No evidence for a core deficit in developmental dyscalculia or mathematical learning disabilities

Irene C. Mammarella[1], Enrico Toffalini[2], Sara Caviola[3], Lincoln Colling[4], & Denes Szűcs[5]

[1] *Department of Developmental and Social Psychology, University of Padova, Italy*

[2] *Department of General Psychology, University of Padova, Italy*

[3] *School of Psychology, University of Leeds, UK*

[4] *School of Psychology, University of Sussex, UK*

[5] *Department of Psychology, University of Cambridge, UK*

Correspondence concerning this article should be addressed to:

Irene C. Mammarella

Department of Developmental and Social Psychology

University of Padova

Via Venezia 8, 35131 Padova, Italy

E-mail: irene.mammarella@unipd.it

Denes Szűcs

Department of Psychology, University of Cambridge, UK

Cambridge University

Downing Street, Cambridge, UK, CB2 3EB

E-mail: ds377@cam.ac.uk

Abstract

Background: Two hypotheses were tested regarding the characteristics of children with mathematical learning disabilities (MLD): (i) that children with MLD would have a "core deficit" in basic number processing skills; and (ii) that children with MLD would be at the end of a developmental continuum and have impairments in many cognitive skills.

Methods: From a large sample (N=1,303) of typically-developing children, we selected a group definable as having MLD. The children were given measures of basic number processing and domain-general constructs. Differences between the observed sample and a simulated population were estimated using Cohen's d and Bayes factors. Receiver operating characteristic curves were plotted and the area under the curve was computed to ascertain the diagnostic power of measures.

Results: Results suggest that the differences between the MLD and control group can be defined along with general characteristics of the population rather than assuming single or multiple "core deficits". None of the measures of interest exceeded the diagnostic power that could be derived via simulation from the dimensional characteristics of the general population.

Conclusions: There is no evidence for core deficit(s) in MLD. We suggest that future research should focus on representative samples of typical populations and on carefully tested clinical samples confirming to the criteria of international diagnostic manuals. Clinical diagnoses require that MLD is persistent and resistant to intervention, so studies would deliver results less exposed to measurement fluctuations. Uniform diagnostic criteria would also allow for the easy cross-study comparison of samples overcoming a serious limitation of the current literature.

Introduction

The literature on developmental dyscalculia, and on mathematical learning disabilities has been dominated in the past by two main hypotheses: (i) the core deficit hypothesis of domain-specific numerical acuity, magnitude representation or number sense (Butterworth, 1999; Dehaene, 2011; Piazza et al., 2010); and (ii) the domain-general hypothesis (Geary, 2004; Passolunghi & Siegel, 2004). In previous studies the frequent use of terms such as “developmental dyscalculia” or “mathematical learning disability” (MLD) largely reflected these two hypotheses and attempted to describe different clinical profiles of children with math difficulties. On the one hand, researchers assumed the existence of *pure* developmental dyscalculia on the grounds of specific endogenous impairments in basic number processing. On the other hand, the MLD term was used for children with impairments in mathematics and general cognitive deficits (not specific to numerical processing) in working memory, visuospatial processing or attention, for instance (Price & Ansari, 2013; Rubinsten & Henik, 2009).

The core deficit hypothesis stemmed from the observation that infants (Izard, Dehaene-Lambertz, & Dehaene, 2008; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004), like other animal species (Cantlon & Brannon, 2005; Hauser, Tsao, Garcia, & Spelke, 2003), can discriminate between small and large numerosities. This prompted the suggestion that a specific deficit in the ability to discriminate numerosities could *cause pure* developmental dyscalculia (Butterworth, 1999; Mazzocco et al., 2011). Other studies focused instead on a more specific weakness in symbolic (Arabic digits) magnitude discrimination ability (De Smedt, Noël, Gilmore, & Ansari, 2013; Mussolin, Mejias, & Noël, 2010). Within this framework, a recent meta-analysis (Schwenk et al., 2017) found that children with developmental dyscalculia showed longer response times than typically developing controls in symbolic magnitude comparison tasks (Hedges’ g = 0.75; 95% CI [0.51; 0.99]) whereas group differences were not that expressed in non-symbolic magnitude discrimination tasks (Hedges' g = 0.24; 95% CI [0.13; 0.36]). The study concluded in favor of the diagnostic importance of the symbolic rather than the non-symbolic magnitude discrimination task.

As for the domain-general hypothesis, many researchers (Geary, 2004; Passolunghi & Siegel, 2004) studied the cognitive profiles of children with MLD, based on the assumption that cognitive impairments in attention, short-term memory, working memory, or executive functions would contribute to explaining the mathematical difficulties observed in children with MLD. A recent meta-analysis (Peng, Wang, & Namkung, 2018) examined cognitive deficits in children with MLD (including phonological processing, processing speed, short-term and working memory, visuospatial skills, attention and executive functions). It concluded that individuals with MLD show varying degrees of impairments in different cognitive skills. Deficits in processing speed and working memory seem to be the most salient and stable cognitive markers of MLD. The study showed that the observed deficits in the various cognitive domains depended on diverse study parameters such as comorbidity in participant groups, screening methods, severity of MLD, and age, and the authors highlighted the discrepancies in the criteria used for selecting MLD (Peng et al., 2018).

In fact, past studies have adopted widely varying criteria when diagnosing MLD (Devine, Soltesz, Nobes, Goswami, & Szucs, 2013; Murphy, Mazzocco, Hanich, & Early, 2007). For example, percentile cut-offs ranging from the 35th (Geary, Hamson, & Hoard, 2000) to the 10th percentile (Murphy et al., 2007) and using mathematical performance equivalent to children one or two years younger (Shalev, 1997; Temple & Sherwood, 2002). In addition, not all studies actually tested children with a previous clinical diagnosis; some simply selected children with a low achievement profile from larger typically-developing samples (i.e., by using only psychometric cut-offs). Table 1 summarizes the characteristics of the samples in the studies included in the two above-mentioned recent meta-analyses (Peng et al., 2018; Schwenk et al., 2017). See also in Table S1 for further details of the criteria used.

Table 1 about here

Based on those recent meta-analyses (Peng et al., 2018; Schwenk et al., 2017) some important characteristics of previous research are worth summarizing: a) 88% of studies used only psychometric cut-offs but no clinical diagnosis to select MLD children from larger typically-developing samples; b) studies used extremely diverse selection criteria, in both studies considering children with or without a previous clinical diagnosis (see Table 1); c) 39% of studies had weak statistical power ($N \leq 20$), most likely overestimating effect sizes characterizing between-group differences (Szucs & Ioannidis, 2017). Considering the above observations, it is unlikely that the available scientific literature provides a clear picture of the cognitive profile of children with MLD.

From a clinical perspective, the DSM 5 (American Psychiatric Association, 2013) suggests criteria for diagnosing children with MLD (cf. Footnote 1). Although "specific learning disorders" have been combined in the light of a dimensional approach, it is still possible to specify which subdomains of learning are impaired in MLD. The manual suggests using the term "specific learning disorder with impairment in mathematics" with reference to MLD. According to the DSM 5 criteria, for individuals to be diagnosed with a specific learning disorder (such as MLD), they should show symptoms of the corresponding impairments for at least six months despite specific interventions targeting their difficulties. Specific learning disorders occur in children with average intellectual abilities and produce lifelong impairments in activities reliant on certain learning skills. Although the DSM 5 recognizes that any cut-off used to differentiate between individuals with and without specific learning disorders is arbitrary, they recommend considering scores at least 1.5 SD below the mean on standardized tests. The manual also specifies that not only psychometric cut-offs, but also the other criteria it mentions (e.g. the persistence of impairments despite specific interventions) should be met.

In the current study we addressed the diagnostic problems discussed above by taking a novel analytical approach. Our main question was whether any core deficit can be observed in MLD children or whether they simply reflect the characteristics of the general population. Using cut-offs commonly adopted in the literature we identified 47 children with MLD and 895 control children from a sample of 1,303 children. To see whether describing MLD groups requires information that cannot be inferred from the general population we adopted a simulation procedure. In step one, we computed the standardized differences between MLD cases and controls along each measure of interest. Then, excluding children with MLD, we simulated a large population that reflected the same set of correlations between all the variables as measured in our sample. We again computed the standardized differences between the simulated MLD cases and controls for all variables. Finally, we compared the standardized differences between the observed and simulated MLD cases. A discrepancy between observed and simulated cases would indicate that examining specific groups provides more information than examining the characteristics of the general population.

To further formalize the above we computed a Bayes factor for each measure of interest. We call our null hypothesis (H_0) the ‘Dimensional Hypothesis’. According to H_0 participants are distributed over a multidimensional space and those with MLD are positioned at the most extreme coordinate positions along *some* dimensions of this space. That is, MLD children would not show a marked discontinuity relative to the whole sample. We call our alternative hypothesis (H_1) the ‘Core Deficit Hypothesis’. According to H_1 the MLD group would have some core deficit, that is, this group would be markedly distinct from the characteristics of the sample population. That is, MLD status could not be explained by simply being at extreme positions along some dimensions. Finally, we calculated the area under the ROC curve (Area Under the Curve: AUC) (Robin et al., 2011), to establish how much basic number processing and domain-general measures contributed to distinguish cases of MLD from controls, in terms of their diagnostic power.

Methods

Participants

The data were gathered as part of a larger research project. A subset of the data is reported in Caviola, Colling, Mammarella, and Szűcs (2020). We tested 1,303 children from 73 different classes in three different school years (Grade 2: N = 435; Grade 4: N = 408; Grade 6: N = 460). Children with incomplete data, apart from the Weber fraction (which cannot always be calculated), were excluded from any further analyses. Children with intellectual disabilities have been excluded. However, those with other neurodevelopmental disorders were not excluded, but our ethical constraints prevented us from obtaining details of their diagnoses. We selected two groups of children using the inclusion and exclusion criteria commonly adopted in previous research. For inclusion in the MLD group, we considered a cut-off corresponding to the 10th percentile (Murphy et al., 2007) of the scores for mathematics performance; exclusion criteria were scores lower than one SD below average for reading performance or fluid intelligence. The “control” group included all remaining children in the sample, with the exclusion of those with a score lower than one SD below average in mathematics or reading performance or fluid intelligence. Based on these criteria, 47 (M_{age} = 121.68, SD = 19.97; 18 M) children were classified as belonging to the MLD group, and 895 (M_{age} = 120.11, SD = 21.00; 461 M) were classified as controls. Thus, we used data from 942 children here. The study was approved by Padova University’s ethics committee on psychology research. The written informed consent of parents or guardians was obtained for all children before any data collection. The supplementary material contains further details of our sample.

Materials

A full description of the symbolic, non-symbolic magnitude comparison and working memory tasks used in this study is available in the Supplementary materials. Here we provide an abbreviated account.

Math and reading achievement

Math achievement. Ability in mathematics was assessed by means of the standardized tasks available in the AC-MT batteries (Cornoldi & Cazzola, 2004; Cornoldi, Lucangeli, & Bellina, 2012). Different components of mathematical ability are assessed by mean of different subtests. The following subtests were used for the analysis: approximate calculation (i.e., the child has to indicate the closest approximation of a calculation result among a series of alternatives); retrieving combinations and numerical facts; complex mental and written calculation; transcoding (i.e., the child has to write down in Arabic format a series of numbers that the experimenter speak aloud). The reliability was good for all subtests, as the test-retest correlation coefficients range $.70 < r_s < .79$ for grades 2 and 4, and $.72 < r_s < .83$ for grade 6.

Reading task. Reading speed was assessed by mean of standardized tasks from a clinical battery for the assessment of Developmental Dyslexia and Dysorthographia-2 (DDE-2) (Sartori, Job, & Tressoldi, 1995). Two subtests were used. The first consisted of a real word reading task, while the second subtest consisted of a pseudo-word reading task. The reading speed in syllables per second from the two subtests were used (reliability measures are $r > .77$). This choice was made because Italian is a transparent language, and thus errors are relatively less frequent.

Mathematics and reading scores were derived as factorial scores from a confirmatory factor analysis (CFA) model combining mathematics and reading measures (multigroup CFA model for MLD vs control participants was performed and it suggested both configural and metric, albeit not scalar, invariance; all details are reported in the Supplementary material–results section).

Fluid intelligence

Reasoning ability. As a measure of nonverbal fluid reasoning, we used the Cattell Culture Fair Intelligence Test (Cattell & Cattell, 1981). It consists of four series of items, each of which must be completed in a limited time. The series of items include: series completion, odd-one-out, matrices, topology. Each series presents items of increasing difficulties. The whole Cattell task includes a total of 46 multiple-choice items (α = .66).

Magnitude representation/comparison

The following tasks were administered using the Psychtoolbox for Matlab (Brainard, 1997) on a laptop computer.

Non-symbolic magnitude comparison task (ANS task). A non-symbolic magnitude comparison task was used to measure the children's approximate number system. This task was previously used by Szűcs and colleagues (Szűcs, Devine, Soltesz, Nobes, & Gabriel, 2013). In this task, the children are required to compare the numerosity of two sets of black dots presented on a white background. Specifically, they must tell which of the two sets contains a larger number of dots. To do so, they must press either one of two buttons on the keyboard (i.e., the one on the same side of the larger set). A total of 160 trials were presented in a randomized order. One of the two dot arrays contained 16 dots while the other dot array contained a smaller (i.e., 8, 10, 12, 13, or 14 dots) or a larger (i.e.,18, 20, 22, 26, or 32) number of dots. Thus, the dot patterns differed by ratios 0.5, 0.62, 0.74, 0.81 and 0.88. The trials were presented in randomized order. Moreover, in half of the trials the numerical and visual information (specifically, convex hull) was congruent, while in the other half it was not. Each trial started with a fixation cross presented for 1000 ms, followed by the stimulus. Stimuli disappeared as soon as a response was given by the child. An interstimulus interval of 500 ms was used between trials. Twenty practice trials were administered before the 160 experimental trials. In keeping with Szűcs et al. (2013), we calculated the average task accuracy, the response times, and the Weber fraction (*w*), as the final scores. Since the presentation was brief and not counting strategy was possible, it was expected that most variability would come from accuracy and the Weber fraction (w), while response times would not be much informative. The internal consistency, measured with the Cronbach's alpha calculated on the subtotals divided by ratio, was good, α = .97 for RT, α = .87 for accuracy.

Symbolic magnitude comparison (SNC) task (Dénes Szucs, Devine, Soltesz, Nobes, & Gabriel, 2014). In this task, children are presented with one Arabic digit in the middle of the computer screen, and they are required to rapidly indicate whether it is smaller or larger than 5 (by pressing a different button on the keyboard). The digits are differently classified as follows: close to 5 (digits 4 and 6, also named "distance-one condition", because they differ by only one from 5) or far from 5 (digits 1 and 9, also named "distance-four condition", because they differ by four positions from 5). During the presentation, a fixation marker is shown for 200 ms and it is followed by a 1000 ms pause. Subsequently, the digit is presented on the screen. After the response is given, or after a maximum of 3000 ms, the digit disappears from the screen. An interval of 400 ms was placed between trials. Eight practice trials were administered before the experimental trials. The latter consisted of two blocks of 40 trials each. As the final scores, we calculated: the overall task accuracy, the average reaction times (RTs) and the distance effect in both accuracy and RTs. Since the stimuli are very simple (i.e., single digits), it is expected that accuracy would approach ceiling, and that task variability would come mostly from RT data. The internal consistency, measured with the Cronbach's alpha calculated on the subtotals divided by digit, was good, $\alpha = .97$ for RT, $\alpha = .73$ for accuracy.

Working memory

We used two simple memory span tasks to assess short-term memory: a word span task that involved verbally repeating increasingly long series of words; and a matrix span task to measure spatial short-term memory, in which the children had to recall the different positions of blue cells appearing briefly on a grid shown in the middle of a screen. Two tasks were used to assess working memory, one dual verbal and the other a dual spatial task (Giofrè, Mammarella, & Cornoldi, 2013) that involved concurrently performing a primary and a secondary task, which necessitated manipulating as well as recalling the stimuli. These tasks were administered using E-Prime Professional Software 2.0 (Psychology Software Tools, Inc., Sharpsburg, PA).

Simple-span tasks. We used two short-term memory (STM) span tasks, which were originally adapted by Giofrè and colleagues (Giofrè, Mammarella, & Cornoldi, 2013). The verbal STM task consisted of word span task in which the child is required to repeat a series of words presented orally at a rate of 1 item per second. The series are presented with an increasing level of difficulty, i.e., from the shortest (2 item) to the longest one (8 items). Words must be repeated in forward order and no time limit is imposed (α = .70). The spatial STM task consisted of a matrices span task (adapted from Hornung, Brunner, Reuter, & Martin, 2011). The children must memorize and subsequently recall the positions of coloured cells that appear briefly (1 sec) in different positions on a 4 x 4 grid presented in the middle of the screen. After the presentation the children must use the mouse to select the locations corresponding to the previously seen coloured cells. The number of coloured cells in each series ranged from 2 to 8 (α = .83).

Complex-span tasks. Two dual tasks were used to measure WM (Engle, 2010). Both are characterized by the request to concurrently perform a primary and secondary task. The verbal WM task consisted of several series of word lists presented aurally. Each series included an increasing number of word lists, each of which was composed by four medium-to-high frequency words. The length of the series ranged from 2 to 6 word lists, and there were 3 sets of series for each series length (for a total of 15 series). As the primary task, the child was required to recall the last word of each list within the presented series (immediately after the completion of that series), in forward order. As the secondary task, the children were asked to press the space bar every time they heard an animal noun among the presented words. The final raw score is the number of last words correctly recalled in the right order, and it ranges from 0 to 60. This task showed adequate psychometric properties (α = .69) and good predictive power (Giofrè et al., 2013).

The spatial WM task was adapted from Mammarella and Cornoldi (2005), and it consisted once again of a matrices span task. The structure of the task resembled that of the verbal WM task, but using spatial material. The matrices were composed by a 4 x 4 grid of empty cells, and they were presented on the middle of the computer screen. In all matrices, seven cells were grey, and nine cells were white. The task consisted of a series of lists of matrices. As for the verbal WM task, the number of to-be-remembered items increased from 2 to 6, and there were 3 sets of series for each series length. The task was organised into sets of two grids in which a black dot appeared and disappeared on the grid. In each set, the children had to press the spacebar if the dot was presented on a grey cell while at the same time remembering the last position of the dot (i.e., the third position for each set). The final raw score is the number of last dot positions correctly recalled in the right order, and it ranges from 0 to 60 (α = .82).

Procedure

The children were tested at their own school, in three sessions held between the end of January and May 2018. Children were tested once in groups and twice in individual sessions. Group sessions were used for administering the Fluid intelligence task and some subtests from the Math achievement batteries (according the administration manual). The order of test administration was counterbalanced across classes. Following the group session, two individual sessions, lasting approximately 50 minutes each, were used for administering the Reading tasks, the remaining tasks of the Math batteries, and all the computerized tasks (two Magnitude comparison tasks and four Working memory tasks). Both paper-and-pencil and computerized tasks were equally divided and counterbalanced across the two individual sessions.

Data analysis

In step one, we calculated the standardized difference (Cohen's d) between the MLD and control groups for each variable of interest (accuracy in distinguishing symbolic and non-symbolic magnitudes, verbal short-term and working memory, spatial short-term and working memory). Cohen's d and its 95% Bayesian credible interval (BCI) were calculated using the MCMC sampling algorithm implemented in the “brms” package (Bürkner, 2018) in R. For each variable, we fitted a linear model with group as the predictor and the measure of interest as the response variable. The posterior distribution of Cohen’s d was calculated as the ratio between the group coefficient (equivalent to the between-group difference) and the SD of the residuals (equivalent to the pooled SD). Default non-informed priors were used, so the point estimates of Cohen’s d were exactly the same as those obtained via analytical calculation.

In step two, we estimated the same series of standardized differences under the hypothesis that the MLD group simply reflects the global characteristics of the sample population. We simulated a very large population (N = 1,000,000) that reflected the exact same set of Pearson’s correlations between all the variables measured in our sample of children. In this phase, children previously placed in the MLD group were excluded to avoid any inflated correlations (see Tables S6 and S7 in the Supplementary materials – results section - showing that this had a negligible effect on the correlations, however). We set all simulated variable distributions to reflect the same asymmetries and kurtoses as those seen in our sample. To simulate data that matched both the correlation matrix and the vectors of asymmetries and kurtoses, we used the “simulateData” function of the “lavaan” package of R (see Supplementary material for details on the efficacy of the simulation). To calculate the simulated standardized differences, we repeated the same procedure as in step one on the simulated sample, using the exact same categorizing criteria (based on cut-offs) to identify simulated MLD and “control” groups.

To assess the “dimensional” hypothesis that the characteristics of the MLD group simply reflect the characteristics of the population after selection criteria have been applied, we can consider the discrepancy between the observed and simulated Cohen’s d values. The larger the discrepancy, the more the variable is likely to be associated with MLD in a way that cannot easily be inferred from the characteristics of the general sample population.

We used a t-test-like approach to quantify how strongly the data support the “dimensional” hypothesis H_0 as opposed to the “core deficit” hypothesis H_1 (in other words, the notion that MLD is characterized by core deficits that cannot be anticipated under H_0). We adopted a procedure that resembles the “ttestBF” function of the “BayesFactor” package (Morey & Rouder, 2015) in R. For each variable of interest, we computed a Bayes factor (BF) to compare the likelihoods of the standardized difference being observed under each hypothesis. H_0 was modelled as a normal distribution centered on the simulated Cohen’s d. The standard deviation for the H_0 model was derived from a bootstrap distribution with 10,000 iterations on our observed sample (calculating Cohen’s d under H_0 at each iteration). The model for H_1 was similar to the one used in the above-mentioned “ttestBF” function, i.e., it was modelled as a Cauchy distribution with the location parameter set to the same value as for the H_0 model, and a scale parameter of $\sqrt{2}/2$. There is consequently more support for H_1 the greater the distance of the observed values from the central H_0 point. The models were fitted using the “brms” package in R and compared using its “bayes_factor” function, which enabled us to calculate both univariate BFs (i.e., on models separately fitted on each variable), and a multivariate BF for models simultaneously including all dependent variables of interest.

In addition, we calculated the AUC (Robin et al., 2011) for each variable of interest to ascertain their discriminatory/diagnostic power with regard to the MLD group as opposed to the control group. The AUC indicates how a continuous variable performs as a binary classifier, and it is especially useful in diagnostic settings. We calculated the AUC with the “auc” function of the “pROC” package in R, which uses the trapezoidal rule (as is generally the case for AUC). Using the AUC has the following advantages: it directly expresses diagnostic power; and it is robust against normality violations because it is calculated on non-parametric ROC curves.

We used the following rules to interpret the effect sizes: for Cohen’s d, any value up to around .50 (medium effect size) was interpreted as unlikely to indicate that the variable of interest represents a “core deficit” in the MLD group; a “core deficit” should coincide with a large standardized difference, e.g., a Cohen’s d >.80. Note that, under normality, a Cohen’s d of .80 corresponds to an AUC of about .71, which is still of limited use as a diagnostic classifier (it would only identify around 70% of true-positives, while it would also generate around 38% of false-positives). See Zhu *et al.*(Zhu, Zeng, & Wang, 2010) on how to interpret AUC sizes.

Results

All variables were standardized by school year, using the means and SDs calculated in our sample of children. This was done to eliminate any difference relating to level of schooling. Supplementary results reported descriptive statistics (Tables S21 to S5), correlations among variables, as well as asymmetry and kurtosis of each variable (Table S6)

Standardized differences and diagnostic power

Figure 1 shows the Cohen’s d values observed for all variables of interest, along with the simulated Cohen’s d values and the Bayes factors (BF) that quantify the likelihood of H_1 over H_0. A BF of more than 1 provides evidence to support H_1, while a BF of less than 1 is in favor of H_0. We interpreted any BF between 1/3 and 3 as indicating only “anecdotal” evidence, and BFs beyond those limits as indicating “moderate” (or stronger) evidence (Lee & Wagenmakers, 2013).

Figure 1 about here

Figure 1 shows that virtually the entire cognitive profile of standardized differences supports the Dimensional Hypothesis. In other words, practically every single standardized difference between the MLD and control group reflects the way in which the cognitive variables are related with math, reading, and fluid intelligence scores in the non-MLD population. This is evident by the fact that for almost all variables, the observed (black dots with error bars) and simulated (empty circles) Cohen's d values were remarkably similar. The BFs supported the Dimensional Hypothesis (i.e., H_0) with at least a "moderate" degree of evidence in all but three cases (and even in these cases it pointed in favor of H_0). Two of these relative exceptions were "Spatial STM" and "Spatial WM". They were even less impaired in the MLD group than predicted under the dimensional hypothesis, however, so they could not be defined as core deficits. Lastly, "Accuracy symbolic" was the only variable that may come close to deserving the definition of a core deficit (i.e., a deficit that is large, and larger than predicted under a dimensional hypothesis). "Accuracy symbolic" was a strongly skewed variable, however, with a clear ceiling effect, so in principle it may not be appropriate as a psychometric measure because it is biased, measuring differences only in the very lowest portion of the distribution. The evidence was nonetheless against H_1 even for this variable.

In short, none of the variables considered revealed large standardized differences between the two groups (Cohen's d was always below .70), and none of these differences deviated from those predicted under a purely Dimensional Hypothesis (ΔCohen's d never exceeded .25). The multivariate BF was < .0001, indicating extremely strong evidence in favor of H_0 at the multivariate level.

Figure 2 shows the AUCs for all the variables measured. The AUC is reported, both as observed and as predicted under the dimensional hypothesis. All AUCs are below .70, indicating a weak diagnostic power for all variables. There were also no major discrepancies between the observed and simulated AUCs, again supporting a dimensional interpretation.

Figure 2 about here

Discussion

In this study we aimed to examine whether differences between typically-developing children and those with MLD (identified by means of psychometric cut-offs without a clinical diagnosis) reflect global characteristics of the population considered as a whole (Dimensional Hypothesis), or whether the MLD group is characterized by core deficits that cannot be inferred from the global parameters that describe the rest of the population (Core Deficit Hypothesis). We also aimed to test the diagnostic power of basic number processing (i.e., non-symbolic and symbolic magnitude comparison tasks) and domain-general measures (i.e., verbal and visuospatial short-term and working memory) in the diagnosis of MLD.

After selecting cases of MLD from a large sample of children by using psychometric cut-offs and employing widely accepted criteria (Murphy et al., 2007), we simulated a very large population that reflected the characteristics of the remaining sample. We computed Cohen’s d expressing the difference between the MLD and “control” groups for a series of variables of interest, for both the observed sample and the simulated population. Our findings suggested that none of the measures of basic number processing or domain-general abilities could identify “core deficits” in our children with MLD. Rather, all differences between the MLD and control groups were more likely to reflect the global characteristics of the population sampled (see also Peters & Ansari, 2019 for an opinion paper). As in previous studies, the discrepancy between the MLD and the control group was the largest in symbolic magnitude comparison accuracy (De Smedt et al., 2013), and in working memory measures, especially in verbal working memory (Peng et al., 2018).

To test the diagnostic power of basic number processing (i.e., non-symbolic and symbolic magnitude comparison tasks) and domain-general measures (i.e., verbal and visuospatial short-term and working memory), we used the AUC (Robin et al., 2011) as a binary classifier to see how much each measure of interest enabled us to distinguish cases of MLD from controls. Our findings again suggested that none of the measures of interest were good classifiers of children with MLD (and none substantially exceeded the diagnostic power that could be inferred by means of a simulation from the general population). In fact, none of the basic number processing measures (i.e., non-symbolic and symbolic magnitude comparison tasks) considered in our study, previously linked to putative core deficits in children with MLD (Butterworth, 2011; De Smedt et al., 2013; Mussolin et al., 2010), exceeded an AUC of .70, indicating weak classification performance (Zhu et al., 2010).

Our findings seem to suggest that looking for a "core deficit" in children with MLD is simplistic (Astle, & Fletcher-Watson, 2020). Individuals with MLD may have deficits in both basic number processing and in domain-general cognitive skills, but neither of these are necessarily present. From another perspective, math performance may correlate differently with various basic number processing and domain-general measures in the general population, and the cognitive profile seen in MLD may simply reflect this set of correlations after specific psychometric selection criteria have been applied. This is the clear message emerging from our results. Hence, rather than comparing artificially created subgroups it may be more fruitful to position children in a multi-dimensional measurement space including both domain-general and domain-specific measures (Szűcs, 2016). According to this view MLD children would occupy different moderate to extreme positions in this smoothly changing (without abrupt changes predicted by core deficit models) multi-dimensional space. Different cognitive impairments in children with MLD could then vary as a function of comorbidity, severity of the disorder, age, and possibly other factors too, which need to be investigated more thoroughly in further studies (Peng et al., 2018).

A limitation of our study is that we tested a large sample of children and used widely accepted criteria (psychometric cut-offs only) to demonstrate that groups of children selected in this way reflected the characteristics of the whole population sampled. However, ideally, we should have also tested a large group of children already with a clinical diagnosis of MLD, using the same basic number processing and domain-general measures. Including such a group would have allowed us to conclude with greater certainty about the presence or absence of core deficits in children with persistent mathematical deficits. Thus, based on our findings we want to enforce the importance of study MLD more in depth. Further studies should apply our simulation approach to confirm our findings in children with a clinical diagnosis of MLD. Another limitation is that we cannot exclude that the MLD population may include sub-populations characterized by different, independent deficits. This would violate dimensionality in a way that cannot be detected by our analyses. Future research could investigate this hypothesis using cluster and latent profile analysis. Correctly detecting the number of latent classes in LPA requires no less than several hundred observations, regardless of the assumptions or method used, even with between-class separations as large as Cohen's d = .8 (Tein et al., 2013).

In summary, our study mainly demonstrated that the psychometric cut-off procedure commonly adopted to identify MLD children in principle leads to the same conclusions than just considering the whole population and examining the properties of children at the extreme positions of diagnostic variables. In fact, selecting MLD groups in the common way provides inferior results because of the overestimation of the effect sizes due to the typically low sample sizes (Szucs & Ioannidis, 2017). It is notable that although, the DSM-5 (American Psychiatric Association, 2013) suggests criteria for identifying children with MLD, previous research studies were highly inconsistent in selecting children with MLD. This is unfortunate given that ideally researchers and clinicians should investigate the same populations, in order to reach conclusive findings. A substantial difference between children diagnosed by psychometric cut-offs and those diagnosed by clinical diagnosis is the presence of significant deficits interfering with both academic and daily-life activities in the latter. Moreover, according to the DSM 5, these deficits should be persistent (see also Mazzocco & Mayers, 2003). In other words, obtaining weak math performance in a single assessment does not necessarily mean that a child really shows a specific learning disorder. Notably, the clinical diagnosis of MLD requires that children show sustained (at least 6-month long) and treatment-resistant MLD. For this reason, the criterion of the persistence of deficits in math performances should be considered in selecting children with MLD, in order to advance research in this field.

Thus, our findings offer a new starting point for future research in the field of MLD. First, the criteria offered by the international diagnostic manuals should be used for identifying children with MLD. It is particularly important to consider the persistence of impairments despite tailored interventions. A dimensional approach should also be used to study children with MLD, considering the presence of comorbidities within specific learning disorders or with other neurodevelopmental disorders. Risks and protective factors, and changes in symptoms over the lifetime should also be investigated. Though more expensive in terms of time and resources, studies with larger samples should be conducted to increase statistical power, to produce more precise population-level effect size estimates, to lower the risk of overestimating effect sizes and to avoid detecting falsely statistically significant findings (Gelman & Carlin, 2014; Szucs & Ioannidis, 2017).

In conclusion, the present study suggests that using only psychometric cut-offs to distinguish groups of children with MLD from typically developing children is a suboptimal approach that is unlikely to further advance our knowledge about MLD. Diagnosing MLD groups this way may simply result in MLD groups reflecting the structure of cognitive abilities (i.e., the overall set of correlations among cognitive variables) observed in the typically developing population. Future research should focus either on typical populations or on samples of MLD with a more comprehensive clinical diagnosis, following shared and recognized criteria, such those suggested by the international diagnostic manuals.

Footnotes

1 - From this point on, we use the term MLD to refer children with specific learning disorders with impairments in mathematics, without distinguishing between MLD and developmental dyscalculia.

2 - It is worth noting that previous research based on the assumption of general cognitive deficits in MLD never expected to find a core deficit. That is why we considered some general cognitive measures, such as short-term and working memory, for comparison in the present study.

References

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. In *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*. https://doi.org/10.1176/appi.books.9780890425596.744053

Astle, D. E., & Fletcher-Watson, S. (2020). Beyond the Core-Deficit Hypothesis in Developmental Disorders. *Current Directions in Psychological Science, 29(5),* 431–437. https://doi.org/10.1177/0963721420925518

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411.

Butterworth, B. (1999). *The Mathematical Brain*. London UK: Macmillan.

Butterworth, B. (2011). Foundational numerical capacities and the origins of dyscalculia. In *Space, Time and Number in the Brain* (pp. 249–265). https://doi.org/10.1016/B978-0-12-385948-8.00016-5

Cantlon, J. F., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16507–16511. https://doi.org/10.1073/pnas.0506463102

Cattell, R. B., & Cattell, A. K. S. (1981). *Measuring intelligence with the culture fair tests. Institute for Personality and Ability Testing [Italian edition: Misurare l'intelligenza con i test "Culture Fair"]*. Florence, Italy: Giunti OS.

Caviola, S., Colling, L., Mammarella, I., & Szűcs, D. (2020). Predictors of mathematics in primary school: magnitude comparison, verbal and spatial working memory measures. *Developmental Science*. https://doi.org/10.1111/desc.12957

Cornoldi, C., & Cazzola, C. (2004). *AC-MT 11-14: Test for Assessing Calculation and Problem*

Solving Skills [“Test AC-MT 11-14 - Test di Valutazione delle Abilità di Calcolo e Problem Solving”]. Trento, Italy: Erickson.

Cornoldi, C., Lucangeli, D., & Bellina, M. (2012). *AC-MT 6-11: Test for Assessing Calculation and Problem Solving Skills [“Test AC-MT 6-11 - Test di Valutazione delle Abilità di Calcolo e Problem Solving”]*. Trento, Italy: Erickson.

De Smedt, B., Noël, M. P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children’s mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48–55. https://doi.org/10.1016/j.tine.2013.06.001

Dehaene, S. (2011). The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition. In *The number sense How the mind creates mathematics rev and updated ed*.

Devine, A., Soltesz, F., Nobes, A., Goswami, U., & Szucs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction*, *27*, 31–39. https://doi.org/10.1016/j.learninstruc.2013.02.004

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15. https://doi.org/10.1177/00222194040370010201

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology*, *77*(3), 236–263. https://doi.org/10.1006/jecp.2000.2561

Gelman, A., & Carlin, J. (2014). Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Giofrè, D., Mammarella, I. C., & Cornoldi, C. (2013). The structure of working memory and how it relates to intelligence in children. *Intelligence*, *41*(5), 396–406. https://doi.org/10.1016/j.intell.2013.06.006

Hauser, M. D., Macneilage, P., & Ware, M. (1996). Psychology numerical representations in primates. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 1514–1517. https://doi.org/10.1073/pnas.93.4.1514

Hauser, M. D., Tsao, F., Garcia, P., & Spelke, E. S. (2003). Evolutionary foundations of number: Spontaneous representation of numerical magnitudes by cotton-top tamarins. *Proceedings of the Royal Society B: Biological Sciences*, *270*, 1441–1446. https://doi.org/10.1098/rspb.2003.2414

Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PLoS Biology*, *6*(2), e11. https://doi.org/10.1371/journal.pbio.0060011

Lee, M. D., & Wagenmakers, E. J. (2013). Bayesian cognitive modeling: A practical course. In *Bayesian Cognitive Modeling: A Practical Course*. https://doi.org/10.1017/CBO9781139087759

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development*, *82*(4), 1224–1237. https://doi.org/10.1111/j.1467-8624.2011.01608.x

Mazzocco, M. M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school-age years. *Annals of Dyslexia*, *53*, 218-253.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: computation of Bayes Factors for common designs. R package version 0.9.12-2. *BayesFactor: Computation of Bayes Factors for Common Designs*. https://doi.org/10.1007/978-3-642-21037-2

Murphy, M. M., Mazzocco, M. M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with Mathematics Learning Disability (MLD) vary as a function of the cutoff criterion used to define MLD. *Journal of Learning Disabilities*, *40*(5), 458–478. https://doi.org/10.1177/00222194070400050901

Mussolin, C., Mejias, S., & Noël, M. P. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition*, *115*(1), 10–25. https://doi.org/10.1016/j.cognition.2009.10.006

Passolunghi, M. C., & Siegel, L. S. (2004). Working memory and access to numerical information in children with disability in mathematics. *Journal of Experimental Child Psychology*, *88*(4), 348–367. https://doi.org/10.1016/j.jecp.2004.04.002

Peng, P., Wang, C., & Namkung, J. (2018). Understanding the cognition related to mathematics difficulties: A meta-analysis on the cognitive deficit profiles and the bottleneck theory. *Review of Educational Research*, *88*(3), 434–476. https://doi.org/10.3102/0034654317753350

Peters, L., & Ansari, D. (2019). Are specific learning disorders truly specific, and are they disorders? *Trends in Neuroscience and Education*, *17*. https://doi.org/10.1016/j.tine.2019.100115

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., … Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41. https://doi.org/10.1016/j.cognition.2010.03.012

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. https://doi.org/10.1016/j.neuron.2004.10.014

Price, G., & Ansari, D. (2013). Dyscalculia: characteristics, causes, and treatments. *Numeracy*, *6*(1), Article 2. https://doi.org/10.5038/1936-4660.6.1.2

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, Article 77. https://doi.org/10.1186/1471-2105-12-77

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2).

Rubinsten, O., & Henik, A. (2009). Developmental dyscalculia: heterogeneity might not mean different mechanisms. *Trends in Cognitive Sciences*, *13*(2), 92–99. https://doi.org/10.1016/j.tics.2008.11.002

Sartori, G., Job, R., & Tressoldi, P. (1995). *Battery for the assessment of dyslexia and developmental dysorthography [“Batteria per la valutazione della dislessia edella disortografia evolutiva”]*. Florence, Italy: Organizzazioni Speciali.

Schwenk, C., Sasanguie, D., Kuhn, J. T., Kempe, S., Doebler, P., & Holling, H. (2017). (Non-)symbolic magnitude processing in children with mathematical difficulties: a meta-analysis. *Research in Developmental Disabilities*, *64*, 152–167. https://doi.org/10.1016/j.ridd.2017.03.003

Shalev, R. S. (1997). Neuropsychological aspects of developmental dyscalculia. *Mathematical Cognition*, *3*(2), 105–120. https://doi.org/10.1080/135467997387434

Szűcs, D. (2016). Subtypes and comorbidity in mathematical learning disabilities: Multidimensional study of verbal and visual memory processes is key to understanding. *Progress in Brain Research*, *227*, 277–304. https://doi.org/10.1016/bs.pbr.2016.04.027

Szucs, Denes, Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex*, *49*(10), 2674–2688. https://doi.org/10.1016/j.cortex.2013.06.007

Szucs, Dénes, Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2014). Cognitive components of a mathematical processing network in 9-year-old children. *Developmental Science*, *17*(4), 506–524. https://doi.org/10.1111/desc.12144

Szucs, Denes, & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3). https://doi.org/10.1371/journal.pbio.2000797

Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 640-657. https://doi.org/10.1080/10705511.2013.824781

Temple, C. M., & Sherwood, S. (2002). Representation and retrieval of arithmetical facts: Developmental difficulties. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *55*(3), 733–752. https://doi.org/10.1080/02724980143000550

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *Northeast SAS Users Group 2010: Health Care and Life Sciences*.

Table 1

Summary of the characteristics of MLD samples in the studies included in two recent meta-analyses (Peng et al., 2018; Schwenk et al., 2017)

	MLD samples with previous clinical diagnosis	MLD samples with no previous clinical diagnosis
Number of studies	11/90 (12%)	79/90 (88%)
Testing the core deficit hypothesis	2/11 (18%)	18/79 (23%)
Testing the domain general hypothesis	9/11 (82%)	61/79 (77%)
Number of participants		
N ≤ 20	4/11 (36%)	31/79 (39 %)
N ≥ 21, < 40	5/11 (46%)	21/79 (27%)
N ≥ 40	2/11 (18%)	27/79 (34%)
Math criteria for selecting MLD groups		
Math scores ≤ 10^{th} percentile (or 1.5 SD)	4 /11 (37%)	14/79 (18%)
Math scores ≤ 15^{th} percentile (or 1 SD)	0/11	20/79 (25%)
Math scores ≤ 25^{th} percentile	5/11 (45%)	29/79 (37%)
Other criteria	2/11 (18%)	16/79 (20%)
Other abilities controlled for		
Reading skills	10/11 (90%)	59/79 (75%)
IQ	11/11 (100%)	55/79 (69%)

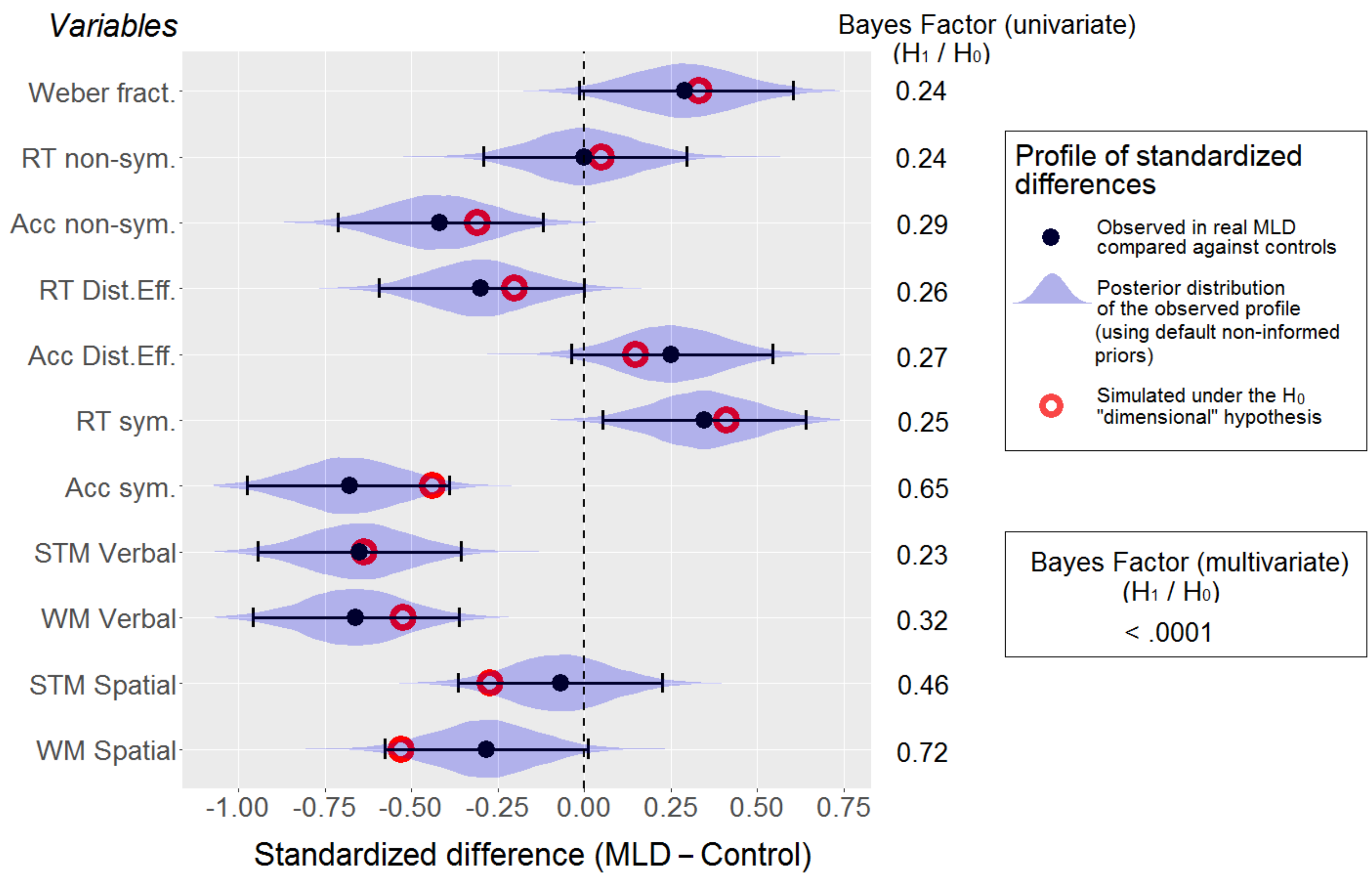

Figure 1. Observed Cohen's d (black dots) between children with MLD and controls, for all variables of interest. Error bars represent 95% BCI of the observed Cohen's d. Violin plots represent their posterior distributions (using default priors). Empty circles represent Cohen's d simulated under

the dimensional hypothesis. Bayes factors (BF) approximate the relative likelihood of the data under the alternative (core deficit) hypothesis over the null (dimensional) hypothesis, both when analyzed separately for each variable (univariate BF) and in the multivariate model.

Note. **Non-symbolic magnitude comparison task**: Weber fract. = Weber fraction; RT non-sym. = response times in non-symbolic magnitude comparison task; Acc. Non-sym. = task accuracy. **Symbolic magnitude comparison task:** RT Dist. Eff. = distance effect in response times; Acc. Dist. Eff. = distance effect in accuracy; RT sym. =average response times in the symbolic magnitude comparison task; Acc. Sym. = proportion of correct responses in the symbolic magnitude comparison task. **Working memory tasks:** STM verbal = accuracy in the verbal short-term memory task; WM verbal = accuracy in the verbal working memory task; STM spatial = accuracy in the spatial short-term memory task; WM spatial = accuracy in the spatial working memory task.

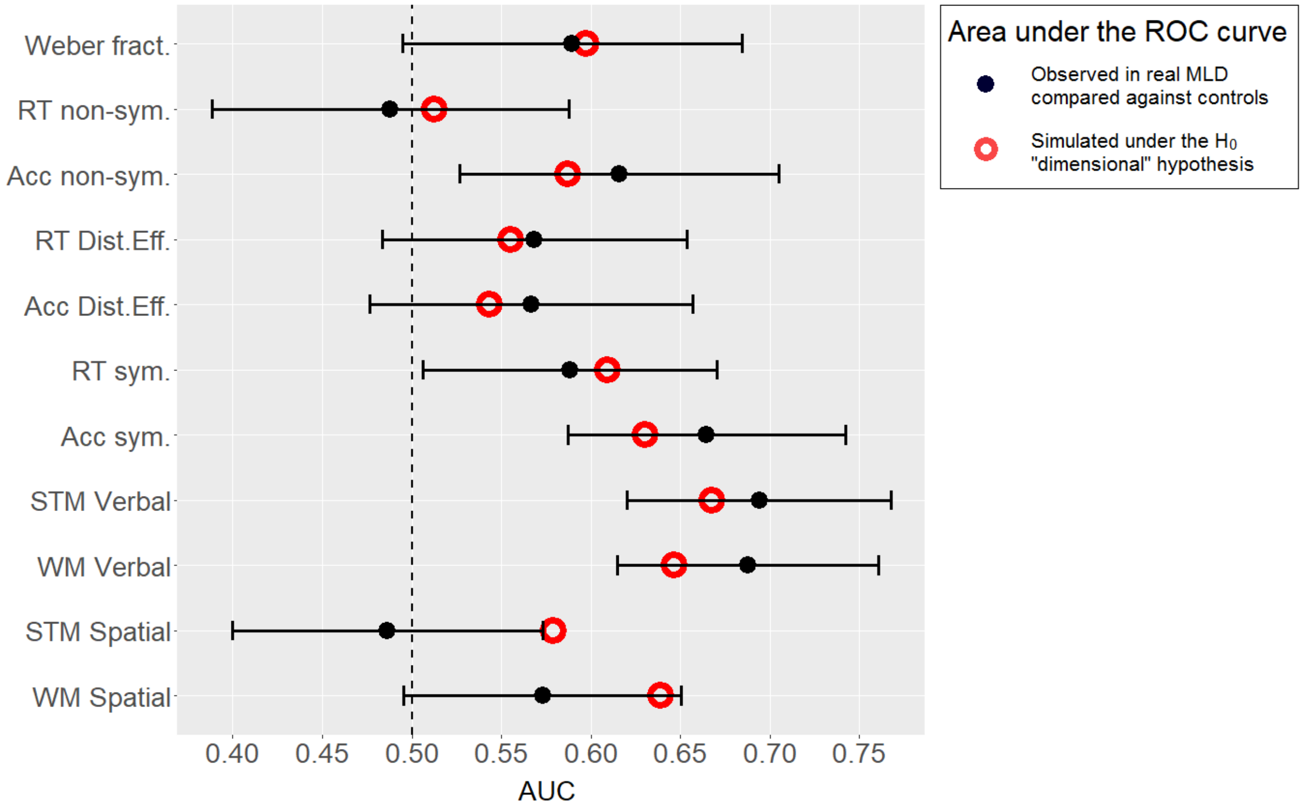

Figure 2. Black dots represent the observed area under the ROC curve (AUC), indicating the diagnostic power of each variable of interest as a binary classifier of group (MLD vs control). Error bars represent 95% confidence intervals computed with 10,000 stratified bootstrap replicates. Empty circles represent the same AUC simulated under the dimensional hypothesis. Chance level is AUC = .50.

Note. **Non-symbolic magnitude comparison task**: Weber fract. = Weber fraction; RT non-sym. = response times of non-symbolic magnitude comparison task; Acc. Non-sym. = task accuracy. **Symbolic magnitude comparison task:** RT Dist. Eff. = distance effect in response time; Acc. Dist. Eff. = distance effect in accuracy; RT sym. =average response times in the symbolic magnitude comparison task; Acc. Sym. = proportion of correct responses in the symbolic magnitude comparison task. **Working memory tasks:** STM verbal = accuracy in the verbal short-term memory task; WM verbal = accuracy in the verbal working memory task; STM spatial = accuracy in the spatial short-term memory task; WM spatial = accuracy in the spatial working memory task.