests · Gender-fair language · Gender roles · Gender

Extended author information available on the last page of the article

Introduction

In educational assessment, standardized spelling tests are widely used tools for evaluating students' writing abilities (Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V., 2015). However, the potential presence of gender stereotypes within these tests raises concerns. Gendered assumptions, even in seemingly neutral educational materials, could affect how young students perceive roles and abilities across genders, possibly influencing their development, perception of gender roles, attitudes, and/or their performance (Buckley et al., 2022; Guichot-Reina & De la Torre-Sierra, 2023). To our knowledge, no study has systematically tested whether gender stereotypes do actually occur in standardized spelling tests. In order to narrow this gap, we conducted two studies of standardized German-language spelling tests for 6- to 16-year-olds.

Gender disparities are well documented across various sectors. Women predominantly work in fields like education (e.g., Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend, 2023; United States Census Bureau, 2023), while they are underrepresented in science, technology, engineering, and mathematics (Destatis Statistisches Bundesamt, 2024; Lewis et al., 2021). For example, 77.2% of students enrolled in educational programs in Germany in 2022 were female, whereas only 22.9% of those enrolled in engineering programs were female (Bundesministerium für Forschung, Technologie und Raumfahrt, 2025). Such differences are not just a matter of interest; there is a debate to which degree they are rooted in societal expectations and reinforced by stereotypes that shape young people's perceptions of their aptitudes and career opportunities. Educational environments, comprising the actual diversity of teachers and caretakers, as well as classroom materials, like textbooks, and standardized tests, play a critical role in conveying and potentially legitimizing such gender norms (e.g., Buckley et al., 2022; Stuve & Rieske, 2018).

Gender stereotypes

Stereotypes concern the "rigidity" (in Greek, stereos means rigid, fixed) of impression formation. Stereotypes are defined as cognitive structures people use when processing social information, and they contain knowledge and beliefs about members of a social group (Kite & Whitley, 2016). Stereotypes can lead people to judge individuals based on the characteristics of the social group to which they are assigned, even if those characteristics do not correspond to the individual's actual qualities and/or are neither accurate nor fair (Katz & Braly, 1933). The usage of stereotypes can lead to simplified information processing. Gender is one of the most prominent characteristics around which stereotypes are formed, alongside age and ethnicity (Fiske, 1998; Six-Materna, 2020), and gender stereotypes are defined as socially shared knowledge about characteristics assigned to women and men (Eckes, 2008). They specifically assign qualities such as competence (e.g., active, dominant, rational, aggressive) more often to men and warmth (e.g., passive, submissive, empathetic, warm-hearted) to women, creating a dual set of expectations for behavior and professional roles (Abele & Bruckmüller, 2016; Cuddy et al., 2008; William & Best, 1990). These stereotypes often include assumptions about suitable occupations, with women as caregivers or educators and men as scientists or engineers (Rice, 2014).

Girls, for instance, are frequently less interested and have less-positive attitudes toward science and technology compared to boys (e.g., Brotman et al., 2008; Kerger et al., 2011), while male students and young men are frequently less attracted to working in education (Middendorff et al., 2017; Stuve & Rieske, 2018). Thus, the impact of stereotypes likely extends beyond simple categorization. Research has repeatedly shown that gender stereotypes have a direct impact on a variety of outcomes on the stereotyped person, including changed interests, information processing, performance, behavior, or self-image (e.g., Hermann & Vollmeyer, 2016; Steinmayr et al., 2019).

Gender stereotypes can influence individuals' attitudes and self-perceptions, as demonstrated in studies on stereotype threat. A stereotype threat theory assumes that people experience a sense of threat when they find themselves in a situation where they (a) fear being judged on the basis of negative stereotypes and (b) fear unintentionally confirming negative stereotypes about their group through their own behavior (Steele & Aronson, 1995). In educational contexts, this effect has been observed among, for example, German students, where girls performed worse in math when gender stereotypes were activated before testing (Hermann & Vollmeyer, 2016). No such effect was found for boys. Other studies have shown that exposing students to counter-stereotypical role models can help alleviate such performance disparities. For example, female high school students in the United States achieved greater comprehension in a science lesson when a female scientist was represented compared to a male scientist (Good et al., 2010). A study with primary school kids in Australia revealed that girls' self-esteem increased after having listened to a story with a same-sex main character compared to girls exposed to an opposite-sex main character; the same pattern was obtained for boys (Ochman, 1996). Hence, there appears to be some generalizability of the effects across different countries and cultures.

Gender stereotypes in learning materials

Recent research has begun to scrutinize gender stereotypes in learning materials more closely. To our knowledge, there were no studies on the occurrence of gender stereotypes in standardized spelling test materials. However, there are findings from related fields. In German-speaking countries, studies have shown that recent schoolbooks provide a more balanced gender representation (Moser & Hannover, 2014). While improvements have been made, subtle biases persist: Women appear less frequently than men in images, especially in math books, and are still depicted in traditional roles more often than men (Moser & Hannover, 2014). In older textbooks, gender biases were more pronounced. Females were underrepresented in images and texts, typically portrayed in traditional roles including domestic work, caring, and low-income jobs, whereas male persons were shown participating in a wide range of activities, including high-income jobs (e.g., Glötzner, 1982; Hellinger, 1980; Ohlms, 1984). Recent studies have also found gender biases in online educational resources, with more men depicted as scientists and more women as teachers (Kerhoven et al., 2016). Such patterns reflect broader societal expectations and reinforce traditional gender roles, subtly guiding young learners' perceptions of professional possibilities.

Despite these developments, research on the presence of gender stereotypes in standardized tests as a significant part of educational materials remains scarce, which is questionable because standardized tests carry a level of authority in educational contexts, often serving as benchmarks in psychological diagnostics and academic assessments. They are often characterized by special test situations and sometimes carried out by people who are strangers to the subjects, like psychologists, therapists, or specialized teachers. These materials influence decisions on educational placement, intervention, and progress monitoring, affecting students from Grade 1 to high school. Some of these tests are administered not only to students with specific learning difficulties but also to entire cohorts for periodic assessments (e.g., Hamburg Spelling Test; May, 2010). Thus, a substantial number of students encounter these materials, some of them also repeatedly (e.g., the prevalence of dyslexia is 3 to 8%; Moll et al., 2014), making it important to examine whether these tests contain latent biases that might impact students' academic self-image and future aspirations.

Following previous research on the consequences of gender stereotypes in the educational context, gender biases in standardized test materials could have profound effects. For instance, girls might see fewer female role models in examples used within tests, which could reinforce stereotypical notions about their career potentials and undermine their confidence in nontraditional fields like science and technology. Conversely, boys might feel pressured to align with traditional male roles and expectations, limiting their openness to roles in education or caregiving. If stereotypical portrayals are presented as part of exceptional situations at school (e.g., school psychological or other educational diagnostics) or officially sanctioned materials (e.g., learning status or progress analyses), students might perceive these representations as objective truths, reinforcing and legitimizing traditional gender roles.

The present research thus aimed to address a notable gap in the research by examining whether standardized spelling tests in German-speaking countries reflect gender stereotypes, specifically per the portrayal of male and female roles within test items. Thus, we aimed to examine whether gender differences—such as those in study program prevalence, gender roles (e.g., parental roles), activities, and work settings—are reflected in the test materials used to assess writing abilities in 6- to 16-year-olds in German-speaking countries (Study 1), and whether these representations reflect common gender stereotypes (Study 2). We opted for spelling tests because they are widely used in educational settings in Germany, contain a particularly large amount of textual material in contrast to other performance tests (e.g., in mathematics), and the involvement with this material is particularly intensive via various senses (listening to spoken text, processing it, converting it into handwriting with motor skills, visually checking what has been written, etc.). In particular, we studied gender stereotypes in tests recommended by the so-called "S3-Leitlinien," the official guideline recommended for the diagnosis of children and adolescents with spelling disorders, which is published by the German Society for Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy (Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V., 2015) to assess spelling difficulties. Diagnostic tests recommended in this guideline for the diagnosis of spelling difficulties are considered to be appropriately valid, reliable, objectively feasible, and standardized.

To the best of our knowledge, this paper describes the first examination of such test materials in German-speaking countries. The results could inform future test designers and emphasize the importance of eliminating biases, including in assessment materials as a subset of education-related materials, to promote a more gender-equitable educational landscape.

The present studies

This paper presents two studies that investigated gender portrayals in standardized spelling tests used to assess spelling deficits in German-speaking countries. To determine the type of representation of men and women in the tests, we conducted two studies. In Study 1, we selected information units that depicted the subjects and analyzed the frequencies of female and male subjects presented in the test material. We also analyzed whether the female and male subjects were presented differently in terms of roles and activities. In Study 2, we examined whether the male and female subjects were presented in a gender-stereotypical way.

Study 1: frequencies of female and male subjects, roles, and activities

In Study 1 we aimed to investigate the frequencies of female and male subjects depicted in spelling test materials. Based on previous research on gender presentation in schoolbooks and learning materials (e.g., Kerkhoven et al., 2016; Moser & Hannover, 2014), we expected that female subjects would be presented less frequently than male subjects. We were also interested in whether female and male subjects were presented differently regarding their assigned roles and activities. Therefore, we analyzed the frequency of presentation of female and male subjects in various categories related to activities such as caring, doing household work, or helping in parental and educational roles. A complete list of all categories is shown in the coding guide in Table 1.

Method

Sample

We selected five standardized spelling tests suited for children in German school grade 1 up to grade 10 (i.e., students 6 to 16 years old) to assess students' spelling abilities when diagnosing spelling deficits. The tests were selected based on a sequential decision process. First, we included the five spelling tests recommended by the current S3 guideline (in the document named *Diagnosis and treatment of children and adolescents with reading and/or spelling disorders*) for the diagnosis of children and adolescents with spelling disorders, which was written by the German Society for Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy (Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V., 2015). The document lists only those spelling tests that are particularly suitable in terms of content and psychometrics. Second, we discussed with practitioners, including educational psychologists from the "Local State School Administration" and teachers, whether those tests were in line with educational standards, frequently used (in terms of diagnosis and instruction), and practical to implement. Based on our sequential decision process, we included the five tests recommended by the current S3 guideline.

We finally included the following five spelling tests: the *Hamburger Schreibprobe* (Hamburg Spelling Probe) (HSP; May, 2010), the *Weingartener Grundwortschatz Rechtschreibtest* (Weingarten Basic Vocabulary Spelling Test) (WRT; Birkel, 2007a, 2007b, 2007c), the *Salzburger Lese- und Rechtschreibtest* (Salzburg Reading and Spelling Test) (SLRT-II; Moll & Landerl, 2010), the *Rechtschreibtest—Neue Rechtschreibregelung* (Spelling Test—New Orthographic Regulation) (RST-ARR; Ibrahimovic & Bulheller, 2017), and the *Deutscher Rechtschreibtest* (German Spelling Test) (DERET; Martinez Méndez et al., 2015; Stock & Schneider, 2008a, 2008b). We used multiple versions of each test because the tests are grade-specific, and some tests have two forms (A and B). A detailed description of the number of tests and the number of items, sentences, and words used from each test is presented in Table 2.

Table 1 Coding guide

Categories	Subcategories	Text examples	Coding rule
Age group	Child	“KLAUS and Robert fight over the ball.”	First names, if the context does not clearly speak for adults, such as grandfather Tino The boys, the girls, the little sister, daughter, son
	Adolescent	“The youth group explored the lake landscape with rented paddle boats. “It was impressive, but also exhausting,” reports an EXCURSION PARTICIPANT.”	First names, if the context does not clearly speak for adults, such as grandfather Tino If it is recognizable from the context and activities, e.g., they have just finished school, they are applying for an apprenticeship
	Adult	“MOTHER goes there regularly.”	Job titles, roles in the family: father, aunt, grandfather, family names
Attributes	Female	“Erika begins to cry.” “An understanding ANIMAL DOCTOR.”	Passive, submissive, empathic, warm-hearted
	Male	“The old man is certainly very wise.”	Active, dominant, rational, aggressive
Sports activities		“REFEREE,” “COACH”	Does the person play sport in any way, e.g., referee, coach, striker, runner…
Physical appearance		“Wreath in hair,” “beautiful skirt”	Description of the appearance including clothing, decoration, body features
Positive attributes		“Reliable,” “paints beautiful picture”	Positive attribute (such as being described as shrewd, reliable, carrying heavy bags, or getting along again) and when activities or processes were managed or worked out
Negative attributes		“Liar,” “has stolen,” “clumsy”	Negative attribute (aggressive, quarrels all the time, clumsy, liar) and when activities or processes were not managed or did not work out
Activities	Parental role	“TOM lifted the baby onto his lap.”	
	Low-status profession	“He waves to a CAB DRIVER, but he doesn’t move.”	
	High-status profession	“Kurt wants to become a paralegal and maybe even a JUDGE or lawyer later on.”	
	Caring professions	“SHE did her first internship as a nursery nurse in a daycare center.” “The teacher…”	

Table 1 (continued)

Categories	Subcategories	Text examples	Coding rule
	STEM profession	“HE enjoys working on the computer.”	
	School homework	“First, of course, we have to do the homework,” MONIKA said.”	
	Cultural	“I went to the theater last night!” Mr. Müller tells his WORK COLLEAGUE.”	
	Building, carrying, constructing	“Then I’ll just paint the window frames in the basement,” he says.”	
	Garden work	“Before HE cut the rose hedges.”	
	Fine arts, painting, crafts	“PETER paints a beautiful picture.”	
	Playing music	“TILO must lock up his drums after the rehearsal.”	
	Active playing	“I’m sure I can climb our thick trees with HER.”	
	Sports activities	“The GOALKEEPER scolds the referee.”	
	Putting on, dressing up	“The MAN puts on his hat.”	
	Walking, hiking	“ONKEL KARL had signed up for a mountain hike.”	
	Communicating	“He asks when HE must be home.”	
	Reading, writing	“ROBERT writes on a new sheet.”	
	Counting	“My little SISTER can only count to three.”	
	Food, drinking, eating	“Would you like a pear?” MOTHER asks me.”	
	Shopping	“On the way back, they bought a dish cleaner for Kurt’s MOTHER.”	
	Listen to someone	“Well, tell me about it!” says mother and sits down at GERTRUD’s bedside.”	
	Sleeping	“In the evening SHE came home tired.”	
	Cleaning, cooking, laundry	“My SISTER INGE baked three cakes, all by herself.”	
	Driving a car	“The impatient car driver moans in exasperation in big city traffic. “if only I had taken the train, I would have reached my destination long ago.”	

Table 1 (continued)

Categories	Subcategories	Text examples	Coding rule
Professional role		"Cab driver," "carpenter," "secretary"	When a profession is mentioned
Parental role		"Father," "mother"	
Leadership role	Professional	"Branch manager," "boss"	
	Leisure	"Club boss"	
Domestic work	Lighter work	"Hangs laundry," "knits," "bakes cookies," "goes shopping"	Less physically demanding
	Harder work	"Digs in the garden," "paints the house"	More physically demanding
Caretaker role	Nurturing	"Feeds baby," "takes care of the children," "watches the children"	Caring and watching children
	Activities	"Animates child to play soccer," "leads bike tour with children"	Doing activities such as outside activities with children

Notes: Words in uppercase letters represent the character of interest

Table 2 Description of the number of versions for each test with the related numbers of tasks, words, and sentences

Tests	Test versions	Number of tasks	Number of sentences	Number of words
HSP	1 Class +	10	11	39
	2 Class +	18	18	53
	3 Class +	18	19	53
	4–5 Class	21	21	58
	5–6 Class	20	21	73
	7–8 Class	22	40	299
	9–10 Class	22	50	361
WRT	2 Class + , Form A	1	49	225
	2 Class + , Form B	1	43	250
	3 Class + , Form A	4	57	395
	3 Class + , Form B	4	67	426
	4 Class + , Form A	7	100	654
	4 Class + , Form B	6	92	708
SLRT-II	2 Class, Form A	24	24	129
	2 Class, Form B	24	24	129
	3–4 Class, Form A	48	48	268
	3–4 Class, Form B	48	48	283
RST-ARR	Age 14 to 61 years	1	95	996
DERET	1–2 Class, Form A	9	15	98
	1–2 Class, Form B	9	14	93
	3–4 Class, Form A	7	17	142
	3–4 Class, Form B	7	18	137
	5–6 Class, Form A	24	33	273
	5–6 Class, Form B	27	40	275

Notes: HSP Hamburger Schreibprobe/Hamburg spelling probe, *WRT* Weingartener Grundwortschatz Rechtschreibtest/Weingarten basic vocabulary spelling test, *SLRT-II* Salzburger Lese–und Rechtschreibtest/Salzburg reading and spelling test, *RST-ARR* Rechtschreibtest–Neue Rechtschreibregelung spelling test–new orthographic regulation, and *DERET* Deutsche Rechtschreibtest/German spelling test

We ordered the material of the selected tests according to the lengths of the material into three task groups when creating the scoring sheet for the raters: (1) single words (which did not form a full sentence), (2) single sentences, and (3) continuous texts. We did not change the test material and included the full material from all tests. We presented the material with the correct answers to be filled in by the children, that is, the material that was usually read aloud by the test administrator and then written in full by the participants or written into text gaps. We excluded pictures or drawings from our analyses. We presented items that occurred more than once in the various test versions only once in this survey, but we considered multiple occurrences in further analyses.

Procedure and materials

Coding of subjects' gender In the first step, seven independent raters coded each subject in the complete test material as "human female," "human male," "gender neutral," "masculine generic," or "non-human." *Masculine generic* in the German language is a male term that can refer to mixed-gender groups, to persons with unknown gender, or when the gender of a person is irrelevant. Our raters (four females, three males) were between 16 and 55 years old ($M = 28.14$, SD = 13.85). They were native German speakers with educational levels ranging from the equivalent of a high school diploma to university degrees.

Coding of activities and roles To test for gender-related differences in the depicted roles and activities, we selected only the specific parts of the test material where the subjects in the first step were coded as female or male characters. Then, we asked four other raters to rate each of these female or male characters within their corresponding text unit (when existing) across various coding systems, considering roles and activities. This entire coding system, which functioned as a coding guide as well, is depicted in Table 1. The coding guide presents the coding system with its categories, and for each category, definitions, text examples, and coding rules are outlined.

This coding guide was developed and refined in accordance with the required standards for qualitative content analysis published by Mayring (2008) to ensure a standardized procedure across raters. Accordingly, in a pre-analysis, an interdisciplinary team that consisted of researchers and educational psychologists selected a random sample of text items from all five spelling tests and simultaneously referred to the existing literature on gender stereotypes in textbooks (e.g., Moser & Hannover, 2014; Rohrmann & Thoma, 1998). Accordingly, we prepared our categorization system and coding guidelines to explain the categories and codes for all raters.

These four raters (two male, two female) were between 22 and 30 years old ($M = 25.75$, $SD = 3.30$) and were undergraduate university students. To ensure a standardized procedure across all raters, all raters underwent an individual standardized training session taught by the same instructor. The training materials and the content of the training were aligned to the coding system. During the training, all raters had to rate the same test material and discuss their decisions afterward with the instructor. To further ensure a standardized approach, the raters were given the opportunity to contact the instructor in case of doubts. Finally, after having completed the coding, each rater had a final discussion with the instructor to clarify ambiguous or open-coding decisions. All four raters coded the entire test material independently. Each text item could be coded only once within a category system, but each could be coded in all categorization systems.

Results

Frequencies of female and male characters in the spelling test material

The interrater reliability for the gender of the subjects was high ($\alpha = .98$). We considered agreement on the selection of the relevant information unit (i.e., subject) and its gender when at least four raters agreed on both the selection and coding. This is the case in 98.6% of the entries. Afterward, four experts (two psychologists with PhDs and two graduate students of psychology) discussed the classifications of the remaining units ($n = 26$) that the raters had not agreed upon until an agreement was reached.

The text material was segmented into 1295 information units that contained a subject. Subjects were coded as *female* characters in 354 (27.3%) cases (e.g., "the mother is feeding the baby," "Anna is buying butter and milk"), as *male* characters in 466 cases (36.0%;

e.g., “the father switches off the light,” “I can’t, I have to go to work, dad is grumbling”), as *gender neutral* in 269 cases (20.8%; e.g., “we need bread,” “I love singing”), as *masculine generic* in 27 cases (2.1%; e.g., “the doctor examines the students,” “we elect a new class speaker”), and as *non-human* in 179 cases (13.8%; e.g., “the tree,” “the lion”). To answer our first research question—whether male characters are more often presented than female characters—we considered only the text units categorized as *female* and *male* ($n=820$). A single comparison revealed that male characters appeared statistically significantly more frequently than female characters, χ^2 (1) = 15.30, $p = <.001$, with a small effect size of Cramer’s $V=.14$. We categorized effect sizes as follows: small effect sizes around .1, medium effect sizes around .3, and large effect sizes around .5; see ESM_1 for a detailed analysis of the gender distributions across the various spelling tests.

Frequencies of roles and activities of female and male characters in the spelling test material

Only the text units whose subject was either coded as a female character or a male character in the previous rating were coded in this section. Here, we applied several categorizations such as occupational and parental roles and various activities (the detailed list of the various categories, and the referred coding system is presented in Table 1). Four raters coded the characters independently. We used a two-step approach to reach a final decision on the coding. First, a category had to be chosen by at least three raters. If this was not the case for an entry, an additional coder (a psychologist) reviewed the material and made a final decision.

At first, the male and female characters were coded per age group across the categories of child, adolescent, or adult. Statistically significantly more male ($n=258$, 55.4%) than female characters ($n=124$, 35.0%) were coded as adults, χ^2 (3) = 34.12, $p<.001$ (see Table 3). A less fine-grained analysis (combining children and adolescents into one category) further supported this finding, with more female ($n=135$, 38.1%) than male entries ($n=127$, 27.3%) rated as children or adolescents, χ^2 (2) = 33.70, $p<.001$. Effect sizes were of small to medium size (Table 3), and interrater agreement was of moderate to substantial size (Table 3). Interrater agreement was calculated with Fleiss kappa, which is recommended when using more than two raters and nominal variables (Landis & Koch, 1977).

We also found that female characters were rated as having statistically significantly more stereotypical female attributes than male characters, while male characters were assigned statistically significantly more stereotypical male attributes than female characters. Female characters were further coded as being less depicted in sports activities and more presented in relation to physical appearance than male characters. Female characters were rated as having more positive attributes than male characters (e.g., “she is also modest, as she chooses the smallest piece for herself”), while no such gender difference was found for negative attributes (e.g., “the text is so awful that the poor girl will probably never understand it”). Considering additional activities, female characters were statistically significantly more frequently presented in relation to drinking and eating and with sleeping activities than male characters, while no gender differences were found for activities like communication, shopping, or driving a car. Effect sizes were of small to medium size (Table 3). Interrater agreement was nearly fair for attributes and of moderate size for the remaining categorizations, whereas the highest agreement was obtained for negative attributes and sports activities (Table 3). A detailed list of the distribution across activities is displayed in the Appendix (ESM_2).

Table 3 Frequencies, percentages, and comparisons of age groups, roles, and activities assigned to female and male subjects in various spelling tests

	Female (*N* = 354)		**Male** (*N* = 466)		**Chi^2-test**				**Interrater reliability**
	Numbers		Numbers						
	Absolute	Relative	Absolute	Relative	χ^2	*df*	*p*	Cramer's *V*	
Age group					34.12	3	< .001	.20	.47
Child	69	19.5%	70	15.0%					
Adolescent	66	18.6%	57	12.2%					
Adult	124	35.0%	258	55.4%					
None mentioned	95	26.8%	81	17.4%					
Age group					33.70	2	< .001	.20	.47
Child and adolescent	135	38.1%	127	27.3%					
Adult	124	35.0%	258	55.4%					
None mentioned	95	26.8%	81	17.4%					
Attributes					92.78	2	< .001	.34	.19
Female	98	27.7%	24	5.2%					
Male	21	5.9%	80	17.2%					
None mentioned	235	66.4%	362	77.7%					
Sports activities					15.51	1	< .001	.14	.52
Mentioned	12	3.4%	50	10.7%					
None mentioned	342	96.6%	416	89.3%					
Physical appearance					4.00	1	.046	.07	.30
Mentioned	13	3.7%	7	1.5%					
None mentioned	341	96.3%	459	98.5%					
Positive attributes					10.28	1	< .001	.11	.33
Mentioned	97	27.4%	84	18.0%					
None mentioned	257	72.6%	382	82.0%					
Negative attributes					0.48	1	.490	.02	.50
Mentioned	112	31.6%	137	29.4%					
None mentioned	242	68.4%	329	70.6%					
Further activities					125.91	24	< .001	.39	.39
Parental role	30	8.5%	23	4.9%					
Low-status profession	7	2.0%	33	7.1%					
High-status profession	4	1.1%	48	10.3%					
Caring professions	21	5.9%	2	0.4%					
STEM profession	0	0%	5	1.1%					

Notes: Fully display of further activities can be found in ESM A; *N* = 820, text units that contain a male or female subject. Cramer's *V* was used as the effect size of the *Chi*^{2}-tests with small effect sizes around .1, medium effect sizes around .3, and large effect sizes around .5. Interrater reliability was calculated using Fleiss kappa with fair agreements around .2, moderate agreements around .4, and substantial agreements around .6

Table 4 Frequencies, percentages, and comparisons of roles assigned to adult female and male subjects in various spelling tests

	Female ($N = 124$)		**Male** ($N = 258$)		**Chi^2-test**				**Interrater reliability**
	Numbers		Numbers						
	Absolute	Relative	Absolute	Relative	χ^2	*df*	*p*	Cramer's *V*	
Professional role					12.63	1	<.001	.18	.76
Mentioned	19	15.3%	84	32.6%					
None mentioned	105	84.7%	174	67.4%					
Parental role					13.97	1	<.001	.19	.79
Mentioned	46	37.1%	50	19.4%					
None mentioned	78	62.9%	208	80.6%					
Leadership role					14.52	2	<.001	.20	.56
Yes—professional	0	0%	20	7.8%					
Yes—leisure	0	0%	8	3.1%					
No	124	100%	230	89.1%					
Domestic work					29.74	2	<.001	.28	.33
Lighter work	14	11.3%	1	0.4%					
Harder work	0	0%	8	3.1%					
None mentioned	110	88.7%	249	96.5%					
Caretaker role					32.89	2	<.001	.29	.37
Nurturing	27	21.8%	10	3.9%					
Activities	1	0.8%	11	4.3%					
None mentioned	96	77.4%	237	91.9%					

Notes: $N = 382$ text units that contained a male or female adult subject. Cramer's *V* was used as the effect size of the *Chi*^{2}-tests with small effect sizes around .1, medium effect sizes around .3, and large effect sizes around .5. Interrater reliability was calculated using Fleiss kappa with fair agreements around .2, moderate agreements around .4, and substantial agreements around .6

In the following analyses, we concentrated only on the female and male characters coded as adults ($n = 382$) for the application of specific roles (Table 4). Here, female characters were statistically significantly more frequently rated in parental roles, engaging in light household work (e.g., shopping, baking, cooking), or playing nurturing caretaker roles than male characters. On the other hand, male characters were more often coded in professional roles, leadership roles (professional and leisure), and heavier domestic household work (e.g., working in the garden). Effect sizes were of small to medium size (Table 4). Interrater agreement was substantial for professional roles and for parental roles and of moderate size for leadership roles, while for domestic work and for caretaker roles, a fair agreement was obtained (Table 4). A deeper analysis of the professional roles demonstrated for all characters ($n = 820$; Table 3) that female characters were more often coded in the context of a caring profession than male characters, while male characters were rated as being more in a science, technology, engineering, or mathematics (STEM) professional context than female characters. We included all male and female characters for this analysis to avoid losing information. For example, some youths were depicted working at an internship, and for many characters, age was not mentioned ($n = 176$). However, our analyses of adults only emphasized the reported gender distribution of the professional contexts: for the caring profession, nine female characters (7.3%) were presented and one male (0.4%). For the STEM professions, no female character was presented and one male (0.4%).

Discusson

To the best of our knowledge, this was the first in-depth analysis of potential gender biases in German-language spelling tests. As German society is continuously moving towards more equality between men and women, we wanted to see how they are portrayed in standardized psychological measures mainly used for 6- to 12-year-olds but also for up to 16-year-olds.

First, there was a difference in the portrayals of men and women in the materials: The study found that male subjects were more often presented than female subjects in the spelling test material (36% vs. 27.3%). Female characters were more likely to be portrayed in parental roles, light domestic homework, nurturing caretaker roles, and caring professions, whereas male subjects were more likely to be portrayed in professional roles, leadership roles, heavier domestic homework, and STEM professions.

While in some cases, the size of the effects was comparatively small, our findings suggest the presence of a general gender bias in the spelling test material, with male subjects being more often represented and additionally portrayed in more stereotypical roles. This is consistent with previous research on gender stereotypes in general and in particular on gender bias in educational materials. In line with results on gender stereotypes in general, female subjects were rated as having more female attributes such as "empathic" or "emotional" and were assigned more positive attributes than male subjects, which is both in line with research that references the stereotype content model (Cuddy et al., 2008). According to gender stereotype beliefs about typical gender-related jobs, females compared to male subjects were not at all depicted in leadership roles and were less often presented in a profession. Regarding results on gender stereotypes in learning material, our results fit well into the existing body of research. For example, Sadker and Sadker (1994) found that male characters were overrepresented in textbooks and other educational materials, and that they were often portrayed in more powerful and active roles than female characters, contributing to an inequitable impression. Also, more recent studies on learning materials (e.g., schoolbooks; Moser & Hannover, 2014) showed that girls and women are still underrepresented in math schoolbooks, and that across all schoolbooks, men are more often presented in terms of their profession than women. For online educational resources in Kerhoven et al. (2016), this gender difference was also obvious and mirrored our results: More men were depicted in science professions than women, and women were more often shown in teaching professions than men.

What are the practical implications of these findings? The gender bias in the spelling test material found in this study could potentially have a number of negative consequences, particularly for female students. Due to the effect of stereotype threat (Steel & Aronson, 1995), female students' information processing performance and related behavior as well as interests might have been impacted. First, gender bias in the spelling tests could reinforce traditional gender stereotypes and limit female students' aspirations. For example, if female students are constantly exposed to images of males in professional and leadership roles, the female students may come to believe that these roles are not open to them or just represent the society "as it is," with relatively rigid assigned roles for men and women. Second, gender bias could lead to female students feeling less confident in their abilities, as had already been demonstrated by the stereotype threat approach (e.g., Good et al., 2010; Hermann & Vollmeyer, 2016) and, relatedly, later career choices. If female students constantly see themselves being portrayed in stereotypical roles, they may come to believe they are less capable than male students. The material studied, which is not used as frequently as textbooks but may have a higher status due to the special conditions of use, could contribute further to this. The question arises whether the construction of materials for spelling tests requires the development of some additional guidelines such as those developed and put forward for the assessment of sex bias and sex fairness in career interest inventories by the NIE (National Institute of Education, 1975) which, therewith, made the used materials more inclusive.

Study 1 had several limitations. First, we restricted the test materials under investigation to those recommended for use in current guidelines (S3; Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V., 2015). However, more spelling tests exist, and these might contain additional useful information. As shown, some of the ratings and categorizations were not as clear as we hoped for. However, one might expect this variation in agreement level. For very simple decisions, such as whether or not to name an occupation, the interrater reliability was correspondingly high. For more difficult decisions, such as for the quality of attributes like positive attributes, the interrater reliability was lower, which in turn might reflect less rigid stereotyping. Also, for some categories like physical appearance, the tests provided little and rather implicit evidence. Hence, there may have been some misclassifications. Finally, Study 1 could not help us say much about how these roles are *perceived*. One open question was whether people truly perceive the portrayal of activities in which men and women are displayed as typical for a man or a woman. So, what happens if we reduce gender-related cues and ask participants to rate how typical it is doing what is described in the item for a man or a woman?

Study 2: stereotype presentation of male and female subjects

Study 2 was based on the results of Study 1. We aimed at analyzing whether the subjects in the spelling tests who were coded as male and female characters in the first study are presented in a gender-stereotypical way. We relied on previous research on the formation and representation of gender stereotypes (Cuddy et al., 2008; Eckes, 2008; Rice, 2014).

Gender stereotypes are defined by socially shared knowledge about the characteristics attributed to women and men. Such knowledge includes socially shared beliefs about typical female or male behavior (e.g., stereotype content mode; Cuddy et al., 2008) and socially shared knowledge about typical activities and contexts in which women or men operate (e.g., Rice, 2014). By psychologically operationalizing gender stereotypes in our study, we considered the context in which the subjects operated and expected the contexts to fit the gendered attributes and activities. In particular, we assumed that (1) female subjects would be presented as more typically female than male, and (2) male subjects would be presented as more typically male than female. We explored these questions with a subset of the test material from Study 1 and took a quantitative approach.

Method

Sample

A total of 160 raters (123 females, 36 males, one non-binary person) provided evaluations for the selected test materials from Study 1. The raters were between 17 and 66 years old

($M = 26.41$, $SD = 11.63$), and their highest educational level was either a university degree (23.8%), high school diploma (64.4%), junior high school diploma (4.4%), or a completed apprenticeship (5.6%). Some raters (1.9%) did not answer this question. The majority of the raters were university students (83.8%).

Procedure

We recruited our raters on social media (e.g., Facebook/Meta) and via physical notices in university buildings. We collected our data in an online survey with SoSci Survey (Leiner, 2014). Our research was designed and conducted in accordance with the code of good practice in online-based testing (Coyne & Bartram, 2006). Raters first answered questions about the gender-stereotypical presentation of our subjects. Then, they provided demographic information about themselves. Participation was voluntary and not compensated, but psychology students could receive course credits. Completion of the questionnaires took approximately 40 min.

Materials

We selected a subset of 134 items from the spelling test material of Study 1 that had already been categorized as containing either a female or male subject (out of 820 total items). We applied the following criteria for the item selection: (a) subjects had to be presented within a word segment in order to provide contextual information (not just single words); (b) the length of the word segments was restricted to a maximum of three sentences in order to reduce reading time; and (c) the spelling test in which the items were embedded had to have a wide range of application. Considering the suitability of test levels, only items stemming from three spelling tests (HSP, SLRT-II, DERET) met these selection criteria.

To conceal gender information, we masked the gender of the subjects in the selected 134 items. For example, we replaced given names, the words "man" and "woman" with the word "person," and the words "father" and "mother" with the word "parent." We did not change any other phrases that did not convey gender information.

For each of the gender-masked 134 items, the raters had to provide two ratings: "How typically male does person A look to you?" and "How typically female does person A look to you?" For each rating, raters used a digital slide control that ranged from 0% ("not at all") to 100% ("very strong").

Results

We measured gender stereotypes with a combination of two direct assessment methods (Eckes, 2020). First, we used the percentage estimation method by asking our participants to rate how typically the neutral person acted or behaved as a female and as a male. Then, we applied the idea of the diagnostic quotient, which means that we could compare the neutral agreement (i.e., we assumed no gender biased ratings and that agreement should be 50%) with the gender-biased agreement (i.e., the actual responses of our participants, and they should be lower or higher than 50%). Dividing the actual responses of our participants by the neutral agreement yielded values below 1 (less typical), around 1 (no bias), or above 1 (more typical). In more than half of the estimates (8 out of 16), these quotients were lower than 1, indicating that participants were not neutral but biased. We expected these biases to be based on participants' gender-related beliefs and stereotypes, as we asked for gender-related agreements in our direct assessment method. Furthermore, the mean differences in agreement between typically male and typically female agreements for both genders differed between half and 1 standard deviation across all tests, and these differences were statistically significant in all cases (Table 5).

We calculated *t*-tests for paired samples to analyze whether the gender-masked word segments whose subjects have been categorized as female before would be rated higher for being presented as typically female than for being presented as typically male. And for gender-masked word segments whose subjects have been categorized as male before, we tested whether they were rated as being presented more typically male than female.

Our results supported our assumptions across all spelling tests and for each spelling test, in particular (Table 5). Gender-masked word segments whose subjects had been categorized as female before were perceived as being presented statistically significantly more typically female than typically male, while gender-masked word segments whose subjects had been categorized as male before were perceived as being presented statistically significantly more typically male than typically female, with predominantly large to medium effect sizes (Table 5). The content chosen by the test authors conveyed these stereotypes, even if the raters did not have any information about the gender of the character in the word segment.

Discussion

Study 1 established that there are differences in the way male and female characters are represented in German spelling tests. Study 2 expanded on this finding: Even when we masked the gender-identifying information in the materials (e.g., names, roles such as mother/father), the participants rating the content rated the portrayed person as being more typically a male/female in accordance with the intended rating (i.e., a caring person was perceived as more typically a woman than a man, and a person working a STEM job was perceived as more typically male). These findings were supported across all spelling tests and for each spelling test, in particular. Hence, the content presented thus seemed to be suitable for activating common gender-related associations or even stereotypes. However,

Table 5 Means, standard deviations, and comparisons (paired *t*-tests) of stereotype female and male ratings for male and female subjects in various spelling tests

	Gender of the subject	Rated as typical female, *M* (SD)	Rated as typical male, *M* (SD)	*T*-test			
				t	df	*p*	*d*
Across all tests	Female	0.48 (0.19)	0.35 (0.17)	19.34	159	< .01	1.52
	Male	0.37 (0.18)	0.47 (0.18)	15.15	159	< .01	1.20
HSP	Female	0.46 (0.19)	0.39 (0.18)	9.13	159	< .01	0.72
	Male	0.38 (0.17)	0.49 (0.18)	13.19	159	< .01	1.04
SLRT-II	Female	0.47 (0.21)	0.32 (0.20)	19.53	159	< .01	1.54
	Male	0.39 (0.21)	0.41 (0.21)	4.00	159	< .01	0.32
DERET	Female	0.51 (0.20)	0.34 (0.17)	17.88	159	< .01	1.41
	Male	0.35 (0.18)	0.49 (0.19)	17.37	159	< .01	1.37

Notes: $N = 160$; d = effect size for paired-samples *t*-tests

HSP Hamburger Schreibprobe/Hamburg spelling probe, *SLRT-II* Salzburger Lese-und Rechtschreibtest/Salzburg reading and spelling test, *DERET* Deutsche Rechtschreibtest/German spelling test when interpreting the results, the gender composition of the current sample must be considered. Given that statistically significantly more male than female raters were included does not necessarily mean that the gender of the rater functioned as a confounding variable. Given that the male and female subsamples were recruited from the same context, we do not expect that a higher number of male raters would provide different results. Nevertheless, future research should especially focus on the sampling method according to the gender of participants. In addition, we did not rely on specific dimensions of gender stereotypes that might be relevant in educational contexts and rather decided to measure gender stereotyping at a general level in order to draw a general conclusion about whether gender stereotyping is present in licensed spelling tests. Our decisions were based on the amount and variety of information; across all spelling tests we selected 134 out of 1295 information units that contained a subject. We suggest that future research should focus on fewer spelling tests and search for specific gender dimensions that may be relevant to educational settings, such as occupational roles.

Such revisions should also take the grammatical gender system of the German language into account. The gender assignment of a noun (i.e., *der* for masculine, *die* for feminine, and *das* for neutral) is often based on its semantic category, but it can also be based on the noun's ending or its historical usage. For nouns that refer to people, the gender assignment is often based on the person's biological sex. For example, the noun (der) *Koch* (a cook) is traditionally masculine, while the noun (die) *Köchin* (a female cook) is feminine. Importantly, if such a term is used in a German-language spelling test, it is clear to the reader whether the authors refer to a male (Koch) or a female (Köchin). However, in spoken German, some nouns in the standard grammatical rules are used in their masculine version but refer to the whole group, especially in the plural form: For example, the noun *Lehrer* (male form of the noun for teacher; female = Lehrerin) is frequently used in the spoken language to refer to both male and female teachers (e.g., a newspaper headline saying that teachers are going on strike would typically refer to them as "Lehrer"). Many people argue that this is incorrect and disrespectful to females and non-binary people and suggest the use of newer variants like *Lehrer:innen* (the ":" should denote that teachers of all genders are signified; one variant is to use an "*": *Lehrer*innen*), or to use a gender-neutral term like *Lehrperson* or *Lehrkraft*. Others argue that, at least, the masculine and feminine version should be used (i.e., talking about *Lehrer* and *Lehrerinnen* if referring to all teachers). These issues around changes in language use are current hot topics in German-speaking countries (e.g., there are civic as well as political and governmental initiatives for or against gender-sensitive language use in official documents or during public broadcasts [1].).

In the context of the spelling tests used, in particular, in official institutions such as schools or conducted by special professional groups with a certain formal reputation such as psychologists, the gender assignment of nouns can be problematic because it can reinforce traditional gender stereotypes. For instance, if spelling test materials only show men in STEM professions and women in caregiving professions, this can send the message to students that these professions are more appropriate for one gender than the other—or could be experienced as exclusive—if items would use the masculine term only even if referring to a whole group of professionals. Considering the findings from our two studies together, it seems advisable to take special care when revising such test material in the future.

[1] In June 2025, German Federal Minister of Education, Karin Prien, has banned civil servants from using gender-neutral language in documents (i.e., using the asterisk).

There is, of course, a broader discussion. Should not the test materials reflect what many children will experience at home or in the educational sector? Even a wealthy country such as Germany has not achieved equality (e.g., in terms of caring or social professions). Hence, such contents could help students connect the spelling test material to their own lives and experiences, making it potentially more engaging and relatable for them and possibly also less irritating or discordant. The discovered differences potentially also describe the noticeable differences in the roles that men and women typically still play, but it might help students develop their critical thinking skills as they learn to question and challenge traditional gender stereotypes. Regarding the latter, it must be said that this (i.e., an intervention) is not the aim of spelling tests. It could even distract from the aim of spelling tests when counter-stereotypical material causes irritation during assessments.

In short, counterarguments against this "test what you see" approach seem to outweigh what would potentially speak in its favor; reinforcing gender stereotypes may have negative impacts. Test materials that send the even-subtle message to students that certain roles and activities are more appropriate for one gender versus the other must be avoided (e.g., Wetzel & Hell, 2012). Moreover, irritation can also arise when gender stereotypes are still present in educational or psychological materials in a society that debates and questions them openly. Therefore, revisionists of these test materials should take these considerations into account. However, this is not a call to avoid gender assignments at all, although we do recommend a balance (i.e., a woman can be a mother cooking *but also* an engineer, and a man can be a father changing diapers *and* a pilot). In the end, we argue for an equal representation of male and female characters in a variety of roles, including both stereotypically male and female roles.

General discussion

Are students confronted with gender stereotypes when being tested for spelling disorders with standardized spelling tests? The two studies described here provided evidence that the answer to this question could be "yes." The results of Study 1 outline that across all five considered tests, male compared to female subjects were more often presented in the spelling test material. Furthermore, female compared to male subjects were more often portrayed in parental roles, light domestic work, nurturing caregiver roles, and caring occupations, while male compared to female subjects were more likely to be displayed in professional roles, leadership roles, heavy domestic work, and STEM occupations. The results of Study 2 are built upon Study 1's findings. Even when the gender-identifying information in the materials was masked (e.g., information on roles such as son/daughter), participants' ratings of the content aligned with the intended rating. These findings were corroborated across all spelling tests and for each spelling test in particular.

One might argue that standardized tests that assess students' spelling abilities might hardly have an impact on students' formation of gender stereotypes because students are not regularly exposed to these tests since they are often not part of regular curricula, and when exposed, only for a short testing period. Nevertheless, when tested, students seem to intensely process the words and texts in order to understand them properly in special test situations. This prolonged process has also been reported by parents who have kids suffering from dyslexia (Plume & Warnke, 2007). Moreover, one might argue that only a few students experience these tests. However, some tests are designed to be used for all students during regular in-class curricula, like the HSP. Furthermore, a gender-stereotype presentation in test material might further reinforce the testing teacher or educational psychologist to act and transmit gender stereotypes in testing situations.

Of course, we could not determine whether the gender bias in the analyzed spelling test materials was intentional or unintentional. However, the findings do suggest that there is a need to be more aware of the potential for gender bias in test materials and their potential impact. What we also do not know at present is whether gender bias is present in test material from other educational contexts, such as the assessment of dyscalculia. Here, too, mathematical operations are embedded in textual material that may reproduce gender stereotypes. Similarly, the assessment of children's general cognitive ability sometimes requires textual material that may reproduce gender stereotypes. Further research is needed to address this gap. Overall, publishers, psychologists, and educators can play a role in reducing or eliminating gender bias by ensuring that male and female characters are equally represented *and* portrayed in non-stereotypical roles.

This research has provided important insights into the gender bias that can exist in educational and psychological materials. However, further research is needed to explore the full range of practical consequences of this gender bias for students. For example, future research could investigate how gender bias in such materials affects students' performance on the measures (e.g., whether there is a stereotype threat; Hermann & Vollmeyer, 2016), but also whether the repeated exposure in test materials and educational materials may have an impact on self-esteem, career aspirations, and choices. Additionally, future research could examine the effectiveness of different interventions and trainings (e.g., guidelines for psychological assessment courses) for reducing or eliminating gender bias in educational materials, and this is in line with a call for greater consideration of these issues in German-language measures for a different field (i.e., the assessment of vocational interests; National Institute of Education, 1975). A further question for future research is whether such test materials can make students who do not identify with traditional gender roles feel uncomfortable or excluded. Even if authors would favor a "test what you see" approach in the development of test materials for spelling tests, the current set of items did not reflect the current state. For example, items that reflected single-parent households were not represented.

It would be better then, for example, to move entirely to neutral material in which people are no longer the focus in terms of content. This would account for the disadvantages of both the "test what you see" and gender-fair approaches. And this would also keep the door open to generating material that is also culture-fair, at the same time. Accordingly, the suggested guidelines might bring us closer to better helping children and adolescents reach their full potential in terms of making their own decisions per interests, studies, and professions and not being constrained by "reproduced gender stereotypes" that might reduce individual potentials by lowering young people's motivation and self-esteem, essential for performing at one's best. In doing so, the guidelines would not undermine the important calls to bring both more women into the science and technology sector (European Commission, 2014) and more men into the caring and teaching sectors (Stuve & Rieske, 2018) in order to meet the variety of work tasks more appropriately.

Conclusion

Our studies can be read as a cautionary tale about the need for continuous revision of standard tests used in psychological assessment. We found good evidence for a potential gender bias in German-language spelling tests. Male and female characters were disproportionately represented in traditional, stereotypical roles. This also translates into seemingly implicit perceptions of what information items used in such tests contain: Even when gender-identifying information was removed, the independent raters still perceived characters in stereotypical ways. A major concern is that such test materials may negatively impact students' self-perception (according to their identity), career aspirations, and overall learning experience. However, these issues can be relatively easily remedied by adapting the wording of the pertinent items.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10212-025-00994-y.

Author contribution All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Nancy Tandler, Felix Peter, Johanna Rimpf, Teresa Wessels, and René T. Proyer. The first draft of the manuscript was written by Nancy Tandler, 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.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abele, A. E., & Bruckmüller, S. (2016). Agency und Communion: Basisdimensionen der sozialen Kognition [Agency and communion: Basic dimensions of social cognitions]. In H.-W. Bierhoff & D. Frey (Eds.), *Enzyklopädie der Psychologie: Sozialpsychologie, Selbst und Soziale Kognition [Encyclopedia of psychology: Social psychology, self and social cognition]* (pp. 409–427). Hogrefe.

Birkel, P. (2007). *Weingartener Grundwortschatz Rechtschreib-Test für zweite und dritte Klassen [Weingarten basic vocabulary spelling test for second and third grades]* (2nd ed.). Hogrefe Verlag.

Birkel, P. (2007). *Weingartener Grundwortschatz Rechtschreib-Test für dritte und vierte Klassen [Weingarten basic vocabulary spelling test for third and fourth grades]* (2nd ed.). Hogrefe Verlag.

Birkel, P. (2007). *Weingartener Grundwortschatz Rechtschreib-Test für vierte und fünfte Klassen [Weingarten basic vocabulary spelling test for fourth and fifth grades]* (2nd ed.). Hogrefe Verlag.

Buckley, C., Farrell, L., & Tyndall, I. (2022). Brief stories of successful female role models in science help counter gender stereotypes regarding intellectual ability among young girls: A pilot study. *Early Education and Development, 33*(4), 555–566. https://doi.org/10.1080/10409289.2021.1928444

Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend. (2023) Anteil von Frauen und Männern an den erfolgreich abgelegten Abschlussprüfungen für das Lehramt Primarbereich. [Students grouped by age, ISCED education level and gender.] https://www.daten.bmfsfj.de/daten/daten/anteil-von-frauen-und-maennern-an-den-erfolgreich-abgelegten-abschlusspruefungen-fuer-das-lehramt-primarbereich-134444. Accessed 17th June 2025.

Bundesministerium für Forschung, Technologie und Raumfahrt. (2025). Studierende nach Alter, ISCED-Bildungsbereichen und Geschlecht. [Students by age, ISCED education level and gender]. https://www.datenportal.bmbf.de/portal/de/K254.html. Accessed 16th June 2025.

Brotman, J. S., & Moore, F. M. (2008). Girls and science: A review of four themes in the science education literature. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 45*(9), 971–1002. https://doi.org/10.1002/tea.20241

Coyne, I., & Bartram, D. (2006). Design and development of the ITC guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*, 133–142.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Destatis Statistisches Bundesamt. (2024). Mehr als ein Drittel der Studienanfängerinnen und -anfänger im MINT-Bereich sind Frauen. [Over a third of first-year STEM students are women.] https://www.destatis.de/DE/Presse/Pressemitteilungen/2024/01/PD24_N003_213.html. Accessed 17th June 2025.

Deutsche Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e.V. (2015). Diagnostik und Behandlung von Kindern und Jugend-lichen mit Lese- und / oder Rechtschreibstörung [Diagnosis and treatment of children and adolescents with reading and / or spelling disorders]. https://register.awmf.org/assets/guidelines/028-044l_S3_Lese-Rechtschreibst%C3%B6rungen_Kinder_Jugendliche_2015-06-abgelaufen.pdf. Accessed 16th June 2025.

Eckes, T. (2008). Geschlechterstereotype: Von Rollen, Identitäten und Vorurteilen [Gender stereotypes: About roles, identities and prejudices]. In R. Becker & B. Kortendiek (Eds.), *Handbuch Frauen- und Geschlechterforschung* (pp. 171–182). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91972-0_20

Eckes, T. (2020). Messung von Stereotypen [Assessment of stereotpyes]. In L.E. Petersen & B. Six (Eds.), *Stereotype, Vorurteile und soziale Diskriminierung: Theorien, Befunde und Interventionen* (2nd ed., pp. 99–108). Beltz.

European Commission (2014). *Reaching gender equality in science, technology, engineering and mathematics.* https://ec.europa.eu/commission/2014-2019/moedas/announcements/reaching-gender-equality-science-technology-engineering-andmathematics_en. *Accessed 5 Sept 2023*

Fiske, S. T. (1998). Stereotyping, prejudice and discrimination. In D. T., Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 357–414). McGraw-Hill.

Glötzner, J. (1982). Ist ein mathematisches Weib wider der Natur? Heidi häkelt Quadrate, Thomas erklärt die Multiplikation. Rollenklischees in neuen Mathematikbüchern [Does a mathematical female contradict nature? Heidi crochets squares and Thomas explains multiplication. Clichés in recent mathematics books]. In I. Brehmer (Ed.), *Sexismus in der Schule. Der heimliche Lehrplan der Frauendiskriminierung [Sexism at school. The hidden curriculum of women's discrimination]* (pp. 150–158). Beltz.

Good, J. J., Woodzicka, J. A., & Wingfield, L. C. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology, 150*(2), 132–147. https://doi.org/10.1080/00224540903366552

Guichot-Reina, V., & De la Torre-Sierra, A. M. (2023). The representation of gender stereotypes in Spanish mathematics textbooks for elementary education. *Sexuality & Culture, 27*(4), 1481–1503. https://doi.org/10.1007/s12119-023-10075-1

Hellinger, M. (1980). For men must work and women must weep: Sexism in English language textbooks used in German schools. *Women's Studies International Quarterly, 3*, 267–275.

Hermann, J. M., & Vollmeyer, R. (2016). Stereotype threat in der grundschule [Stereotype threat in primary school]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 48*, 42–49. https://doi.org/10.1026/0049-8637/a000143

Ibrahimovic, N., & Bulheller, S. (2017). *Rechtschreibtest RST-ARR (3rd ed.).* Pearson Assessment & Information.

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

Kerger, S., Martin, R., & Brunner, M. (2011). How can we enhance girls' interest in scientific topics? *British Journal of Educational Psychology, 81*(4), 606–628. https://doi.org/10.1111/j.2044-8279.2011.02019.x

Kerkhoven, A. H., Russo, P., Land-Zandstra, A. M., Saxena, A., & Rodenburg, F. J. (2016). Gender stereotypes in science education resources: A visual content analyses. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0165037

Kite, M. E., & Whitley, B. E. (2016). *Psychology of prejudice and discrimination (3rd ed.).* Routledge.

Landis, J. R., & Kock, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Leiner, D. J. (2014). SoSci Survey (Version 2.4.00-i) [Computer software]. https://www.sosciurvey.de. Accessed 5 Apr 2020

Martinez Méndez, R., Schneider, M., & Hasselhorn, M. (2015). *DERET 5–6+ Deutscher Rechtschreibtest für fünfte und sechste Klassen* [DERET 5–6+ German spelling test for fifth and sixth grades.]. Hogrefe Verlag.

May, P. (2010). *HSP 1–9 Diagnose orthografischer Kompetenz HSP 1–9* [Diagnosis of orthographic competence]. Ernst Klett Verlag.

Mayring, P. (2008). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis. Basics and techniques]. Weinheim, Deutschland: Beltz Verlag.

Middendorff, E., Apolinarski, B., Poskowsky, J., Kandulla, M., & Netz, N. (2017). *Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2016* [The economic and social situation of students in Germany2016]. https://www.dzhw.eu/pdf/sozialerhebung/21/Soz21_hauptbericht_barrierefrei.pdf. Accessed 5 Sept 2023

Moll, K., & Landerl, K. (2010). *SLRT-II Lese- und Rechtschreibtest. Weiterentwicklung des Salzburger Lese- und Rechtschreibtests* [SLRT-II reading and spelling test. Further development of the Salzburg reading and spelling test]. Verlag Huber.

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., & Schulte-Körne, G. (2014). Specific learning disorder: Prevalence and gender differences. *PLOS One*, *9*. https://doi.org/10.1371/journal.pone.0103537

Moser, F., & Hannover, B. (2014). How gender fair are German schoolbooks in the twenty first century? An analysis of language and illustrations in schoolbooks for mathematics and German. *European Journal of Psychology and Education, 29*, 387–407. https://doi.org/10.1007/s10212-013-0204-3

National Institute of Education. (1975). Guidelines for the assessment of sex bias and sex fairness in career interest inventories. *Measurement & evaluation in guidance, 8*, 7–11.

Ochman, J. M. (1996). The effects of nongender-role stereotyped, same-sex role models in storybooks on the self-esteem of children in grade three. *Scholarly Journals, 35*, 711–735. https://doi.org/10.1007/BF01544088

Ohlms, U. (1984). "Und drinnen waltet die züchtige Hausfrau…" Das Mädchen- und Frauenbild in Grundschulbüchern ["And inside rules the modest housewife…" The portrayal of girls and women in elementary schoolbooks]. In I. Brehmer & U. Enders-Dragässer (Eds.), *Die Schule lebt – Frauen bewegen die Schule [School is alive—Women set school in motion]* (pp. 131–161). DJI Deutsches Jugendinstitut.

Plume, E., & Warnke, A. (2007). Definition, Symptomatik, Prävalenz und Diagnostik der Lese-Rechtschreib-Störung [Definition, symptomatology, prevalence and diagnosis of reading spelling disorder]. *Monatsschrift Kinderheilkunde, 155*, 322–327. https://doi.org/10.1007/s00112-007-1480-2

Rice, C. (2014). *Science in balance*. http://curt-rice.com/2014/07/24/infographic-are-stereotypes-keeping-women-away-from-science/

Rohrmann, T., & Thoma, P. (1998). *Jungen in Kindertagesstätten: Ein Handbuch zur geschlechtsbezogenen Pädagogik* [Boys in day care centers: A handbook on gender-related education]. Lambertus.

Sadker, M., & Sadker, D. (1994). *Failing at fairness: How America's schools cheat girls*. Charles Scribner's Sons.

Six-Materna, I. (2020). Sexismus [Sexism]. In L.-E. Petersen & B. Six (Eds.), *Stereotype, Vorurteile und soziale Diskriminierung: Theorien, Befunde und Interventionen* (2nd ed., pp. 136–152). Beltz.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Steinmayr, R., Weidinger, A. F., Heyder, A., & Bergold, S. (2019). Warum schätzen Mädchen ihre mathematischen Kompetenzen geringer ein als Jungen [Why do girls rate their mathematical competencies lower than boys]? *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie, 51*, 71–83. https://doi.org/10.1026/0049-8637/a000213

Stock, C., & Schneider, W. (2008a). *DERET 1–2+ Deutscher Rechtschreibtest für das ersteund zweite Schuljahr* [DERET 1–2+ German spelling test for the first and second school year]. Hogrefe Verlag.

Stock, C., & Schneider, W. (2008b). *DERET 3–4+ Deutscher Rechtschreibtest für das dritteund vierte Schuljahr*. [DERET 3–4+ German spelling test for the third and forth school year]. Hogrefe Verlag.

Stuve, O., & Rieske, T. V. (2018). *Männer ins Grundschullehramt* [Men into elementary school teaching]. Coburg.

United States Census Bureau. (2023). *Sex by age by field of bachelor's degree for first major for the population 25 years and over*. https://data.census.gov/table/ACSDT1Y2023.B15011?q=McMullan+Marian+Major+Attorney&g=010XX00US&y=2023. Accessed 17th June 2025.

Wetzel, E., & Hell, B. (2012). NIE-Richtlinien zur Gender Fairness von Interessentests: Sind deutschsprachige Interessentests gender-fair? Eine qualitative Analyse [Are German interest tests gender-fair? A qualitative analysis]. *Zeitschrift Für Arbeits- und Organisationspsychologie, 56*, 37–47.

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes. A multination study (rev.ed.)*. Sage.

Lewis, J., Schneegans, S., & Straza, T. (2021). *UNESCO science report: The race against time for smarter development*. UNESCO Publishing.

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

Nancy Tandler. Department of Psychology at the Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Current themes of research:

Playfulness in youth and children, Assessment of personality in children.

Most relevant publications:

Tandler, N. & Dalbert, C. (2020). Always look on the bright side of students: Does valence of teacher perceptions relate to students' educational performance? *Social Psychology of Education, 23*, 1121–1147. https://doi.org/10.1007/s11218-020-09573-z.

Proyer, R. T., & Tandler, N. (2020). An update on the study of playfulness in adolescents: Its relationship with academic performance, well-being, anxiety, and roles in bullying-type situations. *Social Psychology of Education, 23*, 73–99. https://doi.org/10.1007/s11218-019-09526-1.

Felix Peter. Department of School Psychology, Local State School Administration of Saxony-Anhalt, Halle (Saale), Germany

Current themes of research:

Belief in a just world, climate crises and mental health outcomes.

Most relevant publications:

Peter, F., Dalbert, C., Kloeckner, N., & Radant, M. (2013). Personal belief in a just world, experience of teacher justice, and school distress in different class contexts. *European Journal of Psychology of Education, 28*, 1221–1235. https://doi.org/10.1007/s10212-012-0163-0.

Johanna Rimpf. Department of Psychology at the Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Current themes of research:

No previous publications.

Most relevant publications:

No previous publications.

Teresa Wessels. Department of Psychology at the Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Current themes of research:

No previous publications.

Most relevant publications:

No previous publications.

René T. Proyer. Department of Psychology at the Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

Current themes of research:

Positive psychology, Assessment of personality, Playfulness in adults and youth.

Most relevant publications:

Proyer, R. T., & Brauer, K. (2023). Assessing playfulness: Current challenges and overview. In W. Ruch, A. B. Bakker, L. Tay & F. Gander (Eds.), *Handbook of positive psychology assessment* (pp. 145-161). Hogrefe.

Proyer, R. T., Gander, F., Brauer, K., & Chick, G. (2021). Can playfulness be stimulated? A randomised placebo-controlled online playfulness intervention study on effects on trait playfulness, well-being, and depression. *Applied Psychology: Health and Well-Being, 13*, 129–151. https://doi.org/10.1111/aphw.12220.

Authors and Affiliations

Nancy Tandler[1] · Felix Peter[2] · Johanna Rimpf[1] · Teresa Wessels[1] · René T. Proyer[1]

✉ Nancy Tandler
nancy.tandler@psych.uni-halle.de

[1] Department of Psychology, Martin-Luther-University Halle-Wittenberg, Emil-Abderhalden-Straße 26-27, 06108 Halle (Saale), Germany

[2] Department of School Psychology, Local State School Administration of Saxony-Anhalt, Halle (Saale), Germany