Educational Psychology Review (2023) 35:17
https://doi.org/10.1007/s10648-023-09716-6

META-ANALYSIS

The Influence of Diagnostic Labels on the Evaluation of Students: a Multilevel Meta-Analysis

David J. Franz[1] · Tobias Richter[1] · Wolfgang Lenhard[1] · Peter Marx[1] · Roland Stein[1,2] · Christoph Ratz[1,3]

Accepted: 26 August 2022 / Published online: 8 February 2023

Abstract

Research suggests that children suffering from different types of disorders (learning disorders, behavioral disorders, or intellectual disabilities) are sometimes evaluated differently simply due to the presence of a diagnostic label. We conducted a multilevel meta-analysis of experimental studies (based on data from 8,295 participants and on 284 effects nested in 60 experiments) to examine the magnitude and robustness of such label effects and to explore the impact of potential moderators (type of evaluation, diagnostic category, expertise, student's gender, and amount and type of information). We found a moderately negative overall label effect (Hedges' $g = -0.42$), which was robust across several types of evaluation, different samples, and different diagnostic categories. There was no indication that expertise and the gender of the child moderated the effect. Presenting participants with only a label yielded the strongest negative effect of $g = -1.26$, suggesting that the effect was dependent on the amount of information being presented to participants. We conclude that labeling a child can exacerbate negative academic evaluations, behavioral evaluations, evaluations of personality, and overall assessments of the child. Further implications for theory and future research are discussed.

Keywords Behavioral disorders · Evaluation of children · Intellectual disabilities · Learning disorders · Negative label effects · Teacher biases · Stereotypes · Stigma

✉ David J. Franz
david.franz@uni-wuerzburg.de

✉ Tobias Richter
tobias.richter@uni-wuerzburg.de

[1] Department of Psychology IV, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany

[2] Department of Special Education V, University of Würzburg, Wittelsbacherplatz 1, 97074 Würzburg, Germany

[3] Department of Special Education IV, University of Würzburg, Wittelsbacherplatz 1, 97074 Würzburg, Germany

Introduction

Teachers often have to work with students that face special challenges that are psychological in nature, such as problems with paying attention, problems with understanding the fundamentals of arithmetic and spelling, or difficulties in social interactions. If the severity of such difficulties exceeds certain thresholds, children usually are referred to a psychologist or psychiatrist who initiates formal diagnostic processes. Eventually, a child might be assigned to a diagnostic category, such as attention deficit hyperactivity disorder (ADHD), learning disorder, or conduct disorder. Formal diagnoses are often a necessary step toward allocating resources for remedial interventions, which may include seeking social support and strategies for coping with the problem (Lenhard et al., 2005). However, one possible downside might be that diagnoses can function as labels that amplify teachers' negative expectations about the child (Jussim et al., 1994). For example, a teacher's academic expectation about a student, who faces considerable difficulties in arithmetic and spelling, might become even worse after the teacher is told that the student has been diagnosed with a learning disorder (Minner, 1982; Minner & Prater, 1984; Franz et al., 2021). Simultaneously, it is also possible that the diagnostic label has a positive impact, for example by increasing people's acceptance of the student's problems (Fernald & Gettys, 1980).

The purpose of this meta-analysis was to synthesize the existing experimental literature on effects that such diagnostic labels can have on how children are evaluated. We define a negative label effect as a more negative evaluation of a child that is caused exclusively by the presence of a diagnostic label. For example, if a teacher is confronted with two children that suffer from the exact same problems while only one child is diagnosed, the teacher would evaluate the diagnosed child worse than the undiagnosed one. A positive label effect, in contrast, occurs if a label leads to a more positive evaluation. Since the majority of studies that we analyzed either explores the impact of a learning disorder diagnosis (i.e., a child suffers from considerable difficulties in one or more areas of learning), or a behavioral disorder diagnosis (i.e., a child shows a pattern of disruptive behaviors that cause emotional and social problems), or an intellectual disability diagnosis (i.e., a child suffers from considerably impaired cognitive functioning), we aimed to explore whether label effects can differ between these types of disorders. Furthermore, we aimed to explore whether different evaluators (e.g., students, regular teachers, special education teachers, mental health workers) of the diagnosed children are prone to label effects to the same extent. Finally, we investigated possible moderators of the effect, such as the student's gender, the overarching diagnostic category of the label, or the kind of evaluation (e.g., academic vs. behavioral vs. personality evaluation), and the point in time when the study was carried out. The practical purpose of this analysis was to investigate the extent that diagnostic labels carry a negative surplus-meaning that might lead to disadvantages for the interaction of professionals with students.

Why is this meta-analysis important? The existence of labeling effects in the school context is widely suspected and is almost common sense in quite a few

areas of research, from stereotype research to expectancy effects in the classroom. Against this background, it is somewhat surprising that a systematic meta-analysis on the topic has not yet been conducted. Apart from a comprehensive but older discussion of research on the mental retardation label[1] (MacMillan et al., 1974) and a short narrative review of research from 1970 through 2000 on the effects of the learning disability label (Osterholm et al., 2011), no research synthesis has been published on labeling effects. The purpose of the current research was to fill this research gap. A systematic quantitative synthesis of the available research seems to be especially relevant as the extant studies do not provide an unequivocal and homogeneous picture. Moreover, the literature spans 60 years, during which societal changes have taken place that have also altered our perception of learning disorders, emotional problems, intellectual disabilities, and other diagnoses that children receive. Therefore, quantitative estimates of the labeling effect in general are needed, plus an investigation of conditions that might affect the direction or the magnitude of labelling effects. These results would be highly informative with regard to extant theories that predict labeling effects. Moreover, they would also be of great practical importance. For educational, psychological, and medical practitioners, it is crucial to know whether negative labelling effects occur, how big a problem they are, and what conditions affect their magnitude. Answers to these questions would also provide a starting point for developing effective measures to counter negative labelling effects.

Negative Effects of Mental-Disorder Labels

Although our meta-analysis is mainly focused on label effects caused by diagnoses of learning disorders, behavioral disorders, or intellectual disabilities, it is insightful for a start to examine the large theoretical and empirical literature on negative effects caused by mental disorder diagnoses (i.e., affective, anxiety, eating, personality, and psychotic disorders). Scholars have argued that mental disorder diagnoses can be the cause of stereotypes (i.e., beliefs or cognitive schemas about people suffering from mental illness), prejudice (i.e., evaluative reactions towards mentally ill persons), and discrimination (i.e., overt negative behavior towards the mentally ill) (e.g., Corrigan, 2007; A. B. Fox et al., 2018; Link et al., 1989; Rüsch et al., 2005). Empirical research suggests that people with psychological problems are often perceived to be incapable, childish, weak-minded, or dangerous (Curcio & Corboy, 2019; Jorm et al., 2012; Rüsch et al., 2005).

Several lines of research have yielded evidence for the central role that diagnostic labels play in the stigmatization of mentally ill people. For example, Angermeyer and Matschinger (2005) found an association between the self-imposed description

[1] Please note that the label "mental retardation" was accepted language use at that time; it is no longer in use today, for good reasons. Here and in the remainder of the article, we cite the labels originally used in the studies that we describe or discuss in the text. Furthermore, when presenting our own arguments and analyses, we use the modern term "intellectual disability."

of another person as schizophrenic and the tendency to perceive that person as dangerous and unpredictable, which elicited the desire to maintain social distance from the labeled person. In two experiments, female silhouettes were judged to be more alike and similar in weight when they were sorted into categories of eating-disorder labels (Foroni & Rothbart, 2011, 2013), which is evidence that the diagnostic labels fostered a stereotypical perception of the silhouettes. Carrizosa-Moog et al. (2019) and Cutler and Ryckman (2019) reported experimental evidence for negative label effects caused by different clinical labels, such as delusional disorder, schizophrenia, bipolar disorder, major depressive disorder, alcohol use disorder, and epilepsy. In addition, both studies showed that speaking about mentally ill people in a manner that identifies patients with their disorder (e.g., "He is an epileptic" or "She is delusional") can lead to even more negative label effects. Finally, a recent meta-analysis showed that the psychopathic label can lead to harsher punishments, to a higher level of perceived dangerousness, and to a more skeptical view on the amenability to treatment of the perpetrator compared to an assessment of an unlabeled perpetrator (Berryessa & Wohlstetter, 2019).

In sum, negative label effects have been documented for a wide range of mental disorder labels and there is also evidence that people with intellectual disabilities can be the target of similar stigmatization (i.e., the whole process whereby stereotypes lead to prejudice and discrimination; Ditchman et al., 2013). Although mental disorders differ in many respects from typical diagnostic categories that are associated with lower academic performance, such as learning disorders, learning disability, or emotional problems, negative effects have been documented for such labels, too, as will be discussed next.

Negative Effects of Labels in the School Context

One can interpret the well-known Pygmalion effect (Rosenthal & Jacobson, 1968) as a kind of label effect (or a consequence thereof). In a typical study on this effect, students perform better in standardized tests after their teacher had been told that these students have a special potential for developing their cognitive abilities. The sole description of a child as having special potential can change the teacher's behavior toward the child and thereby have a positive impact on the child's academic performance. The flipside of the Pygmalion effect is the so-called Golem effect, which refers to negative effects associated with teacher expectations and the ensuing self-fulfilling prophecy. Although the findings regarding such negative effects of teacher expectations are somewhat mixed (Jussim & Harber, 2005; Madon et al., 2011), studies have produced considerable evidence that they may depend on a host of characteristics ascribed to students, such as their ethnicity, social class, and, most important in the present context, diagnostic labels (for a review, see Rubie-Davies, 2009).

In a typical study on the effects of diagnostic labels in the school context, teachers receive a written vignette or watch a video that portrays the behavior of a child, possibly enriched with additional information. For example, teachers might watch a video or read a written vignette about a child labeled emotionally

disturbed, learning disabled, or behaviorally disordered (or receive the same video or written vignette without the label) and then provide judgments of the child's personality, skills, or further academic development. In several studies of this kind, the label led to a more negative evaluation of the child (e.g., Foster et al., 1975; Foster & Ysseldyke, 1976; Jacobs, 1978; Johnson & Blankenship, 1984; Thelen et al., 2003). Nonetheless, other studies using similar designs found no negative labelling effect (e.g., Cornett-Ruiz & Hendricks, 1993; Fernald et al., 1985; Tournaki, 2003) or negative labelling effects only for specific dependent variables, labels or presentation formats (e.g., Allday et al., 2011; Dukes & Saudargas, 1989; Franz et al., 2021; Shuller & McNamara, 1976). Thus, across studies, the pattern regarding negative labeling effects is somewhat heterogeneous, raising the questions of the generalizability and potential moderating or boundary conditions of the effect.

Generalizability and Potential Moderators of Label Effects

Exploring potential moderators of negative label effects can help to identify the underlying causal factors and to develop effective interventions for mitigating those effects. In the following section, we use extant theory and research to substantiate our research aims and to explain why certain moderators might influence label effects.

Type of Evaluation

Stereotypes connected to labels are likely to influence how children are evaluated (Levy et al., 1998). For example, stereotypes can guide teachers' evaluation of students' academic performance, classroom behavior, and personalities (Rubie-Davis, 2009). Arguably, diagnostic labels and the associated stereotypes are likely to not affect every kind of evaluation to the same extent. For example, the ADHD label might have a bigger impact on behavioral evaluations, due to the strong behavioral stereotype associated with this label (e.g., Jussim et al., 2000), than the dyslexia label, whose associated stereotype is focused on lower academic achievement (e.g., Knight, 2021). The diagnosis of an oppositional defiant disorder might influence the evaluation of personality in other ways than the diagnosis of a mild intellectual disability. To illustrate these considerations with examples from the literature, Rolison and Medway (1985) asked participants to estimate how often a boy's test scores would exceed the school district average on the next 20 tests. The label "educable mentally retarded" had a negative impact on this evaluation whereas the label "learning disability" had not. Allday et al. (2011) found that the label "oppositional defiant disorder" led participants to judge a child's behavior to be more disturbing, while the label "gifted and talented" had the opposite effect. Interestingly, the label "ADHD" had no effect.

Diagnostic Category

The diagnostic labels that are in the focus of this meta-analysis can be sorted into broad different categories, such as learning disorders, behavioral disorders, or intellectual disabilities. Label effects might differ depending on the diagnostic category. For example, intellectual disability labels might cause considerably more extreme effects than labels from the other categories because the diagnosis of an intellectual disability implies that the affected child has a significantly impaired mental functioning in general (Foster & Ysseldyke, 1976; Rolison & Medway, 1985). Beyond that, there might be more nuanced differences between different diagnostic categories. For example, because learning disorder labels often suggest specific difficulties (e.g., the dyslexia label implies only difficulties in the area of reading and spelling), effects of learning disorder might be limited to academic evaluations (Thelen et al., 2003; Franz et al., 2021). Moreover, accommodations for learning disorders are often provided by means of supportive measures in regular schools (such as 504 Educational Plans in the U.S.) rather than special education, which might affect the severity and breadth of the perceived difficulties associated with learning disorders. Behavioral disorder labels and intellectual disability labels, in contrast, suggest difficulties that are more extensive. Consequently, their impact might be less limited (Foster et al., 1980; Parish et al., 1979; Thelen et al., 2003).

Expertise

People who are very knowledgeable about disorders might not rely on the presence of a label as a heuristic for drawing broad conclusions about the child. They might be aware of the complexity of every clinical condition and the fact that each affected child has its unique history and combination of challenges. In contrast, people less educated about clinical conditions might deploy more simple heuristics that lead them to interpret a label as indicative of substantial difficulties in the child. However, one might also argue that specialists (e.g., special education teachers or psychologists) might be more affected by diagnostic labels because routinely relying on diagnoses in their evaluation of children is an important part of their training and their daily practice.

Although expertise can be acquired throughout one's occupational career and can, therefore, vary between different representatives of the same occupation, it is reasonable to assume that different occupations (e.g., regular teaching, special education, and health care) on average come along with different levels of expertise. Accepting occupation as a rough proxy for expertise leads to the following reasoning. If expertise is associated with a reduction of negative label effects, teachers with work experience should be less affected by labeling than teacher students, special education teachers should be less affected than regular teachers, and highly trained psychologists or psychiatrists might be even less affected than special education teachers. However, existing research on these matters is inconclusive. In one group of studies, diagnostic labels negatively affected the evaluation of children by

special education teachers, psychologists, and psychiatrists (J. D. Fox & Stinnett, 1996; Moberg, 1995; Shuller & McNamara, 1976; Sutherland & Algozzine, 1979; Thurman et al., 1994), whereas other studies found no negative label effects in these occupational groups (Graham & Leone, 1987; Javel & Greenspan, 1983; Pfeiffer, 1980). A second group of studies that compared samples with different occupational directly yielded mixed results. Some of these studies provided evidence that negative label effects were stronger in teachers than in psychologists and psychiatrist (Carroll & Reppucci, 1978) and stronger in regular teachers than in special education teachers (Johnson & Blankenship, 1984; Vlachou et al., 2014). However, several studies found no difference between regular teachers and special education teachers (Bianco, 2005; Bianco & Leech, 2010; Gillung & Rucker, 1977; Minner et al., 1987; Salvia et al., 1973; Taylor et al., 1983) or between education students and teachers with work experience (Combs & Harper, 1967; Ohan et al., 2011; Taylor et al., 1983; Thelen et al., 2003), whereas another study suggested that teachers with work experience were even more susceptible than education students (Foster et al., 1980). Moreover, Parish et al. (1979) found no evidence that the educational level and the amount of mainstreaming experience of teachers mattered for teacher's susceptibility to label effects.

Gender of the Student

People might associate certain disorders more with females or with males. A good example for this is ADHD and the associated stereotypes. Given the greater prevalence of ADHD in boys than in girls (with a male/female ratio of 2:1 to 3:1, Cuffe et al., 2005), behavior that is indicative of ADHD might be more strongly associated with boys. Therefore, people might interpret the presence of the ADHD label in boys as more indicative of problematic behavior than the presence of the same label in girls (Fresson et al., 2019). However, the opposite might also be true. Given the lower frequency of ADHD in girls than in boys, the label might come as a surprise when it is given to girls and thereby lead to a more negative evaluation. In line with this explanation, Eisenberg and Schneider (2007) found that negative effects of the ADHD label were more pronounced when girls were evaluated. In contrast, two experimental studies found no evidence for gender differences in the effect of the ADHD label (Batzle et al., 2010; Ohan et al., 2011), and one study yielded inconclusive results (Lee et al., 2019).

Amount and Type of Information

Label effects might be more pronounced when people have little information about a child. When information is relatively sparse, the label might be more salient. Since category salience increases stereotyping (Rees et al., 2020), people might evaluate the child in light of the typical problems associated with a disorder especially when they have no other information than the disorder of the child. Conversely, the impact of the label might be much smaller when there is rich information present. It has been shown that enhancing knowledge about a stereotyped group can reduce

stereotypes about that group (Pettigrew & Tropp, 2008). Furthermore, teachers' expectations can be shaped by a large variety of variables, such as students' socio-economic status, gender, ethnicity, and various personal characteristics (Wang et al., 2018), which further suggests that a label becomes less influential the more additional information is known about a student.

There is some evidence supporting this line of reasoning. Several studies have found negative label effects when participants were presented with short written texts about students or just with the label but found considerably weaker or no label effects when participants watched videotapes depicting the students (Fernald et al., 1985; Fogel & Nelson, 1983; Reschly & Lamprecht, 1979). One could argue that video material provides more comprehensive and more ecologically valid information about children than brief texts do. Consequently, these studies suggest that labels have negative effects only in cases of limited information. This argument finds further (indirect) support in studies that used video presentations only and found no evidence for negative label effects (Cornett-Ruiz & Hendricks, 1993; Yoshida & Meyers, 1975). However, empirical evidence also speaks against a moderating role of the amount and type of information. First, one study that compared written texts and videos found some evidence for negative label effects when videos were used (M. A. Stanley & Comer, 1988). Second, several studies found no differences in label effects between presentation of information via text and video (Foster et al., 1975; Foster et al., 1980; Foster & Keech, 1977; Foster & Ysseldyke, 1976; Jacobs, 1978). Finally, some studies, in which participants were presented with videos only, found negative label effects (Foster et al., 1976; Johnson & Blankenship, 1984; Thurman et al., 1994).

Additional Study Characteristics

The potential role of diagnostic labels in the formation of stigma was extensively discussed in the 1970s and 1980s in psychiatry and special education (e.g., Link et al., 1989; MacMillan et al., 1974), which could have led to more sensitivity regarding negative effects of diagnostic labels. Moreover, effect sizes can also vary depending on publication date because of changing standards in methods or time-specific confounding variables. In some areas of research, the effects found in earlier studies tend to be larger than those found in later studies, which may have failed to replicate the earlier findings (Ioannidis & Trikalinos, 2005). Thus, publication date is a moderator of substantial interest.

In addition, it is also possible that label effects vary to some extent depending on the socio-cultural background of participants. Stereotypes and prejudice about diagnosed children might be more prevalent in some countries than in others, which is why the moderating role of sample nationality should be investigated.

Finally, the study design could have an impact on label effects. For example, deploying a within-subjects design and asking participants to evaluate the same child twice could come along with carry-over effects (i.e., the first evaluation influences the second). In a between design, in which participants evaluate the child only

once, such carry-over effects cannot occur. Consequently, it is important to compare between-subjects and within-subjects designs.

The Present Research

The purpose of the present research was to conduct a comprehensive meta-analysis of experimental studies that investigated label effects caused by psychological diagnoses on the evaluation of students. We concentrated on experimental studies because the net effect of the label can only be isolated by manipulating the presence of a label experimentally while keeping other information constant (e.g., the child's behavior, grades, problems, etc.).

The first aim of this meta-analysis was to test the assumption that a negative effect of diagnostic labels on the evaluation of students exists. The second aim was to clarify the role of potential moderators that might influence the label effect.

Method

Literature Search

The literature search and selection of studies was conducted by the first author using the search string *(label teacher expectation) OR (label teacher) OR (diagnosis label) OR (labeling children with diagnosis)* in the databases PsycInfo and ERIC. Furthermore, the first author also screened every publication that was relevant for the analysis (see Selection Criteria) for citations of further studies. In addition, the first author searched in PsycInfo and Google Scholar for publications that cited studies that were already included in the study pool. The search ended in February 2022.

Selection Criteria

We included studies in the meta-analysis that met the following criteria:

1. The presence of a diagnostic label was manipulated experimentally between or within participants. In addition, there was no confounding factor that covaried with the presence of a label (e.g., grades, performance, behavior, information about the disorder, etc.).
2. The label was stated explicitly for participants in the experimental condition.
3. There was a comparison between a condition with a label present and a condition without a label or with a condition in which the child was labeled as "normal".
4. Children or teenagers (age < 21 years) were the targets of labeling and evaluation.
5. The participants that evaluated the children were adults.
6. The study reported at least one dependent variable that could be interpreted as an evaluation of the child.

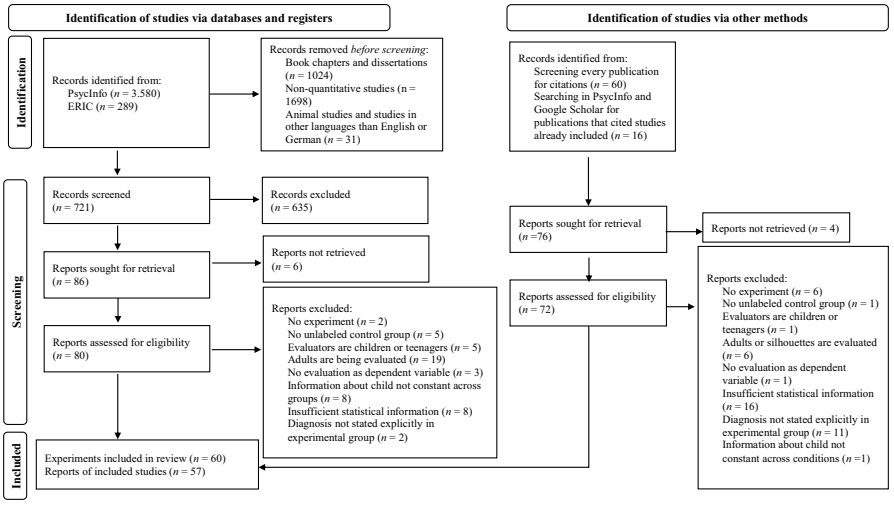

Fig. 1 PRISMA flow diagram of literature search

If a study's sample was composed of adults and children (e.g., Cornett-Ruiz & Hendricks, 1993), we included the study and selectively calculated the effect only for the adult part of the sample. However, for one study with a mixed sample, it was not possible to calculate the effect size for the adults separately because of insufficient information (Thelen et al., 2003). We decided against excluding this study to avoid loss of information and calculated the effect size based on the whole sample.

Several studies that met our inclusion criteria could not be included in the analysis because the reported statistics were insufficient for the calculation of effect sizes. Because most of these studies were published many years ago (the oldest dating back to 1974), it seemed unlikely that we could retrieve the missing information in every case. Considering the general rule of publication ethics to retain data for ten years (American Psychological Association, 2020), we contacted only authors of studies that were published no more than ten years ago. Of the three authors contacted, two responded that they did not have the data anymore and the third did not reply.

Based on the selection criteria, 60 experiments reported in 57 publications were included. Details of the literature search are provided in the PRISMA flow diagram in Fig. 1.

Coded Variables

Moderator variables and additional study characteristics were coded individually for each effect size reported in the articles that met the inclusion criteria.

Type of Evaluation

An initial literature screening yielded a huge variety of different dependent variables. To deal with this complexity, we assigned all effects to one of the following 15 categories:

1. behavioral abnormality: evaluations of abnormalities in behavior or prognoses about future problematic behavior ($k = 63$ effect sizes),
2. performance expectations: expectations about a student's future performance in specific tests or tasks ($k = 9$),
3. willingness to work with the student: assessments of one's own willingness to work with the student in class or on specific tasks ($k = 14$),
4. expectations for academic future: general expectations about the student's academic future (e.g., graduation, success at the university, career success) ($k = 5$),
5. evaluations of social integration or social behavior: assessments of the student's social integration in class or peer group, or assessments of the student's problems regarding social behavior ($k = 10$),
6. evaluations of self-competence in handling the student: assessments of the adult participant's ability to deal with the student's problems ($k = 7$),
7. evaluations of personality: general assessments of student's personality (e.g., via broad trait terms) ($k = 18$),
8. cause of the student's problems: attributions of student's problems to certain factors (e.g., luck, ability, or task difficulty) ($k = 40$),
9. evaluations of student's task performance: assessments of a student's performance in specific tasks (e.g., evaluation of a student's essay) ($k = 6$),
10. recommendations for a gifted program: assessments of a student's eligibility for taking part in programs for gifted children ($k = 10$),
11. recommendations for educational placement: evaluations of the appropriate educational placement for the student (e.g., regular class vs. special education) ($k = 3$),
12. evaluations of treatment strategies: assessments of the usefulness of various treatments for the student ($k = 20$),
13. evaluations of academic skills: assessments of the student's skills that are important for academic success (e.g., intelligence) ($k = 13$),
14. overall assessment: an overall rating of the subject's impression of the child or a global evaluation score composed of various evaluations ($k = 30$),
15. other: evaluations that fit none of the 14 categories ($k = 36$).

However, several of these categories included very few effect sizes, which would be problematic for performing a moderator analysis. Thus, we aggregated some of them. The categories performance expectations, expectations for academic future, evaluations of student's task performance, recommendations for a gifted program, recommendations for educational placement, and evaluations of academic skills were combined into the category academic evaluations (serving as reference category in the meta-regression analyses; $k = 46$). Behavioral

abnormality and evaluations of social integration/social behavior were combined into the category behavioral evaluations ($k = 73$). Willingness to work with the student and evaluations of self-competence to handle the student were combined into the category attitudes towards the child ($k = 21$).

Label

Some of the 32 different labels that we identified in the studies comply with the current terminology, but other labels are no longer used because of their offensive nomenclature. We retained these terms to preserve the original language of the primary studies. The following labels were used in the primary studies: learning disabled or learning disability (serving as reference category, $k = 32$), educable mentally retarded or EMR-class student ($k = 14$), dyslexia ($k = 8$), dyscalculia ($k = 3$), specific learning disability in the language area ($k = 3$), developmental delays and learning problems ($k = 1$), behavior disorder or behaviorally disordered or behaviorally disturbed ($k = 2$), conduct disordered or conduct disorder ($k = 18$), behaviorally/emotionally impaired ($k = 1$), emotionally disturbed or emotional disturbance or seriously emotionally disturbed ($k = 16$), emotional and behavioral disorder ($k = 2$), ADHD ($k = 26$), ADHD with stimulant treatment ($k = 3$), hyperactive syndrome ($k = 6$), hyperkinetic syndrome ($k = 3$), history of hostile aggressive behavior ($k = 6$), oppositional defiant disorder ($k = 2$), mentally retarded or mental retardation ($k = 53$), mild mental retardation or mildly retarded or marginally retarded ($k = 3$), mentally deficient ($k = 1$), developmentally delayed ($k = 2$), minimal brain dysfunction ($k = 5$), mentally retarded or backward children/feeble-minded ($k = 1$), physical disability or physically handicapped ($k = 3$), sexually abused ($k = 21$), gifted or gifted and talented[2] ($k = 8$), socially maladjusted ($k = 3$), schizophrenic or schizophrenia ($k = 1$), cerebral palsied or cerebral palsy ($k = 8$), psychopathic or psychopathy ($k = 19$), speech deficit ($k = 2$), autism disorder ($k = 4$), and Asperger's disorder ($k = 4$).

Diagnostic Category

We assigned every label used in a study to one of the following three diagnostic categories: learning disorders (serving as reference category; $k = 46$), behavioral disorders ($k = 88$), and intellectual disabilities ($k = 77$). All labels that into none of these three categories were assigned to the category "other" ($k = 73$).

We initially planned to include the label as a separate moderator, but we identified 32 different labels in total that were used in the studies with very uneven numbers of effect sizes (see Additional Study Characteristics). Consequently, a meaningful

[2] One might argue that the "gifted" label is different from the other labels because it is indicative of special potential and not indicative of deficiencies. From this might follow that the "gifted" label causes only positive effects. Consequently, one might question the inclusion of this label in our meta-analysis. However, one could also assume that this label is associated with prejudice about gifted children (e.g., gifted children being considered as especially socially incompetent). Moreover, since our focus is on diagnostic labels and "gifted" is a diagnostic label, it seems justified to include it in our analysis.

moderator analysis could not be performed with that many different labels. Therefore, we sorted all labels into more fine-grained subcategories to further search for differences between different types of diagnoses: learning disabilities only (serving as reference category; $k = 46$), combined intellectual disabilities and learning disorders ($k = 15$), behavioral disorders only ($k = 28$), combined behavioral and emotional disorders ($k = 19$), ADHD ($k = 38$), intellectual disabilities only ($k = 60$), mental disorders ($k = 28$), and other ($k = 50$).

Expertise

To explore the different impact diagnostic labels might have on groups with varying degrees of expert knowledge and experience, we coded participants as students (i.e., participants enrolled in university courses, serving as reference category; $k = 97$), non-students (participants were coded as non-students if they had graduated from university or if they were enrolled in university courses but had at least one year of teaching experience; $k = 136$), or mixed (students and non-students; $k = 49$). Orthogonally to this categorization, we further coded participants as teachers only (student teachers included, serving as reference category; $k = 156$), non-teachers only ($k = 105$), or teachers and non-teachers combined ($k = 19$). For two effects, coding was not possible because of missing information. Next, again orthogonally to the previous classifications, we coded whether participants were regular teachers only (all types of teachers, including student teachers but not special education teachers, serving as reference category; $k = 107$), special education teachers only (students of special education or teachers working particularly in special education; $k = 11$), regular and special education teachers combined ($k = 47$), or other ($k = 115$). For two effects, coding was not possible. Finally, and again orthogonally to the previous classifications, we coded participants as mental health workers only (psychologists, physicians, social workers, nurses, serving as reference category; $k = 44$), non-mental health workers ($k = 220$), or mental health workers and other occupational groups ($k = 18$). For two effects, coding was not possible.

Gender of Student

We coded whether the students being evaluated were males only (reference category; $k = 170$), females only ($k = 27$), males and females ($k = 54$), or whether the children's gender was not specified ($k = 33$).

Provided Information

To record the way in which participants were provided with information about the students, we coded whether participants were given vignettes describing the student to be evaluated (reference category; $k = 176$), videos ($k = 40$), only a label ($k = 25$), other stimuli (e.g., photos; $k = 17$), or a combination of stimuli from these four categories ($k = 26$).

Nationality

We coded whether the study was conducted in the U.S. ($k = 252$), Germany ($k = 12$), UK ($k = 5$), China ($k = 1$), Finland ($k = 1$), or Canada ($k = 11$). Because the number of effects sizes from studies conducted outside the US was very small, a meaningful moderator analysis with all countries was not possible. For this reason, we classified the effects as coming from U.S. (reference category; $k = 252$) and non-US samples ($k = 32$) to investigate whether the heavy reliance on samples from the US has an impact on the magnitude of the effects.

Study Design

Combining effect sizes from studies with a between-subjects design and effect sizes from studies with a within-subjects design can be problematic because the extent they are comparable is debatable (Morris & DeShon, 2002). Therefore, meta-analyses should explore whether the two designs result in systematic differences. For this purpose, we coded whether an effect was based on a between-subjects design (reference category; $k = 253$) or a within-subjects design ($k = 31$).

Year of Publication

Negative label effects might change over time, either because of societal changes over the last decades or because of methodological advances in the field of study. To examine whether negative label effects have changed over time, we recorded the year of publication and centered it around the mean (1990) for the meta-regression analysis.

Effect Size Calculation

Polarity

We calculated Hedges' g for every effect. A negative g indicates that a labeled child was evaluated worse than the unlabeled peer (negative label effect), whereas a positive g indicates a more positive evaluation of the labeled child (positive label effect). In most cases, deciding whether an effect was positive or negative was straightforward (e.g., when the probability of failing in a future test was estimated to be higher for a labeled child, the effect was negative). However, additional considerations and specifications were necessary in several cases to determine the polarity of an effect. In measuring participants' evaluation of a child's personality, one study deployed the Big Five personality framework (Baudson & Preckel, 2013). We regarded higher ratings of openness, agreeableness, conscientiousness, and extraversion as positive and higher ratings of neuroticism as negative. In other studies, participants were asked to identify the causes of a child's problems (Kesterson, 2013; O'Donohue & O'Hare, 1997; Severence & Gasstrom,

1977; M. A. Stanley & Comer, 1988; Weisz, 1981). We determined the polarity of these effects in line with considerations based on an attributional theoretical framework (e.g., Allen et al., 2020; Mezulis et al., 2004). If participants attributed a labeled child's problems more to external, local, or unstable causes (e.g., bad luck, effort), we recorded a positive effect. Such an attribution implies that the causes of a child's failure are not permanent, and that the child has potential for improvement. If, in contrast, participants attributed a labeled child's problems more to internal, global, or stable causes (e.g., skill, talent), we regarded the effect as negative, since this implies that the child's condition can hardly be changed.

In two studies, participants' task was to make a recommendation for the educational placement of the child (Javel & Greenspan, 1983; Taylor et al., 1983). These recommendations were based on a continuum that ranged from special education only to complete integration into regular education class. We regarded recommendations that trended toward special education as indicating that from the viewpoint of the participant, the student lacked the necessary skills for regular class. Therefore, we coded these effects as negative. Conversely, recommendations of regular education were regarded as positive.

Finally, in some other studies, participants' task was to provide an evaluation of the appropriateness of certain treatment strategies for the child (Jones & Cauffman, 2008; Murrie et al., 2005; Parish et al., 1979; Rockett et al., 2007; Stinnett et al., 2001). In the context of these studies, we interpreted the recommendation of a certain treatment (e.g., mental health services) for a child with a label (e.g., conduct disorder) as positive because such a recommendation is indicative of participants' belief that the child's condition is treatable.

Within-Subjects Effects

Different methods of calculating within-subjects effects for meta-analyses have been proposed (Borenstein et al., 2009; Lakens, 2013; Morris & DeShon, 2002). Morris and DeShon (2002) discussed two possibilities of standardizing within-subjects effects for the purpose of meta-analysis: using the pooled standard deviation or the pretest standard deviation. Considering that the studies on the label effect are not based on a pre-posttest design, we opted for using the pooled standard deviation approach. We needed the correlation between both measurements for every effect to calculate the effect sizes but none of the articles reported this correlation. Thus, we used formulas by Morris and DeShon (2002) to compute the correlations from the means, standard deviations, sample sizes, and *F*-values for ten effects. For ten effects, this was possible. For studies that did not report enough information to calculate the correlation, we followed recommendations by Morris and DeShon (2002) to estimate a correlation. We meta-analyzed the ten correlations that we had already computed and used the average correlation ($r = .32$) for the computation of within effects from studies that did not report sufficient information to calculate the correlation. Following this procedure, we were able to compute further 18 within-subjects

effects. Finally, we converted all within-subjects effects to the metric of between-subjects effects using formulas proposed by Morris and DeShon (2002).

Between-Subjects Effects

We first computed Cohen's *d* for effects that were based on a between-subjects design. These effect sizes were computed using means and standard deviations or using other statistics (e.g., *F*-values, sample size per group, η^2) in cases in which information was incomplete. Next, we converted all between-subjects effects (and the within-subject effects converted into the metric of between-subject effects) to Hedges' *g* and computed the sampling variances (Borenstein et al., 2009; Lakens, 2013).

For studies that reported that an effect was not significant without reporting sufficient statistics for effect size calculation, we registered the effect to be exactly zero. In cases of effects that were significant according to the authors but could not be computed based on the reported statistics, we excluded the effect from the analysis. The reader should note that excluding significant effects due to missing information while inserting zero for non-significant effects with missing information is a very conservative approach. We opted against imputing a specific value for the significant effects to prevent the mean effect from being biased, for example by potential influences of publication bias or questionable research practices (e.g., Francis, 2012; Rosenthal, 1979; Schäfer & Schwarz, 2019; Simmons et al., 2011).

Coding Procedure

We developed a comprehensive coding manual to instruct two student research assistants about our research questions and the coding strategies. The first author coded all studies, one student assistant coded 55 studies and the second assistant coded two studies. Interrater agreement was satisfactory to excellent (Cohen's κ for categorical data ranging from .69 to .99 and the ICC [absolute agreement] for metric data ranging from .89 to .99), except on two cases, which were the results of misunderstandings. All discrepancies could be resolved through discussion.

Meta-analytical Strategies

Most studies included in this meta-analysis provided more than one effect size. There were multiple effect sizes from the same study because researchers measured several dependent variables on the same sample or because they manipulated the presence of several labels by comparing more than one experimental group to the same control group in one experiment. Multiple effect sizes from one experiment depend on each other, which can lead to biased estimations in meta-analyses (Borenstein et al., 2009; Bosnjak & Viechtbauer, 2009). For this reason, we deemed a multilevel approach to be appropriate for dealing with the dependencies in our data (Assink & Wibbelink, 2016; Cheung, 2019; Moeyaert et al., 2017; Scammacca et al., 2014; van den Noortgate et al., 2013; van den Noortgate & Onghena, 2003).

By implementing a three-level mixed-effects meta-analytic model, we were able to differentiate between sampling variance (Level 1), variance between effect sizes within experiments (Level 2), and variance between experiments (Level 3). Beyond that, we estimated the overall effect with a random-effects model and analyzed the potential impact of moderators with mixed-effects models (Borenstein et al., 2009).

Moderator effects a in meta-analysis are examined by estimating and testing the effect(s) of one predictor or several predictors on the effect size. Moderator effects were analyzed in two ways: (1) We tested the impact of single categorical moderators on the magnitude of labeling effects based on the *Q*-Test (Borenstein et al., 2009, Ch. 19). If this analysis yielded a significant effect, we also estimated and tested the labelling effect in each subgroup. (2) We then estimated and tested the impact of multiple moderators (categorical and metric) in a metaregression model (Borenstein et al., 2009, Ch. 20). For the meta-regression, we selected only those moderators that turned out to be significant in the single moderator analysis. In this way, we could estimate and test the unique effect of each moderator, controlling for the effects of other moderators.

All analyses were conducted with the metafor package in R (Viechtbauer, 2010) and with an adapted version of the R-code and the approach recommended by Assink and Wibbelink (2016).

We applied the Knapp and Hartung (2003) adjustment as recommended by Assink and Wibbelink (2016) for estimating the overall effect and for the moderator analyses. In this analysis, statistical testing of singular coefficients is based on the *t* distribution. Overall tests are based on the *F* distribution, and degrees of freedom equal the number of coefficients (numerator) and the total number of effect sizes minus the number of coefficients in the model (denominator).

For the meta-regression models, we calculated *Q* statistics (Cochran, 1954). Cochran's Q_B is calculated as the sum of the squared deviations of each study's effect size from the overall effect size weighted by the inverse of the within-study variance. It follows a χ^2 distribution (with $df = k$ - 1). Q_B indicates whether the variance of effect sizes deviates significantly from zero. In a similar manner, Q_M can be computed to test whether a significant amount of variance is explained by the model.

We also estimated I^2 within clusters of dependent effects (I^2_{within}), I^2 between effects based on independent samples ($I^2_{between}$), and R^2_{within} and $R^2_{between}$ (Cheung, 2014). I^2_{within} indicates the proportion of the total variability of effects that can be attributed to heterogeneity within clusters of dependent effects, whereas $I^2_{between}$ indicates the proportion of the total variability of effects that can be attributed to heterogeneity between effects based on independent samples. R^2_{within} and $R^2_{between}$ quantify the proportion of variance explained by the predictors within clusters of dependent effects and between independent effects.

Results

We obtained 284 effect sizes that were nested in 60 experiments (57 publications) and were based on 8,295 participants. The papers were published between 1962 and 2021. The typical effect included in the meta-analysis was based on a

Fig. 2 Forest plot of effect sizes for different subgroups. Effect sizes (Hedges' *g*) are depicted with 95% ▶ confidence intervals

between-participants design with an average sample size of 118 ($Mdn = 77$, $SD = 154$, range: 6-1,114). Effect sizes (g) ranged from -4.43, which suggests a markedly more negative evaluation of a labeled child compared to an unlabeled one, to 1.66, which is indicative of a considerably more positive evaluation of the labeled child. Data and R-Code underlying all results and comprehensive forest plots are available at the repository of the Open Science Framework (https://osf.io/g72nt/?view_only=d80b63e8c4084bd28629ed9a81414df8).

Overall Effect of Labeling and Heterogeneity

We found an overall negative label effect size of $g = -0.42$ ($k = 284$, $t(283) = -5.76$, $p < .001$, 95% CI [−0.56, −0.28], Fig. 2). The overall test of heterogeneity was significant, $Q(284) = 2{,}307.87$, $p < .001$, indicating considerable variability of effect sizes between studies and effects. We further estimated a series of unconditional models to test whether the model variance at Level 2 and Level 3 was significant. For this purpose, we first compared a two-level model in which the variance at Level 2 was fixed to zero with the three-level model that included all three levels. Results indicated that the fit of the three-level model was significantly better than the fit of the two-level model (Likelihood-Ratio-Test; $\chi^2(1) = 831.53$, $p < .001$). Next, we compared a two-level model in which the variance at Level 3 was fixed to zero with the three-level model. Again, the fit of the three-level model was significantly better than the fit of the two-level model (Likelihood-Ratio-Test; $\chi^2(1) = 39.55$, $p < .001$). Thus, we found significant heterogeneity both between effects sizes within studies and between studies, suggesting a multilevel approach for the data. Furthermore, we explored how variance was distributed over the three levels of the model by computing I^2 for Level 1 ($I^2 = .0715$), Level 2 ($I^2 = .4559$), and Level 3 ($I^2 = .4726$). From these results, we can conclude that 7.15% of the total variance can be attributed to Level 1, i.e., sampling variance, 45.59% to Level 2, i.e., variance between effects within studies, and 47.26% to Level 3, i.e,. variance between studies.

These variance proportions correspond to estimated variances of 0.03 for Level 1, 0.21 for Level 2, and 0.22 for Level 3. To give an illustration of the magnitude of these variance estimates in relation to the overall labelling effect of $g = -0.42$, we may say that the expected percentage of studies yielding negative label effects and the expected percentage of negative label effects within studies are both approximately 82%, under the assumption that the effect sizes are normally distributed.

Analyses of Single Moderators

Type of Evaluation

We found significant differences between different types of evaluation ($k = 284$, $F(7, 276) = 5.10$, $p < .001$). Label effects were most negative and significantly

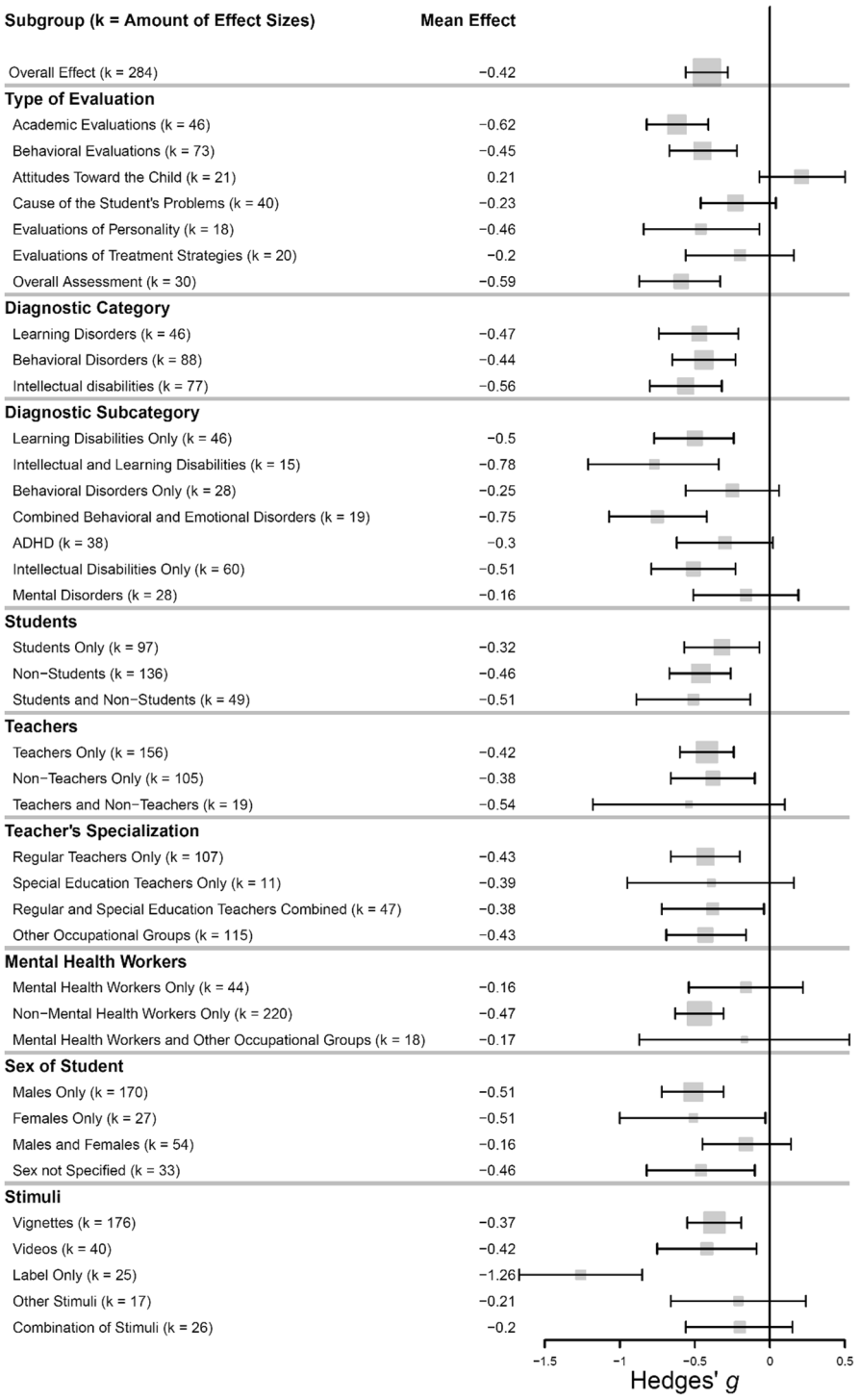
Subgroup (k = Amount of Effect Sizes)
Mean Effect
Overall Effect (k = 284)
−0.42
Type of Evaluation
Academic Evaluations (k = 46)
−0.62
Behavioral Evaluations (k = 73)
−0.45
Attitudes Toward the Child (k = 21)
0.21
Cause of the Student's Problems (k = 40)
−0.23
Evaluations of Personality (k = 18)
−0.46
Evaluations of Treatment Strategies (k = 20)
−0.2
Overall Assessment (k = 30)
−0.59
Diagnostic Category
Learning Disorders (k = 46)
−0.47
Behavioral Disorders (k = 88)
−0.44
Intellectual disabilities (k = 77)
−0.56
Diagnostic Subcategory
Learning Disabilities Only (k = 46)
−0.5
Intellectual and Learning Disabilities (k = 15)
−0.78
Behavioral Disorders Only (k = 28)
−0.25
Combined Behavioral and Emotional Disorders (k = 19)
−0.75
ADHD (k = 38)
−0.3
Intellectual Disabilities Only (k = 60)
−0.51
Mental Disorders (k = 28)
−0.16
Students
Students Only (k = 97)
−0.32
Non−Students (k = 136)
−0.46
Students and Non−Students (k = 49)
−0.51
Teachers
Teachers Only (k = 156)
−0.42
Non−Teachers Only (k = 105)
−0.38
Teachers and Non−Teachers (k = 19)
−0.54
Teacher's Specialization
Regular Teachers Only (k = 107)
−0.43
Special Education Teachers Only (k = 11)
−0.39
Regular and Special Education Teachers Combined (k = 47)
−0.38
Other Occupational Groups (k = 115)
−0.43
Mental Health Workers
Mental Health Workers Only (k = 44)
−0.16
Non−Mental Health Workers Only (k = 220)
−0.47
Mental Health Workers and Other Occupational Groups (k = 18)
−0.17
Sex of Student
Males Only (k = 170)
−0.51
Females Only (k = 27)
−0.51
Males and Females (k = 54)
−0.16
Sex not Specified (k = 33)
−0.46
Stimuli
Vignettes (k = 176)
−0.37
Videos (k = 40)
−0.42
Label Only (k = 25)
−1.26
Other Stimuli (k = 17)
−0.21
Combination of Stimuli (k = 26)
−0.2
−1.5
−1
−0.5
0
0.5
Hedges' g

Table 1 Mean effects and statistics for different types of evaluation

	N	*k*	*g*	*t*(*df*)	*p*	95% CI
Academic Evaluations	26	46	−0.62	−5.87(276)	< .001	[−0.82, −0.41]
Behavioral Evaluations	17	73	−0.45*	−3.87(276)	< .001	[−0.67, −0.22]
Attitudes Toward the Child	10	21	0.21	1.48(276)	.139	[−0.07, 0.50]
Cause of the Student's Problems	8	40	−0.23	−1.67(276)	.096	[−0.46, 0.04]
Evaluations of Personality	7	18	−0.46	−2.35(276)	.019	[−0.84, −0.07]
Evaluations of Treatment Strategies	4	20	−0.20	−1.09(276)	.277	[−0.56, 0.16]
Overall Assessments	16	30	−0.59	−4.44(276)	< .001	[−0.85, −0.33]
Other Evaluations	12	36	−0.37	−2.81(276)	.005	[−0.64, −0.11]

Note. *N* number of experiments, *k* number of effects, *g* Hedges' *g*

Table 2 Mean effects and statistics for diagnostic categories and diagnostic subcategories

	N	*k*	*g*	*t*(*df*)	*p*	95% CI
Diagnostic Categories						
Learning Disorders	17	46	−0.47	−3.50(280)	< .001	[−0.74, −0.21]
Behavioral Disorders	23	88	−0.44	−4.15(280)	< .001	[−0.65, −0.23]
Intellectual Disabilities	20	77	−0.56	−4.56(280)	< .001	[−0.80, −0.32]
Other Disorders	15	73	−0.17	−1.44(280)	.151	[−0.40, 0.06]
Diagnostic Subcategories						
Learning Disabilities Only	17	46	−0.50	−3.76(276)	< .001	[−0.77, −0.24]
Combined Intellectual and Learning Disabilities	7	15	−0.77	−3.48(276)	< .001	[−1.21, −0.34]
Behavioral Disorders Only	7	28	−0.25	−1.60(276)	.110	[−0.56, 0.06]
Combined Behavioral and Emotional Disorders	10	19	−0.75	−4.52(276)	< .001	[−1.07, −0.42]
ADHD	10	38	−0.30	−1.88(276)	.061	[−0.62, 0.02]
Intellectual Disabilities Only	14	60	−0.51	−3.60(276)	< .001	[−0.79, −0.23]
	-0.062	-0.062	−0.049	0.020	0.020	0.020
Mental Disorders	5	28	−0.16	−0.89(276)	.375	[−0.51, 0.19]
Other Disorders	11	50	−0.08	−0.52(276)	.601	[−0.36, 0.21]

Note. *N* number of experiments, *k* number of effects, *g* Hedges' *g*

different from zero for academic evaluations ($g = -0.62$) and overall assessments ($g = -0.59$, Table 1 and Fig. 2). Effects of labeling on behavioral evaluations ($g = -0.45$), on personality evaluations ($g = -0.46$), and on other evaluations ($g = -0.37$) were less negative and differed significantly from zero, too (Table 1 and Fig. 2). Non-significant effects were found for attitudes toward the child ($g = 0.21$), evaluations of the cause of the student's problems ($g = -0.23$), and evaluations of treatment strategies ($g = -0.20$, Table 1 and Fig. 2).

Table 3 Mean effects and test statistics for different types of samples

	N	*k*	*g*	*t*(*df*)	*p*	95% CI
Students Only	23	97	−0.32	−2.56(281)	.011	[−0.57, −0.07]
Non-Students	29	136	−0.46	−4.42(281)	< .001	[−0.67, −0.26]
Students and Non-Students Combined	8	49	−0.51	−2.67(281)	.008	[−0.89, −0.13]
Teachers Only	40	156	−0.42	−4.56(280)	< .001	[−0.60, −0.24]
Non Teachers Only	16	105	−0.38	−2.68(280)	.008	[−0.66, −0.10]
Teachers and Non-Teachers Combined	3	19	−0.54	−1.65(280)	.100	[1.18, 0.10]
Regular Teachers Only	25	107	−0.43	−3.75(279)	< .001	[−0.66, −0.20]
Special Education Teachers Only	5	11	−0.39	−1.41(279)	.160	[−0.95, 0.16]
Regular and Special Education Teachers Combined	11	47	−0.38	−2.18(279)	.030	[−0.72, −0.04]
Other Occupational Groups	18	115	−0.43	−3.13(279)	.002	[−0.69, −0.16]
Mental Health Workers Only	8	44	−0.16	−0.80(280)	.423	[−0.54, 0.22]
Non-Mental Health Workers Only	49	220	−0.47	−5.75(280)	< .001	[−0.63, −0.31]
Mental Health Workers and Other Occupational Groups	2	18	−0.17	−0.48(280)	.635	[−0.87, 0. 53]

Note. *N* number of experiments, *k* number of effects, *g* Hedges' *g*

Diagnostic Category

Overall, we found no significant differences between the four superordinate diagnostic categories ($k = 284$, $F(3, 280) = 2.43$, $p = .066$, Fig. 2). Descriptively, intellectual disabilities yielded the most negative effects ($g = -0.56$, Table 2 and Fig. 5 in the Online Supplement), followed by learning disorders ($g = -0.47$, Table 2 and Fig. 6 in the Online Supplement), and behavioral disorders ($g = -0.44$, Table 2 and Fig. 7 in the Online Supplement). Effects of labels in the residual category were the smallest and the only ones that were not different from zero ($g = -0.17$, Table 2).

Apart from the superordinate category results, the overall test indicated significant differences between diagnostic subcategories ($k = 284$, $F(7, 276) = 2.30$, $p = .027$). Label effects in several diagnostic subcategories were significantly different from zero, with the most negative effects caused by combined intellectual and learning disability labels ($g = -0.77$), followed by combined behavioral and emotional disorder labels ($g = -0.75$), intellectual disability only labels ($g = -0.51$), and learning disability only labels ($g = -0.50$, Table 2 and Fig. 2). The remaining subcategories yielded non-significant effects. Descriptively, effects were the most negative for ADHD labels ($g = -0.30$), followed by behavioral disorder only labels ($g = -0.25$), effects of mental disorder labels ($g = -0.16$), and labels in the residual category ($g = -0.08$, Table 2 and Fig. 2).

Table 4 Mean effects and statistics for gender of student and different amounts and types of information

	N	*k*	*g*	*t*(*df*)	*p*	95% CI
Gender of Student						
Males Only	29	170	−0.51	−4.99(280)	< .001	[−0.72, −0.31]
Females Only	5	27	−0.51	−2.08(280)	.039	[−1.00, −0.03]
Males and Females	15	54	−0.16	−1.06(280)	.291	[−0.45, 0.14]
Gender not Specified	11	33	−0.46	−2.54(280)	.012	[−0.82, −0.10]
Amount and Type of Information						
Vignettes	33	176	−0.37	−4.95(279)	< .001	[−0.55, −0.19]
Videos	12	40	−0.42	−2.48(279)	.014	[−0.75, −0.09]
Label Only	6	25	−1.26	−6.02(279)	< .001	[−1.67, −0.85]
Other Stimuli	9	17	−0.21	−0.93(279)	.352	[−0.66, 0.24]
Combination of Stimuli	5	26	−0.20	−1.12(279)	.263	[−0.56, 0.15]

Note. *N* number of experiments. *k* number of effects, *g* Hedges' *g*

Expertise

We found no significant differences between samples consisting of students and other groups ($k = 284$, $F(2, 281) = 0.51$, $p = .599$). Effects were negative and significantly different from zero for all different groups. Descriptively, the most negative effects were found in samples consisting of students and non-students combined ($g = -0.51$, Table 3 and Fig. 2). The most positive effects were found in samples consisting of students only ($g = -0.32$), whereas effects in samples consisting of non-students were intermediate ($g = -0.46$, Table 3 and Fig. 2).

We also found no differences between teacher samples and other occupational groups ($k = 284$, $F(3, 280) = 0.15$, $p = .929$). Effects were significantly different from zero in samples consisting of teachers only ($g = -0.42$) and in samples consisting of non-teachers only ($g = -0.38$, Table 3 and Fig. 2), with the former effects being more negative than the latter. Effects in samples consisting of teachers and non-teachers combined were the most negative, although they were not significantly different from zero ($g = -0.54$, Table 3 and Fig. 2).

The analysis revealed no significant differences between different types of teachers ($k = 284$, $F(4, 279) = 0.08$, $p = .988$). Negative label effects were very similar and significantly different from zero in regular teachers only ($g = -0.43$), regular and special education teachers combined ($g = -0.38$), and other occupational groups ($g = -0.43$, Table 3 and Fig. 2). Although label effects in special education teachers only were comparable in size, they were not significantly different from zero ($g = -0.39$, Table 3 and Fig. 2).

Finally, no significant differences were found between mental health workers and other groups of participants ($k = 284$, $F(3, 280) = 1.01$, $p = .388$). Label effects in samples of non-mental health workers only were negative and significantly different from zero ($g = -0.47$), in contrast to label effects in mental health workers only ($g = -0.16$) and in samples of mental health workers and other occupational groups combined ($g = -0.17$, Table 3 and Fig. 2).

Gender of Student

Student's gender was not a significant moderator ($k = 284$, $F(3, 280) = 1.36$, $p = .257$). Label effects on the evaluation of boys ($g = -0.51$) and girls ($g = -0.51$) were identical and effects without specified gender were slightly more positive ($g = -0.46$, Table 4 and Fig. 2). Effects in all these three groups were significant, whereas label effects in mixed groups (both males and females) were descriptively more positive and not significantly different from zero ($g = -0.16$, Table 4 and Fig. 2).

We expected that especially the impact of the ADHD label can differ depending on the gender of the labeled child. Therefore, we reran the moderator analyses with the subset of effects that were based on ADHD labels. This analysis, however, also yielded no differences between boys, girls, children without specified gender, and children of both genders ($k = 38$, $F(3, 34) = 0.97$, $p = .416$).

Amount and Type of Information

We found significant differences between the various ways of presenting information to participants ($k = 284$, $F(4, 279) = 5.29$, $p < .001$). Three types of stimuli yielded effects that differed significantly from zero. Experiments in which only a label was mentioned yielded highly negative effects ($g = -1.26$) followed by video-based effects ($g = -0.42$) and vignette-based effects ($g = -0.37$, Table 4 and Fig. 2). Effects based on experiments with a combination of stimuli ($g = -0.20$) and effects based on experiments with other stimuli ($g = -0.21$, Table 4 and Fig. 2) were even more positive and were not significantly different from zero.

Additional Study Characteristics

Nationality

Since there were not enough effects from different countries to investigate the moderating role of specific nationalities, we could only compare US studies with studies from outside the US. We found no significant difference between studies being conducted inside ($k = 252$) and outside the US ($k = 32$).

Study Design

Effects based on a between-subjects design ($k = 253$) or a within-subjects design ($k = 31$) were not significantly different.

Year of Publication

We investigated the impact of publication year because label effects might have changed over time either because of societal developments or because of

Table 5 Multilevel mixed-effect meta-regression models estimating the impact of publication year, amount and type of information, and type of evaluation

	Full Model	Outlier Model 1	Outlier Model 2
Intercept	−0.742***	−0.712***	−0.686***
Publication Year	0.016**	0.014**	0.011*
Amount and Type of Information (Reference Category: Vignette)			
Video	0.233	0.336	0.097
Label Only	−0.792***	−0.536*	−0.126
Other Stimuli	0.575*	0.541*	0.458*
Combination of Stimuli	0.278	0.278	0.364*
Type of Evaluation (Reference Category: Academic Evaluations)			
Behavioral Evaluations	0.211^{+}	0.205^{+}	0.210^{+}
Attitudes Toward the Child	0.871***	0.847***	0.770***
Cause of the Student's Problems	0.447**	0.447**	0.441***
Evaluations of Personality	0.348^{+}	0.195	0.050
Evaluations of Treatment Strategies	0.398*	0.394*	0.412*
Overall Assessments	0.146	0.183	0.299*
Other Evaluations	0.225	0.144	0.248*
Q_B	1996.74***($df = 271$)	1901.31*** ($df = 269$)	1453.73*** ($df = 261$)
Q_M	74.27***($df = 12$)	60.29*** ($df = 12$)	45.56*** ($df = 12$)
I^2_{within}	52.87%	46.01%	45.19%
$I^2_{between}$	39.02%	44.84%	42.38%
R^2_{within}	.02	.25	.45
$R^2_{between}$	.25	.24	.47

Note. $^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed)

methodological changes. Publication year significantly moderated the label effect ($F(1, 282) = 7.81$, $p = .006$). The regression coefficient was positive ($b = 0.013$), indicating that more recent studies found less negative effects.

Analyses of Multiple Moderators: Meta-regression Model

Some of the moderators that had a significant impact on label effects in the singular moderator analysis might overlap to some extent and might therefore be partially redundant. To test whether this might be the case, we estimated a multi-level mixed-effects meta-regression model with the significant moderators from the singular moderator analysis (publication year, type and amount of information, and type of evaluation, Table 5). In this model, publication year was still a significant moderator ($b = 0.016$, $SE = 0.01$, $t(271) = 2.99$, $p = .003$). The model estimated the negative label effect to become more positive about .016 per year, from which follows that the effect was estimated to become more positive 0.16 every decade, and altogether 0.94 over the 59 years that the analyzed research has been published.

Beyond that, the influence of the amount and type of available information on the label effect was somewhat comparable to the results of the singular moderator analysis. Compared with the reference category of vignette-based effects, effects based on the sole mention of a label still were significantly more negative ($b = -0.79$, $SE = 0.22$, $t(271) = -3.54$, $p < .001$). Effects that were based on other types of stimuli were significantly more positive ($b = 0.56$, $SE = 0.26$, $t(271) = 2.25$, $p = .025$) than vignette-based effects. No further significant differences were found.

Finally, the results paralleled the previous moderator analysis regarding different types of evaluation. The label effect on attitudes toward the child ($b = 0.87$, $SE = 0.15$, $t(271) = 5.74$, $p < .001$), on evaluations of the cause of the student's problems ($b = 0.45$, $SE = 0.14$, $t(271) = 3.25$, $p = .001$), and on evaluations of treatment strategies ($b = 0.40$, $SE = 0.19$, $t(271) = 2.07$, $p = .039$) was significantly more positive than the effect on the reference category of academic evaluations. We found no differences between academic evaluations and each of the other types of evaluation.

Outlier Analyses

To detect possible influences of extraordinarily large effects, we estimated the main effect and the meta-regression model again, omitting all effects that differed more than three standard deviations from the mean (first outlier model) and all effects that differed more than two and a half standard deviations from the mean (second outlier model). The first approach led to an exclusion of two effects and a slightly less negative main effect of $g = -0.39$ ($k = 282$, $t(281) = -6.09$, $p < .001$, 95% CI [−0.51, −0.26]). The second approach led to an exclusion of ten effects and an even less negative main effect of $g = -0.34$ ($k = 274$, $t(273) = -7.08$, $p < .001$, 95% CI [−0.42, −0.24]). These results suggest that the main effect was somewhat influenced by exceptionally negative effects.

The first outlier meta-regression model (Table 5) yielded results very similar to the full meta-regression model. Publication year was still a significant moderator of the effect with only a slightly decreased positive influence ($b = 0.014$, $SE = 0.01$, $t(269) = 2.97$, $p = .003$). Effects that were based on the presentation of a singular label ($b = -0.54$, $SE = 0.22$, $t(269) = -2.40$, $p = .017$) were significantly more negative than vignette-based effects, while the presentation of other stimuli ($b = 0.54$, $SE = 0.24$, $t(269) = 2.29$, $p = .023$) yielded significantly more positive effects than the presentation of vignettes. Label effects on attitudes toward the child ($b = 0.85$, $SE = 0.15$, $t(269) = 5.68$, $p < .001$), on evaluations of the cause of the student's problems ($b = 0.45$, $SE = 0.16$, $t(269) = 3.31$, $p = .001$), and on the evaluations of treatment strategies ($b = 0.39$, $SE = 0.19$, $t(269) = 2.10$, $p = .037$) were significantly more positive than effects on academic evaluations. Apart from these findings, there were no significant differences in the first outlier model.

The positive influence of the publication year persisted in the second outlier meta-regression model ($b = 0.011$, $SE < 0.01$, $t(261) = 2.48$, $p = .014$; Table 5). While the presentation of a label only did not differ from the presentation of vignettes anymore, the presentation of other stimuli ($b = 0.46$, $SE = 0.21$, $t(261) = 2.18$, $p = .027$), and the presentation of a combination of stimuli ($b = 0.36$, $SE = 0.15$, $t(261)$

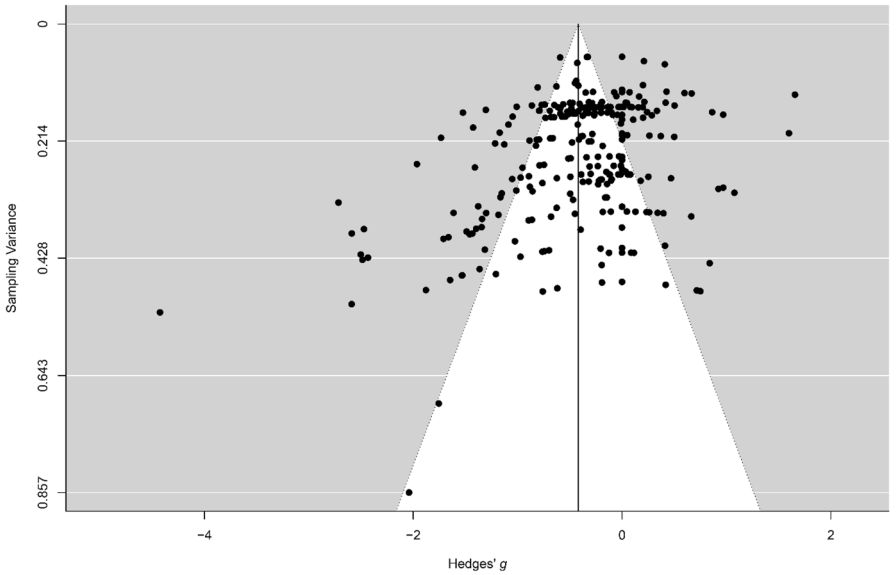

Fig. 3 Funnel plot of individual effect sizes ($k = 284$) against sampling variances

$= 2.37$, $p = .019$) yielded significantly more positive effects than the presentation of vignettes. Regarding the type of evaluation, there were more significant differences compared to the full model and the first outlier model. Effects on attitudes toward the child ($b = 0.77$, $SE = 0.13$, $t(261) = 5.89$, $p < .001$), on evaluations of the cause of the student's problems ($b = 0.44$, $SE = 0.12$, $t(261) = 3.72$, $p < .001$), on evaluations of treatment strategies ($b = 0.41$, $SE = 0.16$, $t(261) = 2.53$, $p = .012$), on overall assessments ($b = 0.30$, $SE = 0.13$, $t(261) = 2.28$, $p = .023$) and on other types of evaluation ($b = 0.25$, $SE = 0.13$, $t(261) = 1.96$, $p = .049$) were more positive than effects of academic evaluations. No other significant differences were found.

In summary, both outlier models suggest that the lack of evidence for a negative impact of labels on attitudes toward the child and on assessments of the cause of the student's problems in the singular moderator analysis cannot be attributed to the influence of outliers. Moreover, the fact that the positive influence of publication year was robust in both outlier models suggests that this trend cannot be attributed to highly negative effects reported by earlier studies. The finding that label effects caused by the presentation of labels only where not significantly different from effects caused by the presentations of vignettes in the second outlier model can be interpreted as evidence that the presentation of labels only yielded unusually negative effects.

Publication Bias

As an initial approach to investigate the possibility that the data was influenced by publication bias, we created a funnel plot showing all effect sizes on the x-axis

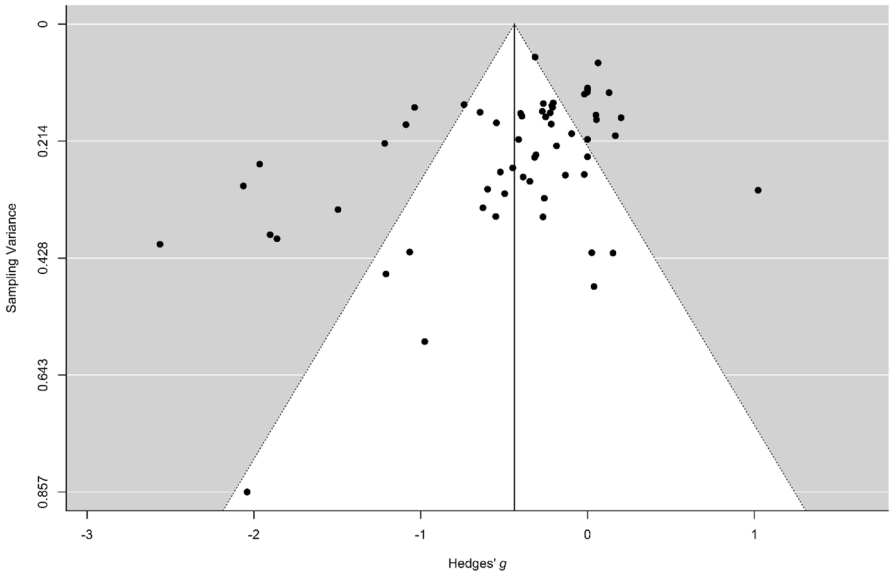

Fig. 4 Funnel plot of effect sizes averaged per experiment ($k = 60$) against sampling variances

against the corresponding sampling variances on the *y*-axis (Fig. 3). The plot suggested the presence of publication bias because it seemed to be asymmetric with more effect sizes on the left side.

We also averaged all effects of every experiment. Removing dependence in the data led to a slightly more negative overall effect of $g = -0.44$ ($k = 60$, $z(59) = 5.69$, $p < .001$, 95% CI [−0.59, -0.29]). We created another funnel plot showing the averaged effect sizes on the *x*-axis against the corresponding sampling variances on the *y*-axis (Fig. 4). A visual inspection of this plot was again indicative of publication bias because the distribution of effect sizes seemed to be asymmetric with more effect sizes on the left side. This impression was supported by a significant rank correlation test ($\tau = -0.29$, $p < .001$) (Begg & Mazumdar, 1994) and a significant regression test ($z = -3.40$, $p < .001$) (Egger et al., 1997). Given the presence of publication bias indicated by both tests, we applied the trim and fill method to estimate the number of studies that might be missing on the right side of the funnel plot because of biased publication (Duval & Tweedie, 2000). This method, however, suggested that no studies were missing on the right side ($SE = 4.06$). Thus, the estimated mean effect was left unaltered.

In summary, visual inspection of both funnel plots and the rank and regression test suggested the presence of publication bias, whereas the trim and fill method did not. Based on these inconsistent results, we can only conclude that there might be an influence of publication bias in our data. However, the reader must keep in mind that the deployed statistical tests are based on the assumption of independent effect sizes and, therefore, might not be suited for the data of this meta-analysis.

Discussion

This meta-analysis sought to answer the questions whether the existing studies yield evidence for an overall negative effect of diagnostic labels on the evaluation of children and whether the type of evaluation, the expertise of the person who evaluates the child, the amount of information, and the type of evaluation or study characteristics moderate this label effect. In response to the first question, our multilevel meta-analysis of experiments on label effects established a moderately negative average effect ($g = -0.42$). In response to the second question, our results show the negative label effect to be robust across several types of evaluations, different types of samples, and different diagnostic categories. We found some evidence that the effect depended on the amount and type of information presented to participants, with information-rich descriptions yielding smaller label effects. We found no indication that participant's expertise and the child's gender moderated the effect. Over the years, the negative label effect weakened. Finally, between- and within-subjects designs and experiments conducted inside and outside the U.S. were not found to differ. In the following, we discuss the theoretical and practical implications of the major findings in more detail.

Theoretical Implications

The main conclusion that can be drawn from the meta-analysis is that diagnostic labels can negatively affect how children are evaluated, across a broad range of diagnostic labels, types of evaluations, and possibly independent of the professional expertise of the evaluators. This conclusion complements the literature on mental health stigma discussed above. The research on mental health stigma suggests that people oftentimes hold stereotypes about people suffering from mental illnesses (Curcio & Corboy, 2019; Jorm et al., 2012; Rüsch et al., 2005) and that diagnostic labels can trigger these stereotypes (Berryessa & Wohlstetter, 2019; Carrizosa-Moog et al., 2019; Cuttler & Ryckman, 2019). The current analysis suggests that labeled children can face comparable stigmatization in the classroom. Teachers sometimes evaluate children struggling with certain challenges more negatively just because these children are diagnosed with a certain condition. This biased evaluation can be triggered by different types of diagnoses, such as learning disorders, behavioural disorders, and intellectual disabilities, and is independent of the gender of the child who is evaluated. Thus, rather than specific diagnostic labels, the simple fact that a child is diagnosed with a disorder can lead to more negative evaluations of the child.

Similarly, the null finding for a moderating role of expertise suggests that negative labelling effects might occur regardless of professional experience or training, implying that even professionals might contribute to the stigmatization of labeled children. This interpretation fits well with an explanation of label effects in terms of stereotypes. Stereotypes are often acquired via socialization (Degner & Dalege, 2013), are therefore deeply rooted in individuals' cognitive system and can exert

their effects through automatic mechanisms as priming (Kidder et al., 2018). Stereotypes triggered by diagnostic labels might exert their effects on evaluations in an automatic fashion rather than affecting controlled processes (see Devine, 1989, for this distinction), which would immunize label effects against modulating effects of professional knowledge at least to some extent. However, it must be noted that some of the occupational categories compared in the moderator analysis were quite small (for example, special education teachers with only 11 effects), which lowers the power of these comparisons, offering an alternative and purely methodological explanation for the null findings.

The meta-analysis also yielded evidence that diagnostic labels can affect several different types of evaluation negatively. Labeling a child can lead to worse academic evaluations (e.g., expecting poor performance in the future), behavioral evaluations (e.g., expecting the child to disrupt the classroom), evaluations of personality (e.g., attributing negative traits to the child), and to a more negative overall assessment (i.e., having a negative overall impression of the child). Somewhat unexpectedly, we found no support for the assumption that diagnostic labels affect the evaluation of treatment strategies (e.g., recommending mental health services). A possible explanation for this null result might be that participants ignore the label and focus on the child's problems when they need to consider whether a child can profit from treatment. Nevertheless, in light of the small amount of effect sizes of treatment evaluations ($k = 20$) and in light of the fact that the evaluation of treatment strategies was not a significant moderator in the meta-regression models, this interpretation is to be treated with caution.

Moreover, the analysis did not support the notion that diagnostic labels influence participants' evaluation of the child and their evaluation of the cause of the students' problems. The former result seems plausible because teachers are equally willing to work with or to help a child facing challenges regardless of whether the child is labeled or not. In contrast, the latter result is puzzling, especially since diagnostic labels had a negative influence on academic evaluations. If people evaluate labeled students' academic skills and their academic future negatively, their attributions should follow a corresponding pattern by being focused on internal and stable causes of students' difficulties. For example, if a teacher expects a student to perform poorly in the future because of the student's diagnosis, it seems likely that the teacher also attributes the student's failure to a permanent lack of skill or talent. Nevertheless, the effect of labels on evaluations of the causes of student's problems was not significant in the moderator analysis, and the effect was significantly more positive than the category of comparison (academic evaluations) in the meta-regression models. Consequently, we could not support this line of reasoning.

The meta-analysis yielded some limited evidence that the amount of information presented to participants moderated the label effect. In the singular moderator analysis and in the first meta-regression model, effects triggered by the sole mention of a label were considerably more negative than effects caused by the presentation of vignettes. If people simply know that a child has received a certain diagnosis, they evaluate the child in a manner consistent with stereotypical notions of intellectual, behavioral, and social problems that are associated with the diagnoses. Moreover, the fact that vignette-based effects were more positive

is evidence that the impact of labels becomes weaker when participants are given additional information, although labels still have a negative effect. However, as a caveat, we have to note that the outlier analyses based on the meta-regression model revealed that the markedly negative label effects in the label-only condition were primarily driven by some extraordinarily large effects. These effects might well be valid effects but, in principle, it is also possible that they are artefacts caused by (unknown) unrelated influences. Future research should clarify the robustness of label effects when no other information is provided.

The analysis of presenting information to participants in other ways revealed inconsistent results. In the singular moderator analysis, no difference was found between effects based on the presentation of vignettes, videos, other stimuli, and a combination of stimuli. However, in the first meta-regression model, effects based on the presentation of videos, a combination of stimuli, and other stimuli trended to be or were significantly more positive than vignette-based effects. This result is tentative evidence that more information about the child can result in a further weakening of label effects. If diagnostic labels affect teachers' expectations, this pattern of findings aligns well with the results from research on teacher expectations and self-fulfilling prophecies (e.g., Jussim & Harber, 2005; Raudenbush, 1984), the effects of which are predominantly occurring in lower grades when the teachers know less about the students and must rely on heuristics. However, since we did not examine self-fulfilling prophecies in this meta-analysis but merely label effects on the evaluation of students, this link to self-fulfilling prophecies remains speculative to some extent.

A critical alternative point of view poses that label effects might be a research artefact, which has been implied by several researchers (Cornett-Ruiz & Hendricks, 1993; Dukes & Saudargas, 1989; Fernald et al., 1985; Reschly & Lamprecht, 1979; Yoshida & Meyers, 1975). According to this line of reasoning, labels only have a negative impact when people lack sufficient information for a valid evaluation of the child. When people are presented just with a few sentences about a child, they will rely on the label as a source of information. However, when they are given more comprehensive information, for example, through the presentation of video material, the label has no effect on people's evaluations. Given that teachers are expected to have detailed information about their students from various sources, there should be no label effects beyond evaluations of artificial vignettes in the psychological laboratory.

We could not find unequivocal support for this line of reasoning. Thus, we deem this interpretation of the results an overstatement. Moreover, several cross-sectional and longitudinal studies have suggested that the negative label effect is a robust phenomenon that affects students in their daily lives (Eisenberg & Schneider, 2007; Knight, 2021; Schwehr et al., 2014; Shifrer et al., 2013; Shifrer, 2013, 2016; Whitley, 2010), and that is not modulated by knowledge about the underlying disorder. These studies complement the results of this meta-analysis by providing more ecologically valid evidence for negative label effects. For this reason, we are confident that negative label effects on the evaluation of children are not a mere research artifact caused by artificial stimuli.

Practical Implications

The disheartening practical implication of this meta-analysis is that diagnosed children suffer a dual burden. They struggle with considerable challenges that serve as the criteria for the diagnosis, and then they must further deal with diagnosis-related stigmatization. We think that the most important practical consequence that should follow from this is that parents, teachers, and mental health workers should be aware of negative label effects. Research is lacking on effective interventions for mitigating negative label effects (see Avenues for Future Research), but we believe that raising awareness of such effects can be an important first step. As long as practitioners lack the awareness of the potential negative influence of a diagnostic label, they will not be able to counteract it. Consequently, addressing negative label effects as a part of the training of teachers and mental health workers can increase the awareness and abate the adverse effects of labeling. On the positive side, we found evidence for a decline of the negative effect over the years, which might be due to a successful professionalization of people working in those fields, as well as general awareness-raising campaigns in society. Efforts of organizations and activists to overcome stigmatization of diagnosed people might have contributed to this weakening of label effects. However, it is also possible that the effect sizes of earlier studies might have been inflated because of publication bias, and more recent studies provide a more realistic estimation of the label effect. Because our efforts of detecting influences of publication bias yielded inconclusive results, we cannot rule out the second explanation.

We further assert that the stigmatization of diagnosed children has some relevance for the complex debate about categorical versus dimensional approaches to mental disorders (e.g., Coghill & Sonuga-Barke, 2012). The question whether mental disorders should be conceptualized as distinct causal entities or not certainly cannot be answered by our results. Nevertheless, the finding that diagnostic labels reliably have a negative impact on the evaluation of children can be interpreted as a downside of the currently widespread categorical approach to mental disorders. As soon as a child is categorized via a diagnosis, evaluations of the child are affected negatively by this categorization. For this reason, such negative consequences of categorical diagnoses are one of many aspects that have to be considered in the debate over categorical versus dimensional approaches to disorders.

Limitations

The results of this meta-analysis entail several limitations. First, several moderator analyses, such as the comparisons of US vs. not US, boys vs. girls, regular teachers vs. special education teachers, and mental health workers vs. other occupational groups, suffered from highly uneven quantities of effect sizes. We cannot rule out the possibility that differences exist between these categories in the population and that the failure to find evidence for these differences might have been due to insufficient power.

Second, in some cases, to assign labels unequivocally to only one diagnostic category was difficult for two reasons. First, in many studies, the authors gave no definition for the diagnosis used. Therefore, inferring the clinical condition from rather vague labels was difficult (e.g., developmentally delayed, socially maladjusted, minimal brain dysfunction). Second, some labels combined clinical aspects of different diagnostic categories (e.g., developmental delays and learning problems, educable mentally retarded). We addressed these problems by assigning labels to more nuanced subcategories to preserve the complexity of some clinical conditions. Moreover, we discussed the assignment of every label in detail, and when we could not agree on a specific assignment, the label was assigned to the residual category. Nevertheless, some label assignments in the meta-analysis are debatable. Finally, a point worth addressing is the fact that some of the labels used in studies are outdated (e.g., emotionally disturbed, educable mentally retarded). It is questionable how the results of these studies would be replicable using present day terminology.

Third, our coding of the amount of information was somewhat limited. We were only able to differentiate between different types of stimuli presented to participants (vignettes, videos, others, combination). Stimulus descriptions varied substantially in scope between studies (e.g., some studies reported the full wording of the vignettes, some studies described the videos at length, and other studies lacked this detailed information), which precluded detailed coding. Consequently, the inconsistent evidence for the moderating influence of the amount of information might be due to these limitations.

Fourth, we encountered several problems in coding studies and extracting effect sizes. A considerable number of studies could not be coded because the statistics were insufficient for calculating effect sizes. Beyond that, some studies reported inconsistencies in the result section. For example, in one study, the degrees of freedom were not transparent or changed without apparent reason in the course of data analysis (Bromfield et al., 1988). In another study, statistics were reported in a manner that was difficult to understand because unusual symbols were used (Tripp & Rizzo, 2006). In some other cases, calculating effect sizes based on means and standard deviations or based on F values yielded considerably different results (Aloia & MacMillan, 1983; Neisworth et al., 1974), or information about sample sizes and degrees of freedom did not match or were even contradictory (Taylor et al., 1983; Thelen et al., 2003). We discussed each problem at length with the goal of addressing problems consistently. For example, we always extracted the information about sample size that was reported in the text, even when information about degrees of freedom in a table suggested otherwise, and we always calculated effects based on means and standard deviations in cases in which a calculation based on F-values yielded different results. Nevertheless, we cannot rule out that our results are biased by these shortcomings.

Finally, a minor limitation is the fact that, in several cases, we had to calculate the statistics needed for effect size calculation indirectly, for example, by calculating weighted means of means or pooled standard deviations across subgroups (J. D. Fox & Stinnett, 1996; Rockett et al., 2007; Rolison & Medway, 1985; Severence & Gasstrom, 1977; Stinnett et al., 2001). To this end, we often based our calculations

on assumptions (e.g., the assumption that subgroups were even in size). This might have led to some inaccuracy in our data.

Avenues for Future Research

Four issues addressed in this meta-analysis require further research. First, we found some evidence that more information leads to weaker label effects, but this evidence was somewhat inconsistent. Consequently, researchers should further explore how different types of information presentation can influence label effects. Second, we found no evidence that label effects are weakened or nonexistent in experts. However, these null results are limited because of highly uneven numbers of effects, with small numbers of effects in some categories. Future research can focus on expertise and label effects to address this gap. In this context, the category of mental health workers deployed in this meta-analysis and in several studies is somewhat problematic because it combines people from different areas of expertise (e.g., psychologists, psychiatrists, nurses, etc.). Therefore, we recommend that future researchers concentrate on comparing clearly separated occupational groups (e.g., comparing teachers and child psychotherapists only). Third, it remains somewhat puzzling that we found no evidence for negative label effects on causal attributions regarding the child's problems. More research is needed on the influence of diagnostic labels on causal attributions. Fourth, there is a lack of research on the effects of specific labels. For example, 32 out of the 46 effects in the learning disability only category in this meta-analysis were based on experiments deploying the broad term "learning disability". Three of the remaining effects were based on the specific learning disability in the language area label, eight on the dyslexia label, and three on the dyscalculia label. Consequently, there is a need for investigations into the different label effects that might be associated with different learning disabilities.

Insufficient power is a ubiquitous problem in psychological research (e.g., Maxwell et al., 2015; T. D. Stanley et al., 2018) and the literature analyzed in this paper is not an exception. Most of the studies in this meta-analysis were based on a between-subjects design with a median sample size of 77 participants. Consequently, the average study in the analysis had only a power of 1-β = .57 (one-tailed testing) or 1-β = .44 (two-tailed testing) to find a significant difference of $g = -0.42$ between two independent means (at $\alpha = .05$). To overcome this limitation, future studies should recruit larger samples: at least 142 (one-tailed testing) or 180 (two-tailed testing) participants to achieve a power of 1-β = .80 and 248 (one-tailed testing) or 298 (two-tailed testing) participants to achieve a power of 1-β = .95.

Although this meta-analysis yielded consistent evidence for a negative effect of diagnostic labels, we do not think that labeling has only negative consequences. Some studies suggested that labels can also have a positive impact by increasing teacher's self-efficacy beliefs (Gibbs & Elliott, 2015) or their willingness to support the child (Jellison & Duke, 1994; Ohan et al., 2011) and by enabling parents to better understand their child's problems (Fernald & Gettys, 1980). We assume that labels can provide parents and teachers with closure. After struggling with the child's challenges for a long period of a time, the diagnosis finally "explains" why

the child faces these challenges. The diagnosis might also raise hopes that the child's condition can be treated effectively. Exploring the boundary conditions of positive and negative label effects should be a primary focus of future research.

Research on ways of counteracting label effects is comparably sparse. There is some very limited evidence that familiarizing participants with rating methods (Graham & Dwyer, 1987; Madle et al., 1980), arranging contact with diagnosed children (Herr, 1975; Herr et al., 1976), or educating people about disorders (Kutcher et al., 2016; Ohan et al., 2008; Parish et al., 1977; Toye et al., 2019) can mitigate negative label effects to some extent. Nevertheless, more research is needed to develop effective interventions for counteracting negative label effects. Such efforts could highly benefit from intensified research on positive label effects.

Conclusion

Our research demonstrates the potentially adverse effects of diagnostic labels on student assessments. Although we acknowledge the necessity of diagnostic labels in the therapeutic and medical domain, we advise to communicate labels only accompanied with thorough explanations to teachers and parents. We recommend raising awareness in society about behavioral and learning disorders, as well as about intellectual disabilities. The decrease in negativity of the label effect over the years provides hope that the ongoing efforts of education on mental conditions have at least been partially effective.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-023-09716-6.

Acknowledgements We would like to thank Janna Teigeler and Cassandra Rosenbaum for their valuable help in study coding.

Funding Information Open Access funding enabled and organized by Projekt DEAL. We received funding for this research from the Human Dynamics Center of the Faculty of Human Sciences at the University of Würzburg.

Data Availability The data and R-Code underlying all analyses are available at the repository of the Open Science Framework (https://osf.io/g72nt/?view_only=d80b63e8c4084bd28629ed9a81414df8).

Declarations

Conflict of Interest The authors declare no competing interests.

References

References marked with an asterisk indicate studies included in the meta-analysis

*Allday, R. A., Duhon, G. J., Blackburn-Ellis, S., & van Dycke, J. L. (2011). The biasing effects of labels on direct observation by preservice teachers. Teach Educ Spec Educ, 34(1), 52–58. https://doi.org/10.1177/0888406410380422

Allen, M. S., Robson, D. A., Martin, L. J., & Laborde, S. (2020). Systematic review and meta-analysis of self-serving attribution biases in the competitive context of organized sport. *Pers Soc Psychol Bull, 46*(7), 1027–1043. https://doi.org/10.1177/0146167219893995

*Aloia, G. F. (1975). Effects of physical stigmata and labels on judgments of subnormality by preservice teachers. Ment. Retard, 13(6), 17–21.

*Aloia, G. F., & MacMillan, D. L. (1983). Influence of the EMR label on initial expectations of regular-classroom teachers. Am. J. Ment. Defic., 88(3), 255–262.

*Aloia, G. F., Maxwell, J. A., & Aloia, S. D. (1981). Influence of a child's race and the EMR label on initial impressions of regular-classroom teachers. Am. J. Ment. Defic., 85(6), 619-623.

Angermeyer, M. C., & Matschinger, H. (2005). Labeling - stereotype - discrimination. An investigation of the stigma process. *Soc Psychiatry Psychiatr Epidemiol, 40*(5), 391–395. https://doi.org/10.1007/s00127-005-0903-4

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *Quant. Meth. Psych, 12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

*Batzle, C. S., Weyandt, L. L., Janusis, G. M., & DeVietti, T. L. (2010). Potential impact of ADHD with stimulant medication label on teacher expectations. J. Atten. Disord., 14(2), 157–166. https://doi.org/10.1177/1087054709347178

*Baudson, T. G., & Preckel, F. (2013). Teachers' implicit personality theories about the gifted: An experimental approach. Sch Psychol Q, 28(1), 37–46. https://doi.org/10.1037/spq0000011

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088–1101. https://doi.org/10.2307/2533446

Berryessa, C. M., & Wohlstetter, B. (2019). The psychopathic "label" and effects on punishment outcomes: A meta-analysis. *Law Hum Behav, 43*(1), 9–25. https://doi.org/10.1037/lhb0000317

*Bianco, M. (2005). The effects of disability labels on special education and general education teachers' referrals for gifted programs. Learning Disability Quarte, 28(4), 285–293. https://doi.org/10.2307/4126967

*Bianco, M., & Leech, N. L. (2010). Twice-exceptional learners: Effects of teacher preparation and disability labels on gifted referrals. Teach Educ Spec Educ, 33(4), 319–334. https://doi.org/10.1177/0888406409356392

Borenstein, M., Hedges, L. V., Julian, H. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bosnjak, M., & Viechtbauer, W. (2009). Die Methode der Meta-Analyse zur Evidenzbasierung von Gesundheitsrisiken: Beiträge der Sozial-, Verhaltens- und Wirtschaftswissenschaften. [The meta-analytic method for establishing the evidence base of health risks: Contributions from the social, behavioral, and economic sciences.]. *Zentralblatt fur Arbeitsmedizin Arbeitsschutz Ergon., 59*(11), 322–333. https://doi.org/10.1007/BF03344247

Boucher, C. R., & Deno, S. L. (1979). Learning disabled and emotionally disturbed: Will the labels affect teacher planning? *Psychol Sch., 16*(3), 395–402. https://doi.org/10.1002/1520-6807(197907)16:3<395::AID-PITS2310160316>3.0.CO;2-6

*Bromfield, R., Bromfield, D., & Weiss, B. (1988). Influence of the sexually abused label on perceptions of a child's failure. J. Educ. Res., 82(2), 96–98

Carrizosa-Moog, J., Salazar-Velasquez, L. V., Portillo-Benjumea, M., Rodriguez-Mejia, A., & Isaza-Jaramillo, S. (2019). Does public attitude change by labeling a person as epileptic, person with epilepsy or the acronym PWE? A systematic review. Seizure: European. *J. Epilepsy Res., 69*, 273–278. https://doi.org/10.1016/j.seizure.2019.05.011

Carroll, C. F., & Reppucci, N. D. (1978). Meanings that professionals attach to labels for children. *J Consult Clin Psychol, 46*(2), 372–374. https://doi.org/10.1037/0022-006X.46.2.372

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychol. Methods, 19*(2), 211–229. https://doi.org/10.1037/a0032968

Cheung, M. W.-L. (2019). A Guide to Conducting a Meta-Analysis with Non-Independent Effect Sizes. *Neuropsychol. Rev., 29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129.

Coghill, D., & Sonuga-Barke, E. J. S. (2012). Annual research review: Categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders – implications of recent empirical study. *J. Child Psychol. Psychiatry, 53*(5), 469–489. https://doi.org/10.1111/j.1469-7610.2011.02511.x

*Combs, R. H., & Harper, J. L. (1967). Effects of labels on attitudes of educators toward handicapped children. Except. Child., 33(6), 399–403. https://doi.org/10.1177/001440296703300607
*Cornett-Ruiz, S., & Hendricks, B. (1993). Effects of labeling and ADHD behaviors on peer and teacher judgments. J. Educ. Res., 86(6), 349–355. https://doi.org/10.1080/00220671.1993.9941228
Corrigan, P. W. (2007). How clinical diagnosis might exacerbate the stigma of mental illness. *Soc Work, 52*(1), 31–39. https://doi.org/10.1093/sw/52.1.31
Cuffe, S. P., Moore, C. G., & McKeown, R. E. (2005). Prevalence and correlates of ADHD symptoms in the national health interview survey. *J. Atten. Disord., 9*(2), 392–401. https://doi.org/10.1080/21622965.2018.1430576
Curcio, C., & Corboy, D. (2019). Stigma and anxiety disorders: A systematic review. *Stig and Health, 5*(2), 125–137. https://doi.org/10.1037/sah0000183
Cuttler, C., & Ryckman, M. (2019). Don't call me delusional: Stigmatizing effects of noun labels on people with mental disorders. *Stig and Health, 4*(2), 118–125. https://doi.org/10.1037/sah0000132
Degner, J., & Dalege, J. (2013). The apple does not fall far from the tree, or does it? A meta-analysis of parent–child similarity in intergroup attitudes. *Psychol. Bull., 139*(6), 1270–1304. https://doi.org/10.1037/a0031436
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *J Pers Soc Psychol, 56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5
Ditchman, N., Werner, S., Kosyluk, K., Jones, N., Elg, B., & Corrigan, P. W. (2013). Stigma and intellectual disability: Potential application of mental illness research. *Rehabil. Psychol., 58*(2), 206–216. https://doi.org/10.1037/a0032466
*Duke, J., & Prater, G. (1991). Preschool teachers' expectations of preschoolers labeled developmentally delayed: A pilot study. (ERIC Document Reproduction Service No. ED355047)
*Dukes, M., & Saudargas, R. A. (1989). Teacher evaluation bias toward LD children: Attenuating effects of the classroom ecology. Learn Disabil Q, 12(2), 126–132. https://doi.org/10.2307/1510728
Duval, S. J., & Tweedie, R. L. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x
Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629
Eisenberg, D., & Schneider, H. (2007). Perceptions of academic skills of children diagnosed with ADHD. *J. Atten. Disord., 10*(4), 390–397. https://doi.org/10.1177/1087054706292105
*Fernald, C. D., & Gettys, L. (1980). Diagnostic labels and perceptions of children's behavior. J. Clin. Child Psychol., 9(3), 229-233. https://doi.org/10.1080/15374418009532996
*Fernald, C. D., Williams, R. A., & Droescher, S. D. (1985). Actions speak louder…: Effects of diagnostic labels and child behavior on perceptions of children. Prof Psychol Res Pr, 16(5), 648–660. https://doi.org/10.1037/0735-7028.16.5.648
Fogel, L. S., & Nelson, R. O. (1983). The effects of special education labels on teachers' behavioral observations, checklist scores, and grading of academic work. *J. Sch. Psychol., 21*(3), 241–251. https://doi.org/10.1016/0022-4405(83)90019-5
Foroni, F., & Rothbart, M. (2011). Category boundaries and category labels: When does a category name influence the perceived similarity of category members? *Soc Cog, 29*(5), 547–576.
Foroni, F., & Rothbart, M. (2013). Abandoning a label doesn't make it disappear: The perseverance of labeling effects. *J. Exp. Soc. Psychol., 49*(1), 126–131. https://doi.org/10.1016/j.jesp.2012.08.002
*Foster, G., Algozzine, B., & Ysseldyke, J. (1980). Classroom teacher and teacher-in-training susceptibility to stereotypical bias. Pers Guid J., 59(1), 27–30. https://doi.org/10.1002/j.2164-4918.1980.tb00478.x
*Foster, G., & Keech, V. (1977). Teacher reactions to the label of educable mentally retarded. Educ. train. ment. retard, 12(4), 307–311.
Foster, G., & Ysseldyke, J. (1976). Expectancy and halo effects as a result of artificially induced teacher bias. *Contemp. Educ. Psychol., 1*(1), 37–45. https://doi.org/10.1016/0361-476X(76)90005-9
*Foster, G., Ysseldyke, J. E., & Reese, J. H. (1975). "I wouldn't have seen it if I hadn't believed it". Except. Child., 41(7), 469–473. https://doi.org/10.1177/001440297504100701
*Foster, G., Schmidt, C. R., & Sabatino, D. (1976). Teacher expectancies and the label "learning disabilities". J. Learn. Disabil., 9(2), 111–114. https://doi.org/10.1177/002221947600900209
Fox, A. B., Earnshaw, V. A., Taverna, E. C., & Vogt, D. (2018). Conceptualizing and measuring mental illness stigma: The mental illness stigma framework and critical review of measures. *Stig and Health, 3*(4), 348–376. https://doi.org/10.1037/sah0000104

*Fox, J. D., & Stinnett, T. A. (1996). The effects of labeling bias on prognostic outlook for children as a function of diagnostic label and profession. Psychol Sch., 33(2), 143–152. https://doi.org/10.1002/(SICI)1520-6807(199604)33:2<143::AID-PITS7>3.0.CO;2-S

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychon Bull Rev, 19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

*Franz, D. J., Lenhard, W., Marx, P., & Richter, T. (2021). Here I sit, making men in my own image: How learning disorder labels affect teacher student's expectancies. Curr Psychol. https://doi.org/10.1007/s12144-021-02250-0

Fresson, M., Meulemans, T., Dardenne, B., & Geurten, M. (2019). Overdiagnosis of ADHD in boys: Stereotype impact on neuropsychological assessment. *Appl. Neuropsychol. Child, 8*(3), 231–245. https://doi.org/10.1080/21622965.2018.1430576

*Gibbs, S., & Elliott, J. (2015). The differential effects of labelling: How do 'dyslexia' and 'reading difficulties' affect teachers' beliefs. Eur. J. Spec. Needs Educ., 30(3), 323-337. https://doi.org/10.1080/08856257.2015.1022999

Gillung, T. B., & Rucker, C. N. (1977). Labels and teacher expectations. *Except. Child., 43*(7), 464–465. https://doi.org/10.1177/001440297704300712

*Graham, S., & Dwyer, A. (1987). Effects of the learning disability label, quality of writing performance, and examiner's level of expertise on the evaluation of written products. J. Learn. Disabil., 20(5), 317-318. https://doi.org/10.1177/002221948702000513

*Graham, S., & Leone, P. (1987). Effects of behavioral disability labels, writing performance, and examiner's expertise on the evaluation of written products. J Exp Educ, 55(2), 89–94. https://doi.org/10.1080/00220973.1987.10806439

*Guskin, S. L. (1962a). The perception of subnormality in mentally defective children. Am. J. Ment. Defic, 67(1), 53-60.

*Guskin, S. (1962b). The influence of labelling upon the perception of subnormality in mentally defective children. Am. J. Ment. Defic., 67(3), 402-406.

*Guskin, S. L. (1963). Measuring the strength of the stereotype of the mental defective. Am. J. Ment. Defic., 67(4), 569-575.

Herr, D. E. (1975). Camp counseling with emotionally disturbed adolescents. *Except. Child., 41*(5), 331–332. https://doi.org/10.1177/001440297504100504

Herr, D. E., Algozzine, B., & Eaves, R. C. (1976). Modification of biases held by teacher trainees toward the disturbingness of child behaviors. *J. Educ. Res., 69*(7), 261–264. https://doi.org/10.1080/00220671.1976.10884893

Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *J. Clin. Epidemiol., 58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

*Jacobs, W. R. (1978). The effect of the learning disability label on classroom teachers' ability objectively to observe and interpret child behaviors. Learn Disabil Q, 1(1), 50–55. https://doi.org/10.2307/1510963

*Javel, M. E., & Greenspan, S. (1983). Influence of personal competence profiles on mainstreaming recommendations of school psychologists. Psychol Sch, 20(4), 495-465. https://doi.org/10.1002/1520-6807(198310)20:4

Jellison, J. A., & Duke, R. A. (1994). The Mental Retardation Label: Music Teachers' and Prospective Teachers' Expectations for Children's Social and Music Behaviors. *J. Music Ther., 31*(3), 166–185. https://doi.org/10.1093/jmt/31.3.166

Johnson, L. J., & Blankenship, C. S. (1984). A comparison of label-induced expectancy bias in two preservice teacher education programs. *Behav. Disord., 9*(3), 167–174. https://doi.org/10.1177/019874298400900305

*Jones, S., & Cauffman, E. (2008). Juvenile psychopathy and judicial decision making: An empirical analysis of an ethical dilemma. Behav Sci Law, 26(2), 151-165. https://doi.org/10.1002/bsl.792

Jorm, A. F., Reavley, N. J., & Ross, A. M. (2012). Belief in the dangerousness of people with mental disorders: A review. *Aust N Z J Psychiatry, 46*(11), 1029–1045. https://doi.org/10.1177/0004867412442406

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Pers. Soc. Psychol. Rev., 9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3

Jussim, L., Madon, S., & Chatman, C. (1994). Teacher expectations and student achievement. In L. Heath, R. S. Tindale, J. Edwards, E. J. Posavac, F. B. Bryant, E. Henderson-King, . . . J. Myers

(Eds.), Social psychological applications to social issues. Applications of heuristics and biases to social issues (Vol. 3, pp. 303–334). : Springer US. https://doi.org/10.1007/978-1-4757-9238-6_16

Jussim, L., Palumbo, P., Chatman, C., Madon, S., & Smith, A. (2000). Stigma and self-fulfilling prophesies. In T. F. Heatherton, R. E. Kleck, M. R. Hebl, & J. G. Hull (Eds.), *The social psychology of stigma* (pp. 374–418). Guilford Press.

*Kedar-Voivodas, G., & Tannenbaum, A. J. (1979). Teachers' attitudes toward young deviant children. J. Educ. Psychol., 71(6), 800–808. https://doi.org/10.1037/0022-0663.71.6.800

*Kesterson, J. S. (2013). The effects of labeling and teacher knowledge of autism on attributions made about students with autism spectrum disorders. (Order No. AAI3525623). Available from APA PsycInfo®. (1428018235; 2013-99131-029). https://search.proquest.com/docview/1428018235?accountid=15156.

Kidder, C. K., White, K. R., Hinojos, M. R., Sandoval, M., & Crites Jr., S. L. (2018). Sequential stereotype priming: A meta-analysis. *Pers. Soc. Psychol. Rev., 22*(3), 199–227. https://doi.org/10.1177/1088868317723532

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Stat Med, 22*, 2693–2710. https://doi.org/10.1002/sim.1482

Knight, C. (2021). The impact of the dyslexia label on academic outlook and aspirations: An analysis using propensity score matching. *Br J Educ Psychol, 17*. https://doi.org/10.1111/bjep.12408

Kutcher, S., Wei, Y., Gilberds, H., Ubuguyu, O., Njau, T., Brown, A., Sabuni, N., Magimba, A., & Perkins, K. (2016). A school mental health literacy curriculum resource training approach: Effects on Tanzanian teachers' mental health knowledge, stigma and help-seeking efficacy. *Int. J. Ment. Health Syst., 10*(50), 1–9. https://doi.org/10.1186/s13033-016-0082-6

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Front. Psychol., 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

*Lee, K. W., Cheung, R. Y. M., & Chen, M. (2019). Preservice teachers' self-efficacy in managing students with symptoms of attention deficit/hyperactivity disorder: The roles of diagnostic label and students' gender. Psychol Sch, 56(4), 595–607. https://doi.org/10.1002/pits.22221

Lenhard, W., Breitenbach, E., Ebert, H., Schindelhauer-Deutscher, J., & Henn, W. (2005). Psychological benefit of diagnostic certainty for mothers of children with disabilities: Lessons from Down syndrome. *Am. J. Med. Genet., 132A*(2), 170–175.

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *J Pers Soc Psychol, 74*(6), 1421–1436. https://doi.org/10.1037/0022-3514.74.6.1421

Link, B. G., Cullen, F. T., Struening, E., Shrout, P. E., & Dohrenwend, B. P. (1989). A modified labeling theory approach to mental disorders: An empirical assessment. *Am. Sociol. Rev., 54*(3), 400. https://doi.org/10.2307/2095613

MacMillan, D. L., Jones, R. L., & Aloia, G. F. (1974). The mentally retarded label: A theoretical analysis and review of research. *Am. J. Ment. Defic., 79*(3), 241–261.

*Madle, R. A., Neisworth, J. T., & Kurtz, P. D. (1980). Biasing of hyperkinetic behavior ratings by diagnostic reports: Effects of observer training and assessment method. J. Learn. Disabil., 13(1), 30-33. https://doi.org/10.1177/002221948001300108

Madon, S., Willard, J., Guyll, M., & Scherr, K. C. (2011). Self-fulfilling prophecies: Mechanisms, power, and links to social problems. *Soc. Personal. Psychol. Compass, 5*(8), 578–590. https://doi.org/10.1111/j.1751-9004.2011.00375.x

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am Psychol, 70*(6), 487–498. https://doi.org/10.1037/a0039400

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychol. Bull., 130*(5), 711–747. https://doi.org/10.1037/0033-2909.130.5.711

Minner, S. (1982). Expectations of vocational teachers for handicapped students. *Except. Child., 48*(5), 451–453. https://doi.org/10.1177/001440298204800509

*Minner, S. (1989). Initial referral recommendations of teachers toward gifted students with behavioral problems. Roeper Review: A J on Gifted Education, 12(2), 78-80. https://doi.org/10.1080/02783198909553240

*Minner, S. (1990). Teacher evaluations of case descriptions of LD gifted children. Gift Child Q, 34(1), 37-39. https://doi.org/10.1177/001698629003400108

Minner, S., & Prater, G. (1984). College teachers' expectations of LD students. *Acad Ther, 20*(2), 225–229. https://doi.org/10.1177/105345128402000213

*Minner, S., Prater, G., Bloodworth, H., & Walker, S. (1987). Referral and placement recommendations of teachers toward gifted handicapped children. Roeper Rev, 9(4), 247–249. https://doi.org/10.1080/02783198709553064

*Moberg, S. (1995). Impact of teachers' dogmatism and pessimistic stereotype on the effect of EMR-class label on teachers' judgments in Finland. Educ Train Autism Dev Disabil, 30(2), 141–150.

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *Int. J. Soc. Res. Methodol., 20*(6), 559–572. https://doi.org/10.1080/13645579.2016.1252189

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods, 7*(1), 105–125. https://doi.org/10.1037//1082-989X.7.1.105

*Murrie, D. C., Cornell, D. G., & McCoy, W. K. (2005). Psychopathy, conduct disorder, and stigma: Does diagnostic labeling influence juvenile probation officer recommendations? Law Hum Behav, 29(3), 323–342. https://doi.org/10.1007/s10979-005-2415-x

*Neisworth, J. T., Kurtz, P. D., Jones, R. T., & Madle, R. A. (1974). Biasing of hyperkinetic behavior ratings by diagnostic reports. J. Abnorm. Child Psychol., 2(4), 323–329.

*O'Donohue, W., & O'Hare, E. (1997). How do teachers react to children labeled as sexually abused? Child Maltreatment, 2(1), 46–51. https://doi.org/10.1177/1077559597002001005

Ohan, J. L., Cormier, N., Hepp, S. L., Visser, T. A. W., & Strain, M. C. (2008). Does knowledge about attention-deficit/hyperactivity disorder impact teachers' reported behaviors and perceptions? *Sch Psychol Q, 23*(3), 436–449. https://doi.org/10.1037/1045-3830.23.3.436

*Ohan, J.L., Visser, T.A.W., Strain, M.C., & Allen, L. (2011). Teachers' and education students' perceptions of and reactions to children with and without the diagnostic label "ADHD". J. Sch. Psychol., 49(1), 81–105. https://doi.org/10.1016/j.jsp.2010.10.001

Osterholm, K., Nash, W. R., & Kritsonis, W. A. (2011). Effects of labeling students "learning disabled": Emergent themes in the research literature 1970 through 2000. FOCUS on Colleges, Universities & Schools, 6(1), 1–11.

*Parish, T. S., Dyck, N., & Kappes, B. M. (1979). Stereotypes concerning normal and handicapped children. J. Psychol., 102(1), 63–70. https://doi.org/10.1080/00223980.1979.9915095

Parish, T. S., Eads, G. M., Reece, N. H., & Piscitello, M. A. (1977). Assessment and attempted modification of future teachers' attitudes toward handicapped children. *Percept Mot Skills, 44*(2), 540–542. https://doi.org/10.2466/pms.1977.44.2.540

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *Eur. J. Soc. Psychol., 38*(6), 922–934. https://doi.org/10.1002/ejsp.504

Pfeiffer, S. I. (1980). The influence of diagnostic labeling on special education placement decisions. *Psychol Sch., 17*(3), 346–350. https://doi.org/10.1002/1520-6807(198007)17:3<346::AID-PITS2310170311>3.0.CO;2-N

Publication manual of the American psychological association. (2020). *Bias-free language* (7th ed. ed.). American Psychological Association. https://doi.org/10.1037/0000165-000

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *J. Educ. Psychol., 76*(1), 85–97. https://doi.org/10.1037/0022-0663.76.1.85

Rees, H. R., Ma, D. S., & Sherman, J. W. (2020). Examining the relationships among categorization, stereotype activation, and stereotype application. *Pers Soc Psychol Bull, 46*(4), 499–513. https://doi.org/10.1177/0146167219861431

Reschly, D. J., & Lamprecht, M. J. (1979). Expectancy effects of labels: Fact or artifact? *Except. Child., 46*(1), 55–58. https://doi.org/10.1177/001440297904600110

*Rockett, J. L., Murrie, D. C., & Boccaccini, M. T. (2007). Diagnostic labeling in juvenile justice settings: Do psychopathy and conduct disorder findings influence clinicians? Psychol. Serv., 4(2), 107–122. https://doi.org/10.1037/1541-1559.4.2.107

*Rolison, M. A., & Medway, F. J. (1985). Teachers' expectations and attributions for student achievement: Effects of label, performance pattern, and special education intervention. Am. Educ. Res. J., 22(4), 561–573. https://doi.org/10.3102/00028312022004561

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull., 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart & Winston.

Rubie-Davies, C. (2009). Teacher expectations and labeling. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (pp. 695–707). Springer.

Rüsch, N., Angermeyer, M. C., & Corrigan, P. W. (2005). Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *Eur J Psychiatry, 20*(8), 529–539. https://doi.org/10.1016/j.eurpsy.2005.04.004

Salvia, J., Clark, G. M., & Ysseldyke, J. E. (1973). Teacher retention of stereotypes of exceptionality. *Except. Child., 39*(8), 651–652. https://doi.org/10.1177/001440297303900807

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Rev. Educ. Res., 84*(3), 328–364. https://doi.org/10.3102/0034654313500826

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Front. Psychol., 10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Schwehr, E., Bocanegra, J. O., Kwon, K., & Sheridan, S. M. (2014). Impact of children's identified disability status on parent and teacher behavior ratings. *Contemp. Sch. Psychol., 18*(2), 133–142. https://doi.org/10.1007/s40688-014-0014-x

*Severence, L. J., & Gasstrom, L. L. (1977). Effects of the label "mentally retarded" on causal explanations for success and failure outcomes. Am. J. Ment. Defic., 81(6), 547–555.

Shifrer, D. (2013). Stigma of a label: Educational expectations for high school students labeled with learning disabilities. *J Health Soc Behav, 54*(4), 462–480. https://doi.org/10.1177/0022146513503346

Shifrer, D. (2016). Stigma and stratification limiting the math course progression of adolescents labeled with a learning disability. *Learn Instr, 42*, 47–57. https://doi.org/10.1016/j.learninstruc.2015.12.001

Shifrer, D., Callahan, R. M., & Muller, C. (2013). Equity or marginalization? The high school course-taking of students labeled with a learning disability. *Am. Educ. Res. J., 50*(4), 656–682. https://doi.org/10.3102/0002831213479439

*Shuller, D. Y., & McNamara, R. J. (1976). Expectancy factors in behavioral observation. Behav. Ther., 7(4), 519–527. https://doi.org/10.1016/S0005-7894(76)80172-4

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci., 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

*Stanley, M. A., & Comer, R. J. (1988). Reacting to mentally retarded persons: The impact of labels and observed behaviors. J Soc Clin Psychol, 6(3-4), 279–292. https://doi.org/10.1521/jscp.1988.6.3-4.279

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull., 144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stinnett, T. A., Bull, K. S., Koonce, D. A., & Aldridge, J. O. (1999). Effects of diagnostic label, race, gender, educational placement, and definitional information on prognostic outlook for children with behavior problems. *Psychol Sch., 36*(1), 51–59. https://doi.org/10.1002/(SICI)1520-6807(199901)36:1<51::AID-PITS6>3.0.CO;2-3

*Stinnett, T. A., Crawford, S. A., Gillespie, M. D., Cruce, M. K., & Langford, C. A. (2001). Factors affecting treatment acceptability for psychostimulant medication versus psychoeducational intervention. Psychol Sch, 38(6), 585–591. https://doi.org/10.1002/pits.1045

*Sutherland, J., & Algozzine, B. (1979). The learning disabled label as a biasing factor in the visual motor performance of normal children. J. Learn. Disabil., 12(1), 8–14. https://doi.org/10.1177/002221947901200103

*Taylor, R. L., Smiley, L. R., & Ziegler, E. W. (1983). The effects of labels and assigned attributes on teacher perceptions of academic and social behavior. Educ. train. ment. retard, 18(1), 45–51.

*Thelen, R. L., Burns, M. K., & Christiansen, N. D. (2003). Effects of high-incidence disability labels on the expectations of teachers, peers, and college students. Ethical Human Sciences & Services, 5(3), 183–193. https://search.proquest.com/docview/620360579?accountid=15156

Thurman, S. K., Brobeil, R. A., DuCette, J. P., & Hurt, H. (1994). Prenatally exposed to cocaine. *J. Early Interv., 18*(2), 119–130. https://doi.org/10.1177/105381519401800201

*Tournaki, N. (2003). Effect of student characteristics on teachers' predictions of student success. J. Educ. Res., 96(5), 310-319. https://doi.org/10.1080/00220670309597643

Toye, M. K., Wilson, C., & Wardle, G. A. (2019). Education professionals' attitudes towards the inclusion of children with ADHD: The role of knowledge and stigma. *J. Res. Spec. Educ. Needs, 19*(3), 184–196. https://doi.org/10.1111/1471-3802.12441

*Tripp, A., & Rizzo, T. L. (2006). Disability labels affect physical educators. Adapt Phys Activ Q, 23(3), 310–326. https://doi.org/10.1123/apaq.23.3.310

van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behav. Res. Methods, 45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educ Psychol Meas, 63*(5), 765–790. https://doi.org/10.1177/0013164402251027

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw., 36*(3). https://doi.org/10.18637/jss.v036.i03

Vlachou, A., Eleftheriadou, D., & Metallidou, P. (2014). Do learning difficulties differentiate elementary teachers' attributional patterns for students' academic failure? A comparison between Greek regular and special education teachers. *Eur. J. Spec. Needs Educ, 29*(1), 1–15. https://doi.org/10.1080/08856257.2013.830440

Wang, S., Rubie-Davies, C., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educ. Res. Eval., 24*(3-5), 124–179. https://doi.org/10.1080/13803611.2018.1548798

*Weisz, J. R. (1981). Effects of the "mentally retarded" label on adult judgments about child failure. J. Abnorm. Psychol., 90(4), 371–374. https://doi.org/10.1037/0021-843X.90.4.371

Whitley, J. (2010). Modelling the influence of teacher characteristics on student achievement for Canadian students with and without learning disabilities. *Int. J. Spec. Educ., 25*(3), 88–97.

*Yoshida, R. K., & Meyers, C. E. (1975). Effects of labeling as educable mentally retarded on teachers' expectancies for change in a student's performance. J. Educ. Psychol., 67(4), 521–527. https://doi.org/10.1037/h0077020