A Meta-Analysis on the Differences in Mathematical and Cognitive Skills Between Individuals With and Without Mathematical Learning Disabilities

Evelyn H. Kroesbergen
Marije D. E. Huijsmans
Radboud University

Ilona Friso-van den Bos
University of Twente

Types of mathematical learning disability (MLD) are very heterogeneous. Lower scores on mathematics and several cognitive skills have been revealed in samples with MLD compared with those with typical development (TD), but these studies vary in sample selection, making it difficult to generalize conclusions. Furthermore, many studies have investigated only one or few cognitive skills, making it difficult to compare their relative discrepancies. The current meta-analysis ($k = 145$) was conducted to (a) give a state-of-the-art overview of the mathematical and cognitive skills associated with MLD and (b) investigate how selection criteria influence conclusions regarding this topic. Results indicated that people with MLD display lower scores not only on mathematics but also on number sense, working memory, and rapid automatized naming compared with those with TD, in general independently of the criteria used to define MLD. A profile that distinguishes people with more serious, persistent, or specific MLD from those with less severe MLD was not detected.

**Keywords:** mathematics education, cognition, individual differences, effect size, meta-analysis, developmental dyscalculia, mathematical learning disability, number sense, rapid automatized naming, working memory

A substantial number of people experience learning difficulties in mathematics. If the difficulties are serious and persistent, such people might be diagnosed
with a mathematical learning disability (MLD). Prevalence estimates of MLD range from 3% to 8% internationally (Butterworth, 2005; Desoete et al., 2004; Geary, 2004; Rubinsten & Henik, 2009). Because of its prevalence and persistence, it is important to gather evidence regarding the (cognitive) characteristics of children and adults with MLD, so that education and interventions can be tailored to their needs. However, empirical research on MLD often includes a broader group of people with mathematical difficulties, such as math performance below a certain criterion (e.g., lowest 10%), regardless of the persistency of their difficulties (Geary et al., 2007; Murphy et al., 2007). The fact that different cutoff criteria are used (ranging from percentile 2 to 40 in the studies analyzed for the current meta-analysis) makes it even more difficult to compare different studies and to answer the question what exactly are the characteristics of MLD, especially on the cognitive level. As a consequence, a multitude of cognitive characteristics for MLD have been described in literature, which makes constructing a clear description of the behavioral and cognitive profile of people with MLD very difficult if not impossible. It is needless to say that this heterogeneity impedes diagnostic practices as well.

Some researchers have, for this reason, argued that a distinction should be made between different types of MLD, based on either behavioral or related cognitive characteristics (e.g., Mazzocco et al., 2011; Piazza et al., 2010). Following such a viewpoint, a specific cognitive profile has indeed been described and related to developmental dyscalculia, which is defined as atypical numerical development despite normal intelligence and educational opportunities. Developmental dyscalculia involves difficulties with very basic numerical processes (Simmons & Singleton, 2008) and is thought to result from problems in the conceptual processing of numbers and numerosities (Willburger et al., 2008). The math difficulties related to developmental dyscalculia are claimed to result from difficulties with number sense (cf. Piazza et al., 2010; Von Aster & Shalev, 2007), which is a domain-specific skill to mathematics.

Number sense is here defined as the capacity to recognize and understand symbolic numbers and nonsymbolic numerosities as well as the ability of mapping the quantities they represent (Dehaene et al., 2003; Geary, 2011a). It should be noted, however, that different studies have used different operationalized approaches to establish number sense (Whitacre et al., 2020). For instance, some studies have laid the focus on the processing of symbolic and nonsymbolic magnitudes (e.g., Landerl et al., 2009; Mammarella et al., 2021), whereas others have emphasized numeracy skills, such as number naming, ordering, and counting (e.g., Ceulemans et al., 2014), or a combination of these two conceptualizations (e.g., Doabler et al., 2020; Shanley et al., 2017). Additionally, some studies have assessed separate construct(s) of number sense, whereas others used multifaceted approaches. Studies that have included at least one measure of nonsymbolic number sense, symbolic number sense, mapping (i.e. combining nonsymbolic and symbolic information), or numeracy were included in the current meta-analysis.

It has appeared from a recent meta-analysis on 19 studies that primary school children with MLD in general perform lower on number sense tasks compared with their typically developing peers, although this deviation depended on differences in operationalization of number sense (Schwenk et al., 2017). The effect
size was especially large for symbolic number sense but was moderate for non-symbolic number sense. However, this finding does not rule out that some people with MLD do have lower number sense scores whereas others do not (but may experience difficulties with other cognitive skills; Swanson & Jerman, 2006), nor does it answer the question whether people with lower number sense scores show a different profile on the behavioral level (i.e., mathematical performance) compared with the people with MLD with average to high number sense scores.

Domain-general cognitive skills that have been suggested to play a role in MLD are working memory and rapid automatized naming (RAN). Working memory has consistently been found to be a relevant predictor of mathematics. The working memory model (Baddeley & Hitch, 1974) describes three systems: The phonological loop and the visuospatial sketchpad are responsible for temporary storage of verbal and visuospatial information, respectively, and the central executive enables processing and recollection of this information. The central executive has three distinct information-processing tasks: inhibition, shifting, and updating (Miyake et al., 2000). Inhibition is defined as the ability to suppress prepotent but undesired behavior, shifting is the ability to adapt behavior to changing circumstances, and updating is the ability to monitor and revise active verbal or visuospatial information. Working memory in the present study is operationalized as the entire storage and processing system as described by Baddeley and Hitch (1974), including the three executive functions as described by Miyake et al. (2000). All working memory systems are used in mathematical problem solving, but the meta-analysis by Friso-van den Bos et al. (2013) showed that executive function updating showed the strongest correlations with mathematics performance. Several meta-analyses showed that working memory differences between people with MLD and typically performing people were solid, especially for verbal working memory (three studies, Johnson et al., 2010; 43 studies, Swanson & Jerman, 2006) and numerical working memory (20 studies, Peng & Fuchs, 2016; 75 studies, Peng et al., 2018).

Convergent evidence from neurocognitive research suggests that domain-general phonological cognitive skills are related to mathematics as well (Dehaene et al., 2003): RAN is involved in quickly accessing verbal codes from long-term memory that correspond to number facts (Donker et al., 2016; Willburger et al., 2008). Lower RAN scores indeed have been observed in children with MLD who had additional reading difficulties (Landerl et al., 2009). Moreover, large effect sizes for lower scores on RAN in children with MLD became evident from a meta-analysis of 28 studies (Swanson & Jerman, 2006).

Although more and more research is conducted on low performance on skills associated with MLD, many studies have investigated only one or few possible predictors, making a direct comparison of results problematic due to methodological variation between studies. Several meta-analyses have been conducted to align these findings, but they addressed only the cognitive (and not mathematical) characteristics of MLD as opposed to TD. Moreover, these studies incorporated either domain-specific cognitive skills or domain-general cognitive skills but not both (Peng et al., 2018; Swanson & Jerman, 2006). The first goal of this meta-analysis is therefore to make a comprehensive state-of-the-art overview of the mathematical and cognitive characteristics of MLD.
Strong effects of the different cognitive skills on mathematics performance in people with MLD have been reported, but variability among studies is huge. As a case in point, for the 20 studies included in the meta-analysis by Peng and Fuchs (2016) on working memory, effect sizes indicating differences between groups with MLD and typically performing groups ranged from 1.39 to 8.05. Moreover, a large effect was found for RAN in the meta-analysis of Swanson and Jerman (2006), although findings from single studies are mixed for alphanumeric RAN (e.g., naming of letters and digits) and non-alphanumeric RAN (e.g., naming of colors and pictures). For instance, an effect for children with MLD has been found for non-alphanumeric RAN (Donker et al., 2016) but also specifically for naming quantities (Willburger et al., 2008). It can thus be questioned to what extent the studied samples are comparable. Indeed, a meta-analysis on lower cognitive scores within MLD has shown that moderating variables, such as seriousness of MLD, age of the sample under study, and type of used instruments, can have a substantial impact on reported associations between cognitive skills and mathematics (Peng et al., 2018). As the selection criteria for MLD largely differ between studies, it is an important question how such selection criteria might influence study outcomes.

A theoretical framework that incorporates those study characteristics can be derived from the literature as well as clinical practice. According to the Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-5; American Psychological Association, 2013) diagnostic criteria for specific learning disabilities require academic skills to be substantially and quantifiably below what can be expected for chronological age (seriousness), and difficulties are present at least for 6 months despite intervention (persistence). Also, no other disorder, including intellectual disability, may be present that could explain the learning difficulties (specificity). Even though those criteria are rather vague, they do provide three potential selection criteria that scientific studies should consider when selecting their MLD sample: seriousness, persistence, and specificity.

First, the seriousness of MLD is commonly determined by the percentile wherein a person scores for a certain—often standardized—math task. Cutoff criteria for MLD have been found to range from below the 5th percentile to as high as the 46th percentile, which makes this approach rather arbitrary (Murphy et al., 2007). Higher cutoffs (>30th percentile) are often referred to as “low achievement” rather than “MLD” (Landerl et al., 2004).

A second consideration associated with selection criteria for MLD is whether persistency is taken into account. Murphy and colleagues (2007) reviewed 22 studies on this criterion, and only one third of those studies assessed MLD at more than one point in time. Yet, it has become evident that although 77% of the children with MLD retained that diagnosis after 2 years, 23% of the children could no longer be identified as having MLD (Toll et al., 2011).

Third, research on MLD tends to exclude other learning (e.g., dyslexia) and behavioral (e.g., attention-deficit/hyperactivity disorder [ADHD]) disabilities in order to align with the specificity criterion during sample selection, whereas others explicitly investigated the comorbidity of MLD. However, because it has become clear that comorbidity rates are high, it is clear that comorbid difficulties are common rather than exceptional. In addition, research on comorbidity finds
Mathematical and cognitive skills in MLD

either common cognitive deficits or distinct patterns of cognitive weaknesses. For instance, the executive function of sustained attention appears to be impaired in children with MLD, whereas the executive function of inhibition seems affected in children with ADHD (Kuhn et al., 2016). Similarly, Landerl and colleagues (2009) specifically found lower scores on number sense in children with MLD as opposed to lower scores on phonology in children with a reading disability. In contrast, Kaufmann and Nuerk (2008) found that impaired numerical processing of children with ADHD could not be explained by weaker executive functions. Instead, other cognitive factors (e.g., reaction speed) may account for the similarities in math difficulties of children with MLD and ADHD. Additionally, Slot and colleagues (2016) identified phonological awareness as a shared cognitive risk factor for MLD as well as reading and spelling disability. Findings on whether to consider MLD as a specific disability or not thus seems to be inconclusive. It can, therefore, be questioned if it is justified to use such a specificity criterion.

Finally, other sample characteristics, such as age, as well as the type of measure used to assess mathematics and related cognitive skills should also be considered as potential factors that can explain variability in study outcomes. More specifically, as children grow older, the demands of mathematics tasks become increasingly complex. This could result in a shift in the strength of the relation between cognitive skills and mathematics performance (e.g., Nelwan et al., 2018; Van de Weijer-Bergsma et al., 2015). Regarding the type of task used in studies, it has been found that MLD is most often linked to difficulties in basic math, that is, simple arithmetic problems and memorizing basic facts (Geary, 2011b; Huijsmans et al., 2020). For number sense, larger effects have generally been obtained for symbolic tasks compared with nonsymbolic tasks (Schneider et al., 2017). For working memory, correlations were the strongest between mathematics and updating (Friso-van den Bos et al., 2013). Finally, an important aspect of tasks measuring predictors of MLD, for both working memory and RAN, is whether the task content is numerical or not (cf. Donker et al., 2016; Peng & Fuchs, 2016).

To summarize, different empirical studies clearly use diverse selection criteria. However, the extent to which those choices impact research findings remains underexplored. Generalizability across empirical studies and the development of diagnostic instruments for clinical practice may be harshly impeded if math performance and related cognitive skills of people with MLD differ as a consequence of study selection criteria. A second goal of this study was therefore to investigate how selection criteria influence conclusions about differences in cognitive skills between people with MLD and typical performers.

Taken together, the present meta-analysis aims to answer the following questions. First, to what extent are people with MLD different from their typically performing peers with regard to their mathematics performance and the associated cognitive skills (number sense, working memory, and rapid naming)? We hypothesized that there would be significant group differences for all of these characteristics but that there would be variability in the between-group effect sizes. The second research question was, To what extent are study outcomes on the cognitive characteristics of MLD affected by sample selection criteria (i.e., seriousness, persistence, specificity, age, and task type)? It was hypothesized that different selection criteria for MLD result in different profiles of performance on
mathematics as well as cognitive skills. More specifically, the people with the most severe, persistent, and specific math difficulties might constitute the MLD subtype developmental dyscalculia and are thus expected to display a core deficit in the domain-specific cognitive skill number sense. This meta-analysis adds to the existing body of literature on cognitive profiles of MLD in that we explicitly investigated mathematical profiles and have employed different selection criteria to define MLD.

**Method**

**Search Procedure and Terms**

A literature search was conducted to identify empirical studies in which a comparison between people with MLD and controls was being made on cognitive characteristics. The literature review for publications was conducted using internet search engines Scopus, Web of Science, PubMed, and PsycInfo. The search was limited to studies in English published between 2000, which is when the *DSM-IV-TR* (American Psychological Association, 2000) was published, and February 2022, which is when the literature search for the current study was completed. Search terms were “dyscalcul*,” “math* learning disab*,” “MLD,” and “math* learning difficult*.” In addition to this electronic literature search, we conducted hand searches to check for additional relevant studies. We performed a hand search of online-first publications and former volumes of relevant journals (*Journal of Educational Psychology, Journal of Experimental Child Psychology, Journal of Learning Disabilities, Learning Disabilities Quarterly, Learning and Individual Differences, and Research in Developmental Disabilities*). We also manually searched the reference lists of relevant previously published meta-analyses on characteristics of MLD (i.e., Johnson et al., 2010; Peng et al., 2018; Peng & Fuchs, 2016; Schwenk et al., 2017; Swanson & Jerman, 2006). Figure 1 shows the flowchart for the literature search and screening.

The initial search for eligible studies was done by the first author. After removal of duplicates, non-English papers, and studies from irrelevant subject areas (such as geology or cancer research, based on descriptors or journal titles), titles and abstracts were double-screened by (a) two trained research assistants and (b) the first author. The interrater reliability (IRR) for the selection phase was 96.8%. In case of discrepancy, studies were retained for the next step of the procedure.

**Inclusion and Exclusion Criteria**

The full texts of 493 studies were assessed for eligibility by the research assistants and the first author. The IRR was 93.7%. After resolving disagreements, 144 studies were selected. The hand search yielded one additional eligible study. In total, 145 studies were selected that met our inclusion criteria; see Supplementary Table A in the online version of the journal.

All empirical studies published between 2000 and 2022 in English were included when a comparison was made between people with and without MLD on any of the cognitive characteristics: number sense, working memory, and rapid naming. Empirical studies that were included were either group comparison studies in which two or more groups were compared, of which at least one was
identified as having MLD; longitudinal studies in which at least one group with MLD was identified; or intervention studies in which these cognitive characteristics were measured in the pretest and compared with a no-MLD control group. Postintervention measures were not included. For longitudinal studies, only
measures taken after group allocation (MLD or not) were included. Articles that did not contain empirical data (e.g., review studies or opinion papers), meta-analyses, and case studies were excluded.

All studies that reported on MLD, mathematical disabilities, or dyscalculia were included, although the selection criteria largely differed, for example, scores on standardized math tests ranging from below the 2nd percentile to below the 40th percentile. Studies just comparing different achievement levels (e.g., low, average, and high achieving) and studies only reporting on groups with combined learning problems (e.g., reading and math) were excluded, as this would make MLD a very broad and ill-defined concept. Studies were excluded if MLD participants were not primarily selected due to low math performance (e.g., groups selected on other characteristics, such as children with Turner syndrome). Studies on (young) children “at risk” but not yet identified as having MLD were also excluded. Studies were included when the control group represented a typically developing group, but if the control group was selected otherwise (e.g., within a sample of ADHD, those without MLD), the study was excluded.

Studies were included that reported on relevant outcome measures (i.e., math and the target cognitive characteristics). If combined measures were used, which made disentangling the separate cognitive characteristics impossible, studies were excluded. If the article did not include enough data to calculate effect sizes ($n = 14$), the authors were contacted, and some were able to provide additional data ($n = 3$). Moderators of the effect sizes were coded by first entering a verbal description of the relevant information into the data set, such as “ADHD excluded” for sample specificity or “listening recall” for type of working memory task. These descriptions were recoded into dichotomous or categorical variables by the first and third authors upon finalizing the data set.

Data Extraction

In a next step, studies were examined in detail, and data extraction took place (see next section). Effect sizes, constituting differences between MLD groups and typically performing groups and being the dependent variable in the current study, were coded as Cohen’s $d$. When several MLD groups, measures, or time points were reported, separate effect sizes were coded for each group, measure, and time point. Variables that were used to select groups (usually a math test) were not included as outcome measures. When the targeted measure of effect size was reported directly, it was registered in a data set. When means and standard deviations per group were reported, Cohen’s $d$ was computed. Finally, if a different measure of effect size was reported instead that indexed a comparable group difference, the reported measure was converted to Cohen’s $d$ following Lipsey and Wilson (2001) and D. Wilson (2022).

The first and third authors conducted the coding. A random sample ($n = 16$) of the studies was double coded. The interrater correlation (Pearson’s $r$) for the main outcomes was 0.996. The IRR for the moderator variables was acceptable (age, IRR = 100%; mathematics, IRR = 97.9%; number sense, IRR = 91.0%; working memory, IRR = 97.1%; RAN, IRR = 100%). IRRs for the sample selection criteria were lower (seriousness, IRR = 75%; persistence, IRR = 85%; specificity, IRR = 70%) and thus were double coded for the whole selection (IRRs, respectively,
were 96.5%, 98.3%, and 94.8%). Disagreements were resolved by consulting the original paper and discussion by the coders.

Data Appraisal

Sample Age
Sample mean ages were recoded into a categorical variable with three categories. Children were participants with a mean age under 12 years, teenagers were between 12 and 17 years of age, and adults comprised the group ages 18 years and older. The cutoffs between these groups were based on the transition from primary to secondary school in many countries and the age at which they have transitioned from secondary school to other education or work.

Type of Mathematics Task
A distinction was made between basic and complex mathematical problems. Basic math included simple arithmetic problems (i.e., addition and subtraction under 100) or basic facts memorization (i.e., multiplication and division of the tables 1 through 10). Complex mathematics included more complex procedures wherein stepwise problem solving is needed (e.g., conversion of measurements; calculations with fractions, percentages, and decimals) as well as word problems. More specifically, basic math mainly requires fact retrieval processes, whereas complex math involves procedural and conceptual processing. In cases in which both basic and complex mathematical operations were included in the same measure, the measure was categorized as consisting of complex mathematical operations.

Type of Number Sense Task
The following categories were coded: (a) nonsymbolic number sense, for example, dot comparison tasks; (b) symbolic number sense, for example, number comparison tasks; (c) mapping, for example, number line tasks; and (d) numeracy, for example, number naming, ordering, or counting tasks. Nonsymbolic number sense in this respect entails the ability to recognize, understand, and compare quantities of different sizes. The same applies for symbolic number sense with Arabic digits or with number words in some studies. The skill of mapping enables one to transfer rather effortlessly between these quantities, digits, and number words. At last, numeracy involves preparatory math skills, such as number naming or writing, ordering, and counting (Dehaene et al., 2003). Both reaction time and accuracy data were used.

Type of Working Memory Task
A categorical variable was made to distinguish between the types of working memory measures, based on the working memory model presented by Baddeley and Hitch (1974; Baddeley, 2007) and its executive functions (Miyake et al., 2000). Processing speed was added to the resulting categories because of the frequent inclusion of this variable in analyses targeting differences between people with MLD and typically performing people. The following categories of working memory measure were included: (a) inhibition, (b) shifting, (c) verbal updating, (d) visuospatial updating, (e) phonological loop, (f) visuospatial sketchpad, and
Inhibition is operationalized as the ability to suppress automatic but undesired behavior; shifting measures, the ability to adapt behavior to changing circumstances; and updating (either verbal or visuospatial) is defined as the ability to monitor and revise active information (Miyake et al., 2000). Being part of the central executive, inhibition, shifting, and updating all refer to the ability of processing and recollecting information from memory. In contrast, the phonological loop (verbal) and visuospatial sketchpad (visuospatial) are involved in the temporary storage of such information (Baddeley & Hitch, 1974). Finally, processing speed is the mere ability to respond quickly to any given cue. However, note that measures of processing speed specifically did not include the rapid naming of items; those effect sizes were coded as RAN. Moreover, a dichotomous variable was created indicating whether each of the effect sizes relied on number-related tasks to investigate whether any variability in differences between groups could be attributed to the use of working memory tasks in which the processing or recall of numbers was required (e.g., counting span tasks and digit recall tasks).

Type of RAN Task

For rapid naming, a variable was created to indicate whether the task included the rapid naming of digits or that of other matter (e.g., colors, letters). These tasks aim to measure the ability to quickly access number (or other) facts from long-term memory (Willburger et al., 2008).

Sample Selection Criteria

Seriousness. A categorical variable was created to distinguish between studies that defined MLD as less than one standard deviation (>15.9th percentile), one to 1.5 standard deviations (<15.9th percentile), 1.5 to two standard deviations (<6.7th percentile), and more than two standard deviations (<2.3th percentile) below the mean on mathematics.

Persistence. Studies were coded based on whether they adhered to the persistence criterion. Persistence was coded when at least two measurements were used (at least 6 months in between the measurements) or when a former diagnosis of at least 6 months earlier was used.

Specificity. Samples with specific MLD were contrasted against samples that may have included individuals with comorbid disabilities (any comorbidity, e.g., dyslexia, ADHD, intellectual disability) in a binary variable. If it was not reported that people with other disabilities were excluded, it was acknowledged that the sample might have included individuals with comorbid difficulties.

Data Handling and Statistical Analyses

Cohen’s $d$ was used as a first indicator of effect size. If it was not given in the article, it was computed using the formula Cohen’s $d = (M_2 - M_1) / SD_{pooled}$, in which $M_1$ and $M_2$ are the means of the MLD and the control group respectively and $SD_{pooled}$ is the pooled standard deviation for the two groups ($\sqrt{(SD_1^2 + SD_2^2) / 2}$). Prior to data analysis, the directions of all effect sizes were checked and
adjusted so that positive effect sizes indicated better performance of the typically performing group (better performance meaning higher accuracy scores or lower reaction time measures, in most cases) and negative effect sizes indicated better performance of the group with mathematical learning disabilities. Next, all effect sizes that exceeded 3 were recoded as 3 because of the reported problems in replicability and reliability that may especially affect very large effect sizes associated with small samples (Ioannidis, 2008). Finally, we performed a computational correction on all effect sizes, Hedges’ \( g \), because of an upward bias observed in small effect sizes, computed as Cohen’s \( d^* \left(1 - \frac{3}{4N-9}\right) \) (Hedges, 1981; Lipsey & Wilson, 2001). The corrected effect sizes were used as a dependent variable.

Analyses were conducted using the software package R, Version 3.5.3 (R Core Team, 2019). Data inspection was performed using the metafor package (Viechtbauer, 2010). Then, the robumeta package was used to address the research questions (Fisher et al., 2017). This package allows for effect sizes to be nested within studies, which accounts for dependency between effect sizes due to the use of the same sample. Separate analyses were performed to analyze group differences in mathematics performance, number sense, working memory, and RAN.

First, unconditional models were estimated, indexing a weighted mean effect size, and variance around the estimated mean. The resulting variance, distributed as a \( \chi^2 \) value, determined whether moderation analyses were executed. In a next step, moderator variables were added to the model to determine whether variance around the mean effect size could be explained using any of the predictor variables. Predictor variables were sample type (seriousness, persistence, specificity), sample age, and a variety of categorizations of tasks used to assess group differences, namely, type of mathematical problem for mathematics, number sense task type, working memory task type, and the inclusion of number tasks for working memory and for RAN. Analyses were repeated with different reference categories until full comparisons could be made whenever a categorical variable consisted of more than two values.

**Results**

**Group Differences in Mathematics Performance and Cognitive Skills**

Four distinct unconditional models were investigated to test to what extent people with MLD are different from their typically performing peers regarding their mathematics performance and the associated cognitive skills (number sense, working memory, and RAN): one for mathematics and one for each of the associated cognitive skills. First, for mathematics performance, the unconditional model was based on 283 effect sizes describing the difference in mathematics performance of groups with MLD and typically performing groups from 145 studies. The weighted mean difference of Hedges’s \( g = 1.52 \) demonstrated the expected significant differences between the typically performing groups and MLD groups, with \( p < .001 \). Variation around the mean effect size was significant, with \( \chi^2(282, N=283) = 2447.86, p < .001 \). The \( F \) value indicated that of this variation, 90% could be attributed to actual differences in effect sizes rather than sampling error.
A sensitivity analysis showed that the outcomes of the analysis were robust against different estimates of within-study effect size correlations; different values of \( \rho \) produced the same outcomes for up to four decimals. Inspection of the funnel plot and tests for funnel plot asymmetry indicated significant funnel plot asymmetry, which suggests that publication bias may have played a role in the estimation of the weighted mean effect size; Egger's linear regression \( z = 10.46, p < .001 \); rank correlation \( \tau = 0.33, p < .001 \). A trim-and-fill procedure yielded an adjusted weighted mean difference of Hedges’s \( g = 1.30, p < .001 \).

The unconditional model indexing differences in number sense between typically performing control groups and groups with MLD was based on 427 effect sizes drawn from 82 studies. It showed that the weighted mean difference of Hedges’s \( g = 0.88 \) was large and significant, with \( p < .001 \). Variation around the mean effect size was also significant, \( \chi^2(426, N = 427) = 2156.06, p < .001 \). Eighty-four percent of this variation could be attributed to actual differences in effect sizes rather than sampling error, as indicated by the \( I^2 \) value. A sensitivity analysis showed that the outcomes of the analysis were robust against different estimates of within-study effect size correlations; different values of \( \rho \) produced the same outcomes for up to four decimals.

The funnel plot and tests for asymmetry indicated significant funnel plot asymmetry, suggesting that publication bias may have played a role in the estimation of the weighted mean effect size; Egger’s linear regression \( z = 8.02, p < .001 \); rank correlation \( \tau = 0.23, p < .001 \). A trim-and-fill procedure yielded an adjusted weighted mean difference of Hedges’s \( g = 0.79, p < .001 \).

Next, the pool of effect sizes targeting working memory differences between MLD groups and typically achieving groups was investigated. The unconditional model of differences between people with MLD and typically performing people was based on 513 effect sizes drawn from 111 studies. On average, there were significant differences of a medium effect size between groups, Hedges’s \( g = 0.68, p < .001 \). There was also significant variation around the mean effect size, \( \chi^2(512, N = 513) = 1694.63, p < .001 \). The \( I^2 \) value indicated that 74% of this variation could be attributed to actual differences in effect sizes rather than sampling error. According to the sensitivity analysis, the outcomes of the analysis were robust against different estimates of within-study effect size correlations; different values of \( \rho \) produced the same outcomes for three to four decimals. The funnel plot and tests for asymmetry showed significant funnel plot asymmetry, suggesting that publication bias may have played a role in the estimation of the weighted mean effect size; Egger’s linear regression \( z = 5.67, p < .001 \); rank correlation \( \tau = 0.16, p < .001 \). A trim-and-fill procedure yielded an adjusted weighted mean difference of Hedges’s \( g = 0.61, p < .001 \).

The analyses regarding differences in RAN between people with and without MLD were based on 69 effect sizes drawn from 29 studies. Consequently, small sample adjustments were made for all analyses. The unconditional model returned a significant medium effect size of Hedges’s \( g = 0.70, p < .001 \). Variation around this mean effect size was significant, \( \chi^2(68, N = 69) = 246.32, p < .001 \). Of this variance, 81% was based on true differences between samples rather than sampling error, as indicated by the \( I^2 \) value. A sensitivity analysis showed that the
outcomes of the analysis were robust against different estimates of within-study effect-size correlations; different values of $\rho$ produced the same outcomes for two or three decimals. The funnel plot and Egger’s linear regression test for asymmetry showed significant asymmetry, which suggests that publication bias may have played a role in the estimation of the weighted mean effect size; Egger’s linear regression $z = 3.19, p < .001$. However, the rank correlation test for funnel plot asymmetry was nonsignificant; $\tau = 0.14, p < .09$. A trim-and-fill procedure yielded an adjusted weighted mean difference of Hedges’s $g = 0.59, p < .001$.

In sum, samples with MLD and typically performing samples differed on mathematics and all three associated variables. Also, for each set of effect sizes, there was variation around the mean effect size, which may be explained by sample or study characteristics. The next sections present analyses investigating whether any of these characteristics contribute to explaining variability in effect sizes. Various models were tested to explore the contribution of these predictor variables in order to get a full understanding of the inclusion of combinations of variables and to compare categories within categorical variables containing more than two categories. The reported models contain only subsets of variables to maintain parsimony in reporting.

**Type of Task as a Predictor Variable**

The distinction between basic and complex mathematical tasks was entered as a predictor of effect sizes in mathematics describing the differences between groups with MLD and typically performing groups. This variable did not explain any variance in effect sizes, $p = .615$.

Then, the effect of the type of number sense task on effect size of group differences in number sense was tested. Different types of tasks did, on average, produce different effect sizes. A comparison between models with different reference categories revealed that numeracy and mapping tasks produced the largest group differences, indexed as Hedges’s $g$ effect sizes. These effect sizes were statistically equal to one another, and mapping produced larger effect sizes than symbolic and nonsymbolic number sense. Other contrasts were nonsignificant. Table 1 shows an overview of regression weights with symbolic mapping as a reference category.

Next, the effect of the properties of working memory tasks on group differences between groups with MLD and typically performing groups (effect sizes)

<table>
<thead>
<tr>
<th>Task</th>
<th>$B$</th>
<th>SE ($B$)</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.00</td>
<td>0.10</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nonsymbolic number sense</td>
<td>–0.58</td>
<td>0.11</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Symbolic number sense</td>
<td>–0.42</td>
<td>0.10</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Numeracy</td>
<td>–0.27</td>
<td>0.15</td>
<td>&lt;.080</td>
</tr>
</tbody>
</table>

Note. The reference category for number sense in the reported model is mapping.
was investigated. The predictor indicating whether the working memory task included the use of any numerical information did not relate to the effect sizes, \( p = .839 \). The type of working memory measures, however, did play a small role in the effect size variability. Specifically, inhibition tasks yielded smaller effect sizes than both verbal updating tasks, \( p = .041 \), and visuospatial updating tasks, \( p = .015 \). Moreover, processing-speed measures yielded larger effect sizes than all other working memory measures (ps < .05) except visuospatial updating tasks.

Finally, it was investigated whether any variability in effect sizes of group differences in RAN was explained by a variable that indicated whether the RAN task contained the rapid naming of digits or not. This variable did not add any explained variance to the pool of effect sizes, \( p = .484 \).

**Type of Sample as a Predictor Variable**

Next, we explored whether any sample selection criteria explained variability in the pools of effect sizes targeting group differences in mathematics performance, number sense, working memory, and RAN. The sample selection criteria we used were age and the three indicators of MLD: seriousness, persistence, and specificity.

In a first subset of models, the seriousness of the delay and persistence of mathematical difficulties were entered as predictors of group differences in mathematics performance between groups with MLD and typically performing groups (effect sizes). The analyses showed that samples in which up to one standard deviation was used as a cutoff criterion for the MLD group had smaller discrepancies with the typically performing control group than samples in which up to two standard deviations was used as a cutoff. Moreover, the addition of an interaction term showed that when selected for persistence, effect sizes associated with groups selected using a two-standard-deviation cutoff became smaller compared with groups selected using a one-standard-deviation cutoff (\( p = .002 \)), counteracting the positive main effect. Upon adding this interaction term, groups selected with cutoffs larger than two standard deviations also showed larger effect sizes compared with one-standard-deviation samples, an effect that was not present without the interaction term. For this model, see Table 2. Analyses with different reference categories for seriousness showed that there were no other differences between categories. Persistence of the mathematical difficulties did not produce main effects in any of the models investigated.

Specificity, as the last of the MLD characteristics, was entered into a separate regression model as a predictor of variance in performance differences between groups with MLD and typically performing groups (effect sizes). When specificity was used as a selection criterion, samples showed larger mathematics discrepancies from control groups compared with when no specificity criterion was used (\( p = .002 \)).

In a next step, age was entered into a model first as a separate predictor and then in interaction with persistence of the mathematical difficulties. Effect sizes were larger for samples consisting of teens than for samples of children, \( p < .001 \). The contrast between samples of children and samples of adults also turned out significant when the interaction with persistence was added to the model, with adults showing larger effect sizes, \( p = .03 \). The interaction between age and
persistence itself did not contribute to explaining variation in effect sizes in any of the models. Table 3 displays the model with the interaction included and samples of children as a reference category.

Then, the effects of seriousness and persistence of the mathematical difficulties on the effect sizes of number sense were explored. Models in which either seriousness or persistence was added as a single variable did not yield any significant relations between the predictors and variance in group differences indexed as effect sizes. However, when an interaction term between both seriousness and persistence was added to the model, the results demonstrated that MLD groups selected with a two-standard-deviation cutoff criterion on mathematics differed more from their typically achieving peers than groups selected with a 1.5-standard-deviation cutoff criterion. Also, effect sizes were larger when samples were selected for persistence but only when samples selected with a 1.5-standard-deviation cutoff were used as a reference category. Moreover, there

### TABLE 2

*Model predicting variability in effect sizes: Group differences in mathematics performance with seriousness and persistence as predictor subset*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.88</td>
<td>0.13</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Seriousness: 1.5 SD</td>
<td>0.06</td>
<td>0.28</td>
<td>.843</td>
</tr>
<tr>
<td>Seriousness: 2 SD</td>
<td>1.08</td>
<td>0.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Seriousness: 2+ SD</td>
<td>0.57</td>
<td>0.13</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Persistence</td>
<td>−0.01</td>
<td>0.18</td>
<td>.964</td>
</tr>
<tr>
<td>Seriousness×Persistence</td>
<td>0.60</td>
<td>0.38</td>
<td>.122</td>
</tr>
<tr>
<td>Seriousness2 × Persistence</td>
<td>−0.60</td>
<td>0.30</td>
<td>.049</td>
</tr>
<tr>
<td>Seriousness2+ × Persistence</td>
<td>0.03</td>
<td>0.52</td>
<td>.954</td>
</tr>
</tbody>
</table>

*Note.* Seriousness reference category = 1 SD; persistence reference category = no.

### TABLE 3

*Model predicting variability in effect sizes: Group differences in mathematics performance with age and persistence as predictor subset*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.04</td>
<td>0.18</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Age: Teen</td>
<td>0.67</td>
<td>0.18</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Age: Adult</td>
<td>0.94</td>
<td>0.43</td>
<td>.031</td>
</tr>
<tr>
<td>Persistence</td>
<td>0.06</td>
<td>0.22</td>
<td>.768</td>
</tr>
<tr>
<td>Persistence×Teen</td>
<td>0.35</td>
<td>0.24</td>
<td>.144</td>
</tr>
<tr>
<td>Persistence×Adult</td>
<td>−0.67</td>
<td>0.56</td>
<td>.237</td>
</tr>
</tbody>
</table>

*Note.* Age reference category = “child”; Persistence reference category = “no”.

persistence itself did not contribute to explaining variation in effect sizes in any of the models. Table 3 displays the model with the interaction included and samples of children as a reference category.

Then, the effects of seriousness and persistence of the mathematical difficulties on the effect sizes of number sense were explored. Models in which either seriousness or persistence was added as a single variable did not yield any significant relations between the predictors and variance in group differences indexed as effect sizes. However, when an interaction term between both seriousness and persistence was added to the model, the results demonstrated that MLD groups selected with a two-standard-deviation cutoff criterion on mathematics differed more from their typically achieving peers than groups selected with a 1.5-standard-deviation cutoff criterion. Also, effect sizes were larger when samples were selected for persistence but only when samples selected with a 1.5-standard-deviation cutoff were used as a reference category. Moreover, there
was a significant interaction effect between the one- versus 1.5-standard-deviation-cutoff groups and persistence, which showed that when MLD groups were selected both for the more stringent cutoff criterion and for persistence of difficulties, they displayed larger differences with their typically developing peers than when only one of these criteria was maintained. The model including this interaction can be found in Table 4. A graphical representation of group differences can be found in Figure 2.

Group differences between groups with MLD and typically performing groups (effect sizes) in number sense did not vary as a function of the specificity of difficulties of the MLD group that was included in studies, $p = .174$, nor did effect sizes vary as a function of the sample age when age was entered as the only

### TABLE 4

Model predicting variability in effect sizes: Group differences in number sense with seriousness and persistence as predictor subset

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$B$</th>
<th>$SE$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.73</td>
<td>0.10</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Seriousness: 1.5 $SD$</td>
<td>−0.19</td>
<td>0.11</td>
<td>.083</td>
</tr>
<tr>
<td>Seriousness: 2 $SD$</td>
<td>0.10</td>
<td>0.15</td>
<td>.503</td>
</tr>
<tr>
<td>Seriousness: 2+ $SD$</td>
<td>−0.10</td>
<td>0.10</td>
<td>.328</td>
</tr>
<tr>
<td>Persistence</td>
<td>−0.11</td>
<td>0.20</td>
<td>.570</td>
</tr>
<tr>
<td>Seriousness: 1.5 $SD \times$ Persistence</td>
<td>0.47</td>
<td>0.23</td>
<td>.047</td>
</tr>
<tr>
<td>Seriousness: 2 $SD \times$ Persistence</td>
<td>0.10</td>
<td>0.25</td>
<td>.697</td>
</tr>
<tr>
<td>Seriousness: 2+ $SD \times$ Persistence</td>
<td>0.10</td>
<td>0.27</td>
<td>.710</td>
</tr>
</tbody>
</table>

*Note.* Seriousness reference category = 1 $SD$; persistence reference category = no.

![FIGURE 2. Mean effect sizes per category of seriousness and persistence.](image)
predictor variable. However, when the hypothesized interaction between age and persistence was added to the model, both age differences and an interaction between age and persistence emerged. The model demonstrated that both children ($p = .001$) and teens ($p < .001$) from MLD samples differed more from their typically achieving peers than adults from MLD samples. Moreover, there was an interaction between age and persistence demonstrating that if samples consisted of adults and were selected for persistence, effect sizes were higher than in both the samples consisting of children ($p < .001$) and those consisting of teens ($p < .001$), counteracting the main effect of age. Moreover, when samples consisting of teens were used as a reference category, persistence explained significant variation in effect sizes, with samples selected for persistence displaying larger effect sizes, $p < .001$. A model including the interaction effect can be found in Table 5.

Next, the variation in effect sizes regarding working memory differences was explored. Again, various models were tested to account for variations in predictor sets and variation of reference categories of categorical variables. Seriousness, as a first predictor variable, explained variance in group differences (effect sizes) in working memory: Groups selected with a cutoff of one or two standard deviations deviated less from their typically performing peers regarding working memory than groups selected with a cutoff of 1.5 or more than two standard deviations (Table 6).

Persistence, when used as the only predictor variable in a model, explained variance in effect sizes ($p = .039$) with groups that were selected based on a persistence criterion, displaying larger deviations from their typically developing peers than groups that were selected without this criterion. None of the contrasts previously reported persisted when an interaction term between seriousness and persistence was added to the model. Also, effect sizes of working memory differences between MLD and typical groups did not vary as a function of specificity of the sample selection, $p = .974$.

Effect sizes for working memory did not vary as a function of the age group comprising the sample when age was entered as the only predictor to the model. However, when an interaction between age and persistence was added, age

### Table 5

Model predicting variability in effect sizes: Group differences in number sense with age and persistence as predictor subset

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$B$</th>
<th>$SE$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.62</td>
<td>0.08</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age: Teen</td>
<td>-0.08</td>
<td>0.08</td>
<td>.338</td>
</tr>
<tr>
<td>Age: Adult</td>
<td>-0.27</td>
<td>0.08</td>
<td>.001</td>
</tr>
<tr>
<td>Persistence</td>
<td>0.11</td>
<td>0.13</td>
<td>.398</td>
</tr>
<tr>
<td>Persistence $\times$ Teen</td>
<td>0.03</td>
<td>0.13</td>
<td>.814</td>
</tr>
<tr>
<td>Persistence $\times$ Adult</td>
<td>0.75</td>
<td>0.16</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Note. Age reference category = child; persistence reference category = no.
differences in effect sizes emerged, with children showing larger working memory discrepancies with their typically achieving peers than teenagers, $p = .049$. Moreover, interaction terms between persistence and age showed that effect sizes were larger for samples of adults than for samples of children when samples were selected for persistence in mathematical difficulties, $p = .028$. Table 7 displays one of the models containing this interaction.

Finally, for the effect sizes of group differences in RAN, there were larger effect sizes for samples selected based on a two-standard-deviation cutoff than for samples selected based on a one-standard-deviation cutoff. Moreover, whereas persistence did not explain variance in effect sizes, $p = .102$, samples selected for specificity produced smaller effect sizes than samples in which no specificity criterion was used, $p = .045$. The interaction between seriousness and persistence was not tested because of the small number of degrees of freedom and the fact that not all combinations of categories were present.

Next, age was entered as a predictor into the model. There were no differences between groups with MLD and typically performing groups (effect sizes) between child and adult samples; teen samples were not present in the data set. However, the results demonstrated that effect sizes were higher for samples in which adults were selected for persistence of mathematics difficulties when an interaction between age and persistence was added to the model, $p = .008$. However, because

### TABLE 6
**Model predicting variability in effect sizes: Group differences in working memory**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$B$</th>
<th>SE ($B$)</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.48</td>
<td>0.07</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Seriousness: 1.5 $SD$</td>
<td>0.20</td>
<td>0.09</td>
<td>.026</td>
</tr>
<tr>
<td>Seriousness: 2 $SD$</td>
<td>0.03</td>
<td>0.08</td>
<td>.717</td>
</tr>
<tr>
<td>Seriousness: 2+ $SD$</td>
<td>0.40</td>
<td>0.15</td>
<td>.008</td>
</tr>
</tbody>
</table>

*Note. The reference category for seriousness in the reported model = 1 $SD$.*

### TABLE 7
**Model predicting variability in effect sizes: Group differences in working memory with age and persistence as predictor subset**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$B$</th>
<th>SE</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.54</td>
<td>0.07</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age: Teen</td>
<td>−0.21</td>
<td>0.11</td>
<td>.049</td>
</tr>
<tr>
<td>Age: Adult</td>
<td>−0.08</td>
<td>0.13</td>
<td>.520</td>
</tr>
<tr>
<td>Persistence</td>
<td>0.14</td>
<td>0.08</td>
<td>.104</td>
</tr>
<tr>
<td>Persistence $\times$ Teen</td>
<td>0.26</td>
<td>0.18</td>
<td>.143</td>
</tr>
<tr>
<td>Persistence $\times$ Adult</td>
<td>0.33</td>
<td>0.15</td>
<td>.028</td>
</tr>
</tbody>
</table>

*Note. Age reference category = child; persistence reference category = no.*
of the small number of degrees of freedom, which also resulted from the small sample size adjustment, this result as well as all other results concerning effect sizes of RAN differences should be interpreted with extreme caution.

Discussion

In this meta-analysis, 147 studies were analyzed that compared samples of persons with MLD with controls on several mathematical and cognitive skills. The goal was to determine which mathematical skills (e.g., basic vs. complex math) and cognitive skills (number sense, working memory, and RAN) characterize the MLD group best. As expected, MLD groups performed worse on mathematical skills than control groups, although no differences in effect sizes were found between basic and complex mathematical skills. This means that, in general, people with MLD have comparable difficulties with basic and complex mathematical skills. Although the difficulties with basic skills are generally mentioned as a core characteristic of MLD (e.g., Geary, 2011b), it is also apparent that difficulties with basic mathematics will influence performance on more complex mathematical tasks (cf. Kleemans et al., 2018). This results in difficulties over the whole spectrum of mathematical skills. An interesting finding was that larger effect sizes were found in studies with teenagers than with children. This finding also seems to suggest that early difficulties affect later difficulties, and for those with serious mathematical difficulties, the differences with the peer group become larger over time. Another explanation might be that school-age children practice their math skills on a daily basis, and especially children with MLD often get more instruction and practice than their peers, to keep the delay as small as possible. Teenagers with MLD, on the contrary, may decide not to elect mathematics as a part of their study programs, with less practice and an even larger deviation from peers as a consequence. This explanation emphasizes the need for daily practicing for children with MLD, because this may reduce their delay in mathematics learning.

A more interesting question, however, is how MLD is related to the cognitive skills. MLD groups differed from the control groups on number sense tasks. The largest differences were found in mapping tasks, for which people have to make a connection between symbolic and nonsymbolic representations of number. The differences between MLD and controls were smaller in symbolic and nonsymbolic tasks, which concerned mainly comparison tasks. These findings show that people with MLD indeed have difficulties with these basic number sense skills, which is fully in line with former research (e.g., Schneider et al., 2017). In addition, the results also show that the difficulties people with MLD have are larger in mapping tasks, which more directly reflect skills that are needed to successfully tackle mathematics problems. Therefore, it may be better if future research incorporates mapping tasks in their study designs instead of the commonly used comparison tasks. This is also more in line with practice. For instance, number line tasks (i.e., mapping) are often included in the math curriculum in primary schools.

Furthermore, MLD groups differed from controls in working memory and RAN performance. This finding confirms that working memory and RAN are on average weaker in people with MLD compared with controls. For working memory, the largest effect sizes were found for processing speed and verbal and
visuospatial updating tasks, and effect sizes for inhibition were the smallest. These results are in line with former research on the relation between working memory and mathematical performance in typical samples (Friso-van den Bos et al., 2013), in which the strongest relations were found with updating and the weakest with inhibition (and shifting). No differential effects were found for tasks with or without numerical content, neither for working memory nor for RAN. This means that we did not find evidence for the hypothesis that people with MLD have difficulties with the processing of numbers per se. Instead, it appears that a general access deficit to information stored in long-term memory underlies the math difficulties of people with MLD (Geary, 2004; Koponen et al., 2017). This finding contradicts a previous meta-analysis (Peng et al., 2018) that showed that difficulties with working memory were more severe in the numerical domain as opposed to the visuospatial domain. Nevertheless, these authors also state, on the basis of an extensive battery of cognitive skills, that “any individual who is identified with MLD is likely to experience both domain-specific (numerical processing) and domain-general cognitive deficits” (Peng et al., 2018, p. 463). Moreover, the fact that effect sizes associated with processing speed were larger than effect sizes associated with most other working memory skills may suggest that processing speed affects mathematical aptitude at a more basic level than most working memory tasks, by indexing the efficiency with which information can be encoded regardless of the context of a specific task (see, e.g., Baddeley, 2007), which may be especially detrimental to MLD groups. To summarize, the cognitive skills of people with MLD are on average weaker compared with people without MLD. There also was an effect of task type, which indicated that some of those cognitive skills (i.e., mapping, processing speed, and updating) are more strongly related to mathematics than others (i.e., comparison and inhibition), but whether the task included operations with numbers did not matter.

A second goal of this meta-analysis was to study how the aforementioned outcomes are affected by sample selection criteria (i.e., seriousness, persistence, and specificity). First of all, we tested whether the seriousness of the math difficulties of the study sample affected the outcomes on the mathematical and cognitive variables. Seriousness as single predictor had consequences for the effect sizes in mathematics and RAN but not number sense. It also predicted effect sizes of working memory differences between MLD groups and typically performing control groups, but contrasts did not uniformly point toward larger deviations in working memory when more stringent selection criteria were used—a finding that was both unexpected and difficult to explain and that may be indicative of a role of factors outside the scope of this analysis, such as strategy selection in mathematics that draws on and facilitates working memory in differential ways for groups with varying levels of difficulties (for an account of the relation between working memory and strategy use, see van der Ven et al., 2012).

Persistence alone also could not explain variances in effect sizes, except for working memory: In samples in which math difficulties persevere, working memory skills are weaker. However, this effect was not robust against the addition of interactions with seriousness to the model, suggesting that part of the variance associated with persistence is also associated with seriousness for this set of effect

20
sizes. For number sense, a significant interaction effect between persistence and seriousness was found: Especially for the group selected on a score between one and 1.5 standard deviations below the mean (e.g., below the 16th percentile), this group showed a larger discrepancy in math scores only when their difficulties were persistent. Such a clear profile or trend in that direction, however, could not be found for the other selection criteria. This may be explained by the following assumption. Persistence may not matter anymore for two possible reasons when one’s math difficulties are extremely severe. First, when math performance is (more than) two standard deviations below the mean, difficulties may always be persistent, because even when a person increases slightly in performance between two assessments, they will still score significantly below the mean (e.g., increase from two standard deviations to 1.8 standard deviations). In that case, there is no variance left to explain in number sense; thus an interaction between seriousness and persistency cannot be obtained. Second, when math performance is extremely impaired, it could be that number sense is affected as well. Number sense has been suggested to be conditional for proficiency in math (Von Aster & Shalev, 2007); thus if math difficulties are severe, number sense is likely to be deficient as well, and vice versa. Seriousness may therefore be more important than persistency when investigating MLD.

Specificity of the sample selection explained variance in effect sizes for both mathematics tasks and RAN. For mathematics, samples selected for specificity showed larger effect sizes than those in which specificity was not taken into account. This suggests that when specificity is a factor, difficulties are consistently visible across a variety of mathematics tasks and not only the tasks that are used for sample selection. For RAN, however, effect sizes for samples that were selected with a specificity criterion were smaller than those of samples selected with no specificity criterion. This may suggest that groups that display comorbid difficulties are at an additional risk of having access deficits, which coincides with research on RAN in groups of children at risk for learning problems in a single domain versus multiple domains (Pauley et al., 2011). This may indicate that the access deficits brought to the surface using RAN tasks rely on other learning difficulties than MLD. Of course, it is unknown whether these studies included people with possible comorbid difficulties at all and, if so, to what extent. Further research is necessary to give insight in the role of the specificity criterion. An additional suggestion for further research is to differentiate between different types of comorbid difficulties (e.g., reading, attention, or intellectual disabilities).

None of the sample selection criteria variables produced consistent and robust effects on the differences between MLD groups and typically performing control groups. Although some effects are visible, such as effects of seriousness on mathematics, working memory, and RAN, these effects are often counteracted by those of persistence or display unexpected patterns, such as a subset of samples selected using more stringent criteria scoring closer to their typically achieving peers in working memory than a subset of samples selected using more lenient criteria. The assertion that samples with MLD can be described using a clear cognitive profile is not supported by these analyses. Similar findings have been obtained in a meta-analysis by Peng et al. (2018), wherein seriousness of MLD also was
related to the severity of some cognitive difficulties but not all of the subset of cognitive characteristics included.

Limitations

We did not find consistent differential effects of sample selection criteria, which could also be an artefact or our method: We included studies from 26 different countries, and educational systems between countries differ considerably. Cutoff criteria are therefore possibly less comparable. If the cutoff percentages were, for example, not based on the population mean but on the study sample, it does make a difference whether mainstream education is inclusive or not (i.e., the percentages of children that are in special education differ greatly across countries). Thus, one should always consider how serious the difficulties are when dealing with MLD, because this could be an indication of underlying (cognitive) factors.

In addition, meta-analyses by default build on the work of previous empirical studies. It is therefore important to take potential reasons for effect size variation across studies into account when conducting a meta-analysis (Alexander, 2020). The studies incorporated in this meta-analysis are conducted in different contexts and with different methods, which may be a threat to the external validity of the present study. We have accounted for this variability by explicitly coding information regarding the contexts and samples on which effect sizes were based and using them as predictors of effect sizes. Any variance in group differences that can be attributed to the differences in methodology that we coded can therefore be meaningfully interpreted. This does not preclude, however, that there may be other sources of methodological variability that were not accounted for in the current study and turn up as error in the presented models.

Another limitation of the present study is that we did not evaluate the reliability of our separate search procedures. Instead, we used a multipronged process including electronic, hand, and ancestral searches. We believe that this confers sufficient confidence that relevant articles were located and could be independently replicated. Last, in contrast to previous meta-analyses on cognitive profiles of MLD, we have mixed domain-specific and domain-general cognitive skills, incorporated mathematical profiles, and employed different gradations to selection criteria for MLD. In this sense, it is an important addition to previous work on the profiles of MLD. However, it should be noted that next to the most common cognitive skills that were included in this meta-analysis (i.e., number sense, working memory, and RAN), we did not include other cognitive skills that might play a role in learning mathematics as well. Further research could lead to a more comprehensive picture of MLD, although our results already converge to a quite clear picture: The seriousness of MLD relates to cognitive factors, but specific profiles could not be detected.

Conclusion

To conclude, people with MLD on average perform weaker on mathematics and all underlying cognitive skills included in this meta-analysis (number sense, working memory, and RAN) compared with typically performing people, independently of the selection criteria used to define MLD. Thus, contrary to formerly
proposed profiles (e.g., Mazzocco, et al., 2011; Piazza et al., 2010; Simmons & Singleton, 2008; Von Aster & Shalev, 2007), we were not able to detect a consistent and specific profile of cognitive mechanisms underlying more serious, persistent, or specific difficulties. Where more serious difficulties in mathematics were associated with larger deviations from typically performing groups on cognitive tasks, it cannot be ruled out that this is—at least partly—explained by more general cognitive difficulties, as tasks such as those used to measure working memory are closely related to general intelligence (e.g., Giofrè et al., 2017).

Therefore, we argue for expanding the criteria for diagnosing MLD, in contrast to the sometimes very specific criteria including different levels of seriousness or specificity that are often used. Every person with math difficulties should be able to receive the support they need to decrease their difficulties, independently of what seem to be arbitrary cutoffs that are commonly used in research and practice, because there is no evidence that the group with more serious, persistent, or specific mathematical difficulties displays a different cognitive profile underlying their math difficulties or has different educational needs compared to a group with less severe MLD. To elaborate, a consistent and specific profile for MLD was not found in the present meta-analysis; thus there appears to be no empirical support for a specific label or diagnosis, such as dyscalculia, that assumes a specific underlying deficit or cognitive profile. This finding, aligned with a recent opinion paper by Peters and Ansari (2020), suggests that future research to learning disabilities could step away from making group comparisons and focus instead on the continuum of academic abilities. For educators, this also implies that MLD should not be seen as a distinct disability that could be distinguished from other forms of mathematical difficulties and that mathematics performance should be assessed on a continuum. Additional instruction or remediation should be given to children scoring low on this continuum, whether it is (yet) persistent or not, to improve their math performance. This means that more attention should be paid to the individual characteristics and educational needs of all students with mathematical difficulties, taking into account their specific strengths and weaknesses. Hence, every person with mathematical difficulties, regardless of how severe their impairments are, should preferably receive additional time and support to (partly) overcome their struggles with learning mathematics. Moreover, instructions and interventions should be tailored to the specific cognitive profiles of children receiving assistance rather than to general conceptions about the cognitive background of children with MLD, because no general cognitive profile for all children with MLD could be detected. After all, we have demonstrated that on average there are cognitive differences between people with and without MLD, but these differences are not the same for all groups in all contexts, let alone for all individuals within the groups.

**ORCID iDs**

Evelyn H. Kroesbergen https://orcid.org/0000-0001-9856-096X
Marije D. E. Huijsmans https://orcid.org/0000-0003-0630-3283
Ilona Friso-van den Bos https://orcid.org/0000-0003-3722-254X
References

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


*McCaskey, U., von Aster, M., Maurer, U., Martin, E., O’Gorman Tuura, R., & Kucian, K. (2018). Longitudinal brain development of numerical skills in typically...


Kroesbergen et al.


Mathematical and cognitive skills in MLD


Kroesbergen et al.


Authors

EVELYN H. KROESBERGEN is a full professor of learning disabilities at Radboud University, Faculty of Social Sciences, Thomas van Aquinostraat 4, 6525 GD Nijmegen, Netherlands; email: evelyn.kroesbergen@ru.nl. The focus of her research is on the typical and atypical development of mathematics and related cognitive factors. A second line of research focuses on creativity and giftedness, also in relation to mathematics.

MARIJE D. E. HUIJSMANS is a special needs coordinator and mathematics expert in primary education at Leerplein055, 2e Wormseweg 80, 7331 VG, Apeldoorn, Netherlands; email: m.d.e.huijsmans@gmail.com. The focus of her dissertation, written at Radboud University, was on explaining variation in mathematics achievement.

ILONA FRISO-VAN DEN BOS is a researcher at the University of Twente, Faculty of BMS, Sections OWK/ELAN, Drienerloalaan 5, 7522 NB Enschede, Netherlands; email: i.friso-vandenbos@utwente.nl. The focus of her current research is on the affordances of virtual reality applications to facilitate situation awareness and other competencies of vocational engineering students. Previously, she worked on research targeting the development, facilitation, and assessment of mathematical skills.