
Article:
Xu, C. orcid.org/0000-0002-6702-3958, Di Lonardo Burr, S., LeFevre, J.-A. et al. (8 more authors) (2023) Development of children's number line estimation in primary school: regional and curricular influences. Cognitive Development, 67. 101355. ISSN 0885-2014

https://doi.org/10.1016/j.cogdev.2023.101355

Cognitive Development 67 (2023) 101355

Contents lists available at ScienceDirect

Cognitive Development

journal homepage: www.elsevier.com/locate/cogdev

Development of children's number line estimation in primary school: Regional and curricular influences

Chang Xu [a,b,*], Sabrina Di Lonardo Burr [c], Jo-Anne LeFevre [b,d], Sheri-Lynn Skwarchuk [e], Helena P. Osana [f], Erin A. Maloney [g], Judith Wylie [a], Victoria Simms [h], María Inés Susperreguy [i,j], Heather Douglas [d], Anne Lafay [k]

[a] School of Psychology, Queen's University Belfast, UK
[b] Department of Psychology, Carleton University, Canada
[c] Department of Psychology, University of British Columbia, Canada
[d] Department of Cognitive Science, Carleton University, Canada
[e] Faculty of Education, University of Winnipeg, Canada
[f] Department of Education, Concordia University, Canada
[g] School of Psychology, University of Ottawa, Canada
[h] School of Psychology, Ulster University, UK
[i] Faculty of Education, Pontificia Universidad Católica de Chile, Chile
[j] Millennium Nucleus for the Study of the Development of Early Math Skills (MEMAT), Chile
[k] Department of Psychology, Université Savoie Mont Blanc, France

ARTICLE INFO

Keywords:
Number line estimation
Mathematics
Students
Primary school children

ABSTRACT

Is the development of number line estimation (NLE) similar across regions? Data from Canada (Quebec, $n = 67$, $M_{age} = 7.9$ years; Manitoba, $n = 177$, $M_{age} = 7.8$ years), Chile ($n = 81$, $M_{age} = 7.9$ years), and Northern Ireland ($n = 171$, $M_{age} = 7.3$ years) were analyzed. Twice, approximately one year apart, students completed a 0–1000 NLE task and other mathematical tasks. Using latent profile analysis, students' estimates were classified as belonging to either a *uniform* or *variable* profile. At Time 1, estimation accuracy differed across regions, but at Time 2, patterns of performance were similar. Regional variations in improvements were related to curricular demands. Moreover, mini meta-analyses of the associations between NLE and other mathematical tasks revealed medium effect sizes. Overall, the NLE task can provide insights into concurrent and longitudinal mathematics achievement, but educational experiences should be considered when comparing performance across regions.

Differences in mathematics competence across countries and regions are often indexed with performance on international mathematics assessments (Mullis et al., 2020). However, these assessments only provide information about group-level differences at a single time point. Moreover, longitudinal research provides information about the development of students' mathematical skill but not all tasks are suitable for students at similar educational stages in different countries because of variations in curriculum expectations and spoken languages. Accordingly, comparing mathematical development over time across countries and educational systems is challenging. To overcome some of these challenges, we recruited students from four regions (three countries), with variability in their years of educational experience, language of instruction, and curricula, and asked them to complete an identical number line

* Correspondence to: School of Psychology, Queen's University Belfast, 18-30 Malone Road, Belfast BT9 5BN, UK.
E-mail address: c.xu@qub.ac.uk (C. Xu).

https://doi.org/10.1016/j.cogdev.2023.101355
Received 1 January 2023; Received in revised form 16 May 2023; Accepted 12 June 2023
Available online 20 June 2023

estimation task. Both across and within regions, we assessed several factors known to relate to number line performance and development, such as age, gender, and socioeconomic status. Our goal was to determine whether the number line task provides useful information about the development of mathematical skills despite levels of diversity that are inevitable in cross-cultural comparisons.

Why focus on number line performance? Performance on the number line estimation task captures important aspects of students' number-system knowledge, such as children's familiarity with numbers, understanding of numerical magnitude and proportional reasoning skills (Barth & Paladino, 2011; Gunderson et al., 2012; LeFevre et al., 2013; Muldoon et al., 2013; Slusser & Barth, 2017), and is strongly associated with current and future mathematics achievement (Nuraydin et al., 2023; Schneider et al., 2018; Siegler & Braithwaite, 2017). In a meta-analysis of over 10,000 participants, Schneider et al. (2018) reported an average correlation of $|r| = .44$ between mathematical competence and number line performance. The correlation was stable across different scoring methods, different measures of mathematical competence, and different versions of the number line task. This stability suggests that the number line task is a good candidate for comparing mathematical development among students who vary with respect to factors that are difficult to equate across groups, such as educational experience, language, and cultural factors.

1. Potential factors to consider in cross-cultural number line estimation research

1.1. Age-related differences in number line estimation

There are many variations of the number line estimation task. However, most researchers have used the classic number-to-position version where students estimate the location of a target number on a horizontal line (Siegler & Opfer, 2003). The line typically has a base-10 scale, with "0" on the left end and "10," "100," or "1000" on the right end. Age-relevant ranges for the number line include 0–10 for preschoolers (Whyte & Bull, 2008; Xu & LeFevre, 2016; Xu, Di Lonardo Burr et al., 2021); 0–100 (Ashcraft & Moore, 2012; Bouwmeester & Verkoeijen, 2012; Dietrich et al., 2016; Geary et al., 2008) and 0–1000 for elementary school students (Ashcraft & Moore, 2012; Gunderson et al., 2012; Laski & Yu, 2014; LeFevre et al., 2013; Siegler & Opfer, 2003; Slusser et al., 2013); and 0–10,000 for middle school (Booth & Newton, 2012), secondary school (Jung et al., 2020), and postsecondary students (Di Lonardo et al., 2020). Children's estimates become more precise as they master number knowledge in the appropriate range (Friso-van den Bos et al., 2015; Gunderson et al., 2012; LeFevre et al., 2013; Muldoon et al., 2013; Praet & Desoete, 2014; Xu, Di Lonardo Burr et al., 2021).

Beyond knowing that children's estimates improve with age, it is important to know *how* they improve. Improvements may reflect a shift in children's mental representation of magnitude from logarithmic to linear (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003) or the selection of more effective strategies, such as the use of appropriate reference points (Barth & Paladino, 2011; Huber et al., 2014; Link et al., 2014; Peeters et al., 2016; Slusser & Barth, 2017). Sometimes trials are averaged such that participants have a single accuracy score associated with their performance, but because accuracy can vary within participants and across trials, analyses that consider individual trial performance can provide further insights into patterns of estimation.

One approach to investigate patterns of estimation is to use latent profile analysis. Latent profile analysis, which is a type of latent variable analysis, is based on the assumption that within a dataset there is a mixture of observations from several mutually exclusive profiles (Lanza & Cooper, 2016). In the case of number line estimation, our mixture of observations is from the various patterns of estimation exhibited by each student. The analysis allows for the identification of groups of students that show similar performance, or patterns, on the latent variables (Oberski, 2016) In contrast to explanations of performance that are based on a logarithmic-to-linear shift model of development, latent variable analysis allows researchers to investigate how students transition from one profile of estimation to another, even when those profiles do not fit linear and logarithmic functions. Latent variable analysis has been used in previous developmental number line research to group participants by the patterns of performance on the number line task and observe changes in these patterns over time, without needing to commit to a particular theoretical perspective before data are analyzed (e.g., Bouwmeester & Verkoeijen, 2012; Xu, 2019; Xu, Di Lonardo Burr et al., 2021). In the present study, we used both this person-centered approach and more typical variable-centered analyses because the two approaches complement each other.

1.2. Educational experiences, language, and number line performance

Educational experiences, such as the age at which children begin compulsory schooling, resources available in schools, and the curriculum (i.e., play-based versus academically-oriented kindergarten programs) vary across and within countries. Family socioeconomic status (SES) is also related to educational experiences (i.e., public versus private schools). Such factors are related to children's number line skills (Ramani & Siegler, 2008; Tikhomirova et al., 2022; Xu, Di Lonardo Burr et al., 2021; Xu et al., 2013).

Torbeyns et al. (2015) compared fraction number line performance for middle-school students (i.e., Grades 6 and 8) in three countries: Belgium ($M_{age} = 11.2$ years and 13.3 years), the U.S. ($M_{age} = 11.4$ years and 13.2 years), and China ($M_{age} = 12.3$ years and 14.3 years). Chinese and Belgian students were more accurate than U.S. students on a 0–1 fraction number line. On a 0–5 fraction number line, Chinese students were more accurate than Belgian students, who were more accurate than U.S. students. However, correlations between number line performance and arithmetic skills were significant in all three countries ($|rs|$ ranging from to.23 for Grade 8 Chinese students to.44 for Grade 6 Belgian students). The overall differences in accuracy of number line estimation for students in these three countries were attributed to differences in educational experiences (i.e., mathematics curricula, teaching training, beliefs about mathematics, and time spent on mathematics), whereas the similar correlations between number line and arithmetic skills were used to support the theory that fraction and whole number knowledge are integrated in development across countries (Torbeyns et al., 2015; Siegler et al., 2011).

Language of instruction is often confounded with educational experience and thus needs to be considered in studies of number line

performance. To separate these influences, Laski and Yu (2014) examined whether knowing Chinese, a language with systematic number word construction, would support students' learning of number line estimation. Students in kindergarten and Grade 2 from two countries – Chinese students in China (M_{age} = 6.1 and 7.9 years) and Chinese-American students attending a bilingual Chinese-English public school in the U.S. (M_{age} = 6.0 and 8.1 years) – completed 0–100 and 0–1000 number line tasks and two-digit, two-addend addition problems. In kindergarten, the Chinese students performed better than the Chinese-American students on both number line and complex addition tasks. In Grade 2, the differences between the groups increased. Overall, Chinese-American and Chinese students were approximately one and two years more advanced, respectively, than non-Chinese-American students in other similar studies (Laski & Yu, 2014). Laski and Yu concluded that language may support students' acquisition of mathematical skills, but because the Chinese-American students were less advanced than the Chinese students, educational experiences are likely also important.

Xu and LeFevre (2018) compared the number line performance of 94 3- to 5-year-old Canadian-born children, half of whose parents had immigrated from China and spoke Chinese. Despite equivalent educational experiences, the Chinese-Canadian children performed better on a range of numeracy measures (e.g., verbal counting, number identification, numeration, and nonsymbolic exact arithmetic) and were better able to effectively use the midpoint on a 0–10 number line. Similarly, Siegler and Mu (2008) found that 29 Chinese students in kindergarten (M_{age} = 5.7 years) performed more accurately on a 0–100 number line than 24 American students (M_{age} = 5.6 years). However, Chinese students outperformed American students on an addition task as well, suggesting that cross-country differences might simply reflect overall mathematical competence that is closely tied to students' educational experiences. These early differences in numeracy skills may reflect not only language differences, but also cultural differences about the importance of early mathematical competence.

In another approach to comparing language effects on number line performance, Helmreich et al. (2011) compared the performance of German- (M_{age} = 7.3 years) and Italian-speaking (M_{age} = 6.9 years) Grade 1 students on a 0–100 number line. The German language uses inversion in two-digit numbers (e.g., four-and-twenty) whereas the Italian language does not (e.g., twenty-four). They found that the German-speaking students made less accurate estimates than Italian-speaking students, especially for numbers where an inversion error would lead to a large estimation error (e.g., placing 82 at 28) compared to numbers where an inversion error would have less impact (e.g., placing 54 at 45). These findings indicate that number line performance may vary with specific language features.

1.3. Gender differences in number line performance

Findings surrounding gender differences in mathematics achievement have been mixed. For example, some studies find no gender differences in semantic number magnitude processing tasks, such as the number line estimation task (Bakker et al., 2019; Rosselli et al., 2009; Zhang et al., 2020). In contrast, other researchers have found gender differences on the number line task, with boys more accurately placing target numbers than girls (Bull et al., 2013; Gunderson et al., 2012; Reinert et al., 2017; Rivers et al., 2021; Thompson & Opfer, 2008; Tian et al., 2022). In a recent study investigating gender differences in children's basic numerical skills, Hutchison et al. (2019) find that although gender differences are the exception, not the rule, number line estimation tasks were the only basic numerical task in which boys showed an advantage over girls.

One explanation for this advantage is related to spatial processing. Gender differences favouring boys have been consistently found in spatial processing (e.g., Halpern et al., 2007; Levine et al., 2016; Voyer et al., 1995). For example, Tian et al. (2022) found that for

Table 1
Demographic Factors and Educational Experience by Region.

	Chile	Quebec	Manitoba	Northern Ireland
N at T1 / T2	87 / 77	81 / 49	182 / 207	180 / 175
Age range (years:months)	6:10–10:5	7:2–10:10	7:2–10:10	6:9–9:1
M_{age} at T1 / T2	7:11 / 9:0	7:11 / 8:11	7:10 / 8:10	7:4 / 8:5
% Boys at T1 / T2	54 / 52	43 / 35	44 / 43	43 / 44
Testing Month at T1 / T2[a]	8 / 8	7 / 7	9 / 9	8 / 7
Years of Preschool/Kindergarten	2	1	1	1
Years of Formal School at T1	2	2	2	2
School SES[b,c]	60% Low 40% High	70% Low 30% High	Working to Middle class	Working to Middle Class
First/Home Language	100% Spanish	62% French	90% English	93% English
Language of Mathematics Instruction	100% Spanish	100% French	38% English 62% French (Immersion)	55% English 45% Irish (Immersion)
Curricular Expectations at T1	Numbers to 100[d]	Numbers to 1000	Numbers to 100	Numbers to 100

[a] Testing month is the median number of months between the start of the school year and when testing took place. The school year starts in September and ends in June in Canada and Northern Ireland; the school year starts in March and ends in December in Chile.

[b] For Quebec, schools were categorized based on a socio-economic background index (Indices de défavorisation). For Chile, schools were categorized based on a vulnerability index for schools (Índice de Vulnerabilidad Escolar, IVE, Agencia de Calidad de la Educación (2015).

[c] In Quebec, Manitoba, and Northern Ireland, all students were in public government-funded schools. In Chile, low-SES students were in government-subsidized schools and high-SES students were in unsubsidized schools.

[d] Count to 1000, but all other number knowledge to 100.

students from kindergarten through Grade 4, gender differences were mediated by students' spatial skills (i.e., proportional reasoning and mental rotation). Tian et al. speculated that the advantage in number line estimation for boys may reflect a more accurate mental representation of numbers or the use of different strategies. In a meta-analysis, Rivers et al. (2021) reported that, after controlling for estimation accuracy, boys were more confident in their trial-by-trial judgements on the number line task than girls. They speculated that girls' lower confidence may be related to their lower self-efficacy in either their perceived math abilities or spatial abilities, or both. Finally, beyond number line estimation, culture may influence gender differences in mathematics, such that the gender gap may be greater in nations with greater gender inequality (Guiso et al., 2008; Hyde & Mertz, 2009). Together, these studies underscore the importance of considering gender in cross-cultural number line estimation research.

In summary, various factors should be considered in comparisons of number line performance across diverse groups. The number line task may be a useful index of mathematical skill, but more work is needed to understand how factors such as age, educational experience, language, and gender are linked to differences in number line performance.

2. The present study

To understand how the factors discussed above are linked to differences in number line performance, we examined and compared patterns of development on the number line task over the course of one year for students from four regions in three countries: the provinces of Quebec and Manitoba in Canada, Chile, and Northern Ireland. These regions were selected because they allowed us to examine differences in age, gender, educational experiences, and language of instruction within and across regions (see Table 1 for details). Notably, these four regions are not the only geographic regions in the world that could provide insights into how various factors may differ within and across cultures. However, because these regions differ in language of instruction, curriculum expectations, and age at which students begin formal schooling, they provide a useful starting point for considering cross-cultural differences in number line estimation. All students completed the same 0–1000 number line task at two time points approximately one year apart. Additionally, students completed a variety of other numeracy and mathematics tasks (i.e., number comparison, number transcoding, mathematics word-problem solving, and arithmetic fluency) that are correlated with number line performance (Helmreich et al., 2011; Schneider et al., 2018).

2.1. Research question 1: Are the patterns of development of number line estimation similar across regions?

We expected students from all four regions to show improvements in their estimation skills from Time 1 to Time 2 (i.e., after one year). To test for quantitative differences (i.e., mean percentage absolute error; PAE) and qualitative differences (i.e., patterns of estimates) in performance across regions and over time, we used ANOVA and latent profile analysis, respectively. Based on previous work by Xu (2019) with primary school students and based on the curriculum for these students (i.e., most students were still learning the numbers to 1000), we expected that two profiles would emerge: (a) a *uniform* profile, where students would have low PAE across the targets, and (b) a *variable* profile, where students would have higher PAEs for smaller target numbers.

Within each region, we also explored the demographic variables (age in months, gender, socioeconomic status, and whether students were enrolled in immersion programs) and mathematical skills (number comparison, transcoding, arithmetic fluency, and word-problem solving) that differentiated students in the two profiles. Moreover, we explored which of these variables were related to students transitioning from, or remaining within, the variable profile. Based on the literature, we expected that older students, boys, students who attended high-SES schools, and students with stronger mathematical skills would be more likely to be in the uniform profile or to be transitioning from the variable to uniform profile. We did not expect profile differences based on immersion status (Xu, Di Lonardo Burr et al., 2022).

2.2. Research question 2: Are the relations among number line estimation and performance on other mathematical tasks similar across regions?

In a meta-analysis, Schneider et al. (2018) reported a medium effect size for the correlation between number line estimation and mathematical competence (cf. Ellis et al., 2021). In the present study, we conducted several mini meta-analyses (Goh et al., 2016) using data from the four regions to examine the relations between number line estimation and various mathematical skills. Moreover, we examined whether some tasks were more strongly related to number line performance than others and compared the strength of correlations across regions. Based on Schneider et al.'s findings, we expected medium effect sizes and similar patterns of relations between number line and mathematical performance in all regions.

3. Method

3.1. Participants

Following ethics approval,[1] school principals were contacted. On approval of the principals, letters were sent home to parents,

[1] Ethics approval was obtained from the Institutional Ethics Review Committees at the Pontificia Universidad Católica de Chile, Concordia University, the University of Winnipeg, and Queen's University Belfast.

inviting students to participate. Interested parents provided informed consent for their child to participate and students provided oral assent prior to each testing session. At Time 1, a total of 496 students participated (M_{age} = 7.8 years, SD = 5 months, 45% boys). Students were tested again approximately one year later at Time 2 (M_{age} = 8.8 years, SD = 6 months, 44% boys). Two one-way ANOVAs were conducted to determine if there were age differences across the four regions at Time 1 and Time 2. There was an effect of age at Time 1, $F(3, 511) = 98.41$, $p < .001$, and Time 2, $F(3, 532) = 75.52$, $p < .001$. At both time points, children from Northern Ireland were significantly younger than children from the other three regions (all $ps < .001$). No other age differences were significant.

Across the four regions, 82 students withdrew from the study between testing points (e.g., changed schools, programs, or illness), leaving 414 students who participated in both testing sessions. Eighty-eight new participants joined the study at Time 2 (see Table 1). To determine whether there were differences between students who completed both waves of testing and those who completed one wave of testing, t-tests and χ^2-tests were conducted on the following demographic and outcome variables: gender, region, number comparison (Time 1 and Time 2), and number line estimation (Time 1 and Time 2). Only region differed for those with complete versus incomplete data such that there were fewer students from Winnipeg and Montreal who participated at both time points than there were from Northern Ireland and Chile. We speculate that this difference is the result of recruitment differences wherein ethics approval and consent from parents had to be obtained at both waves of data collection in Winnipeg and Montreal whereas it only needed to be obtained once in Northern Ireland and Chile. In the subsequent analyses, data for students who participated at either one or both time points were included (N = 584; see the Results for details). Students were individually tested by trained research assistants in a quiet area of the schools.

3.2. Openness and transparency

Various papers based on some of the data from these projects have been previously published (Ellis et al. 2021; Song et al., 2021; Xu, Di Lonardo Burr et al., 2022; Xu, Lafay, et al., 2022; Xu, LeFevre et al., 2021). However, none of the analyses in the previously-published works focused on the development of number line performance. Only the measures used in the current analyses are described here, but a complete description of all measures can be found on the Open Science Framework (OSF) sites for the Language Learning and Mathematics Achievement study in Canada and Northern Ireland and the Home Math Environment and the Development of Math Skills in Chile study at these locations: https://osf.io/428hp/ (Manitoba, Quebec, Northern Ireland) and https://osf.io/enxg5/ (Chile). Data were analyzed using Mplus Version 8.3 (Muthén & Muthén, 1998-2017) and SPSS Version 28.0 (IBM Corp, 2021). The study design and analyses were not pre-registered.

3.3. Materials

3.3.1. Datasets

For each region, detailed information about ages, schooling, curricular expectations, and language is shown in Table 1. Data were collected in two academic years (2017–2018 and 2018–2019).

3.3.1.1. Chile. One dataset included students from Chile in Grade 2 at Time 1 and Grade 3 at Time 2. These data were also included in a meta-analysis conducted by (Ellis et al., 2021). Students were either in low-SES schools that received subsidies from the government or high-SES schools that were not subsidized. All schools were located in the urban metropolitan area of Santiago, the Chilean capital. Given the segregation of the educational system in Chile (Mizala & Torche, 2012), schools catered to students from similar socio-economic backgrounds (either low- or high SES). All students received instruction in Spanish.

3.3.1.2. Canada. Two datasets included Canadian students from two provinces: Manitoba and Quebec. The Manitoba number line data were also included in Xu, Di Lonardo Burr et al. (2022). Although both samples were from Canada, Manitoba and Quebec differ both educationally and culturally such that Manitoba and Quebec had the lowest and highest scores of all Canadian provinces on the PISA mathematics assessment, respectively (OECD, 2018). Participants in both provinces attended public (i.e., government funded) schools and lived in middle-class suburbs of either large or small cities. In Quebec, the language of instruction was French. In Manitoba, students attended either English-instruction or French-immersion programs, in which the language of mathematics instruction was English or French, respectively.

3.3.1.3. Northern Ireland. The fourth dataset included students from Northern Ireland. Students were in their third (Year 3; Time 1) and fourth (Year 4, Time 2) years of primary education. Students either attended English-instruction or Irish-immersion programs in which mathematics instruction was in English or Irish, respectively. Schools were in working- to middle-class neighbourhoods, matched across programs for percentage of students who were eligible for free meals.

3.3.2. Measures

3.3.2.1. Number line estimation. Students completed a 0–1000 number line task using the Estimation Line app on an iPad (https://hume.ca/ix/estimationline/). The length of the line depends on the size of the iPad. In the present study, the line was between 13 and 16 cm across the four regions. A target number was shown and students were instructed to point to a position on the number line where they thought the target number should be located. After the student tapped the number line, a red vertical mark was displayed to show

their estimate. Before beginning the task, the students practiced using the iPad by completing two calibration trials. For these trials they were asked to tap a green target on the number line. After the practice trials, 24 experimental trials were randomly presented one at a time above the line (see an example in Fig. 1). Students were instructed to tap the location on the number line where they think the target number belongs. The number line remained on the screen until the student pressed the "Done" button to proceed to the next trial.

Stimuli were selected such that, with the exception of the endpoints (0–99; 900–999), there were two stimuli per hundreds unit with an equal number of trials on either side of the midpoint. There were slightly more trials around the endpoints and midpoint because benchmark strategies are common for number line estimation (e.g., Ashcraft & Moore, 2012; Di Lonardo Burr & LeFevre, 2021; Luwel et al., 2018; Peeters et al., 2016; Xu, 2019). Target numbers were 6, 18, 59, 97, 124, 165, 211, 239, 344, 383, 420, 458, 542, 580, 617, 656, 761, 789, 835, 876, 903, 941, 982, and 994. No feedback was given on the experimental trials. For each trial, the percentage absolute error (PAE) was calculated using the formula:

$$\text{PAE} = [|\text{Estimated Number} - \text{Target Number}| / 1000] \times 100$$

Scoring was the mean PAE across all trials. Cronbach's α based on the PAE of the individual trials was high across the four regions at both Time 1 (αs ranged from .86 to .91) and Time 2 (αs ranged from .84 to .90). There are other indices of number line performance available, but the relations between number line performance and mathematical skills are similar across these indices (Schneider et al., 2018). Thus, PAE was selected because it is scale-independent and easy to interpret.

3.3.2.2. Number comparison. Students completed a number comparison task using the Bigger Number App on an iPad (https://apps.apple.com/app/bigger-number/id1317619133). Students were presented with two single-digit numbers and instructed to tap the numerically bigger number as quickly as possible. After two practice trials, students were presented with 26 experimental trials in random order (see the list of stimuli in Appendix A). Half of the trials had a small distance between the two numbers (i.e., distances of 1–3) and the other half of the trials had a large distance between the two numbers (i.e., distances of 4–7). Cronbach's αs based on response times on correct trials of individual items for each region (i.e., Manitoba, Quebec, Chile, and Northern Ireland, respectively) were .94, .93, .94, and .88 at Time 1 and .94, .94, .94, and .93 at Time 2. To take both speed and accuracy into account, an adjusted response time (RT_{ADJ}) was calculated using the mean correct response time (RT) for each student, adjusted by adding their proportion of error (PE) scaled by the ratio of the standard deviation of the correct response time (SD_{RT}) and the standard deviation of percentage error (SD_{PE}; Vandierendonck, 2018). This linear integrated speed-accuracy score was calculated using the formula: $RT_{ADJ} = RT + (PE \times [SD_{RT}/SD_{PE}])$.

3.3.2.3. Transcoding. Two transcoding tasks were administered: writing and naming. For the writing task, students heard a number word and were asked to write it down in numeral form. For the naming task, students were shown an Arabic numeral and they were asked to name the number. The number of trials and the stimuli varied across regions (see OSF and Appendix A for details).

At Time 1, students in Manitoba, Quebec, and Northern Ireland completed both the writing and naming tasks.[2] Students in Chile completed only the naming task. Students in Manitoba and Northern Ireland were presented with 20 trials for both the writing and naming tasks: one one-digit number, four two-digit numbers, 12 three-digit numbers, and three four-digit numbers. Students in Manitoba completed all 20 trials whereas for students in Northern Ireland, testing was discontinued after three consecutive errors. Students in Chile were presented with one practice trial (one two-digit number), followed by 15 trials: three three-digit numbers, six four-digit numbers, three five-digit numbers, and three six-digit numbers.

At Time 2, students in Manitoba, Quebec, and Northern Ireland completed the writing task. Students from Northern Ireland and Chile completed the naming task. Students in Manitoba and Quebec were presented with up to 30 trials for the writing task: six trials for each set of three-, four-, five-, six-, and seven-digit numbers. Testing was discontinued when a child incorrectly responded to all trials in a set. Students in Northern Ireland completed 28 trials for the writing and naming tasks without discontinuation rules: one one-digit number, four two-digit numbers, twelve three-digit numbers, seven four-digit numbers and four five-digit numbers. In Chile, for the naming task, students were presented with one practice trial (one three-digit number), followed by 15 trials: three four-digit numbers, six five-digit numbers, two trials for each set of six-, seven-, and ten-digit numbers.

The input and output languages for the transcoding tasks were English (Manitoba and Northern Ireland), French (Quebec), and Spanish (Chile). Scores in all regions were the mean number of correctly transcoded items for both tasks. Given that the tasks varied across regions, the reliability measures varied accordingly. At Time 1, for Manitoba and Northern Ireland, internal reliability was calculated based on the accuracy of individual trials (Manitoba: Cronbach's $\alpha = .94$ and .93 for writing and naming, respectively; Northern Ireland: Cronbach's $\alpha = .71$ for writing, whereas individual trials data were not available for naming). For Chile, the internal reliability was calculated based on the accuracy of the individual trials (Cronbach's $\alpha = .96$). At Time 2, for Manitoba, Quebec and Northern Ireland, internal reliability was calculated based the accuracy of individual trials (writing: Cronbach's $\alpha = .86$, .91, and .90 in Manitoba, Quebec, and Northern Ireland, respectively). For Chile, the internal reliability was calculated based on the accuracy of the individual trials (naming: Cronbach's $\alpha = .94$). For Northern Ireland, individual trials data were not available for naming.

[2] We did not include Quebec transcoding at Time 1 because of administrative differences compared to the other regions. Students in Quebec were assigned to one of four task conditions: target size (large, small) x type of transcoding (writing, naming). Thus, because students did not complete all targets for both writing and naming, a sum score that is equivalent for all could not be obtained.

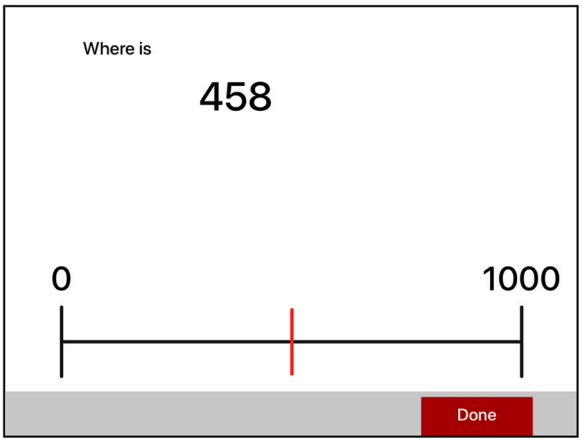

Fig. 1. *Example of a Trial from the Number Line Estimation App.*

3.3.2.4. Word-problem solving. Students in Manitoba and Quebec completed the Applied Problem Solving Test from the KeyMath 3rd Edition (Connolly, 2014) at both Time 1 and Time 2. On each trial, students heard a mathematical word problem accompanied by a visual stimulus (usually a picture), and they provided a verbal response to the question. If they made three consecutive incorrect responses, the task was discontinued. The reported reliabilities for the subset scores of the KeyMath 3rd Edition from the technical report were in the range of .80s (Rosli, 2011). Students in Northern Ireland completed the mathematics reasoning subset of the Wechsler Individual Achievement Test (WIAT-III; Wechsler, 2009) at Time 2. If they made six consecutive incorrect responses, the task was discontinued. The reported reliabilities for the subset scores from the technical report were in the range of .80s and .90s (McCrimmon & Climie, 2011).

Students in Chile completed the Applied Problems subset of the Batería III Woodcock-Muñoz Pruebas de Aprovechamiento (Woodcock-Muñoz Battery III: Achievement Tests; Muñoz-Sandoval et al., 2005). On each trial, students heard a mathematical word problem accompanied by a visual stimulus (usually a picture), and they provided a verbal response to the question. The starting point was dependent on the grade of the child, but baseline was established (Mather & Woodcock, 2005). Testing was terminated when the student made six consecutive errors or failed to respond. The median reported reliability for the Spanish version of the Applied Problems subtest from the technical report was .92 (Schrank et al., 2005).

3.3.2.5. Arithmetic fluency. Students in Manitoba, Quebec, and Northern Ireland completed a paper-and-pencil arithmetic fluency task (Chan & Wong, 2019). For each of three subtests (i.e., addition, subtraction, and multiplication), problems were arranged in a 20 × 3 matrix. Students were given one minute per page to solve as many problems as possible. Students were instructed to start with column 1 and not to skip any problems. At Time 1, students in Manitoba and Quebec completed the addition and subtraction subsets. The reported reliability for the arithmetic fluency task is .97 (Chan & Wong, 2019). Students' performance on addition and subtraction was highly correlated at Time 1 for both regions, *ps* < .001.

At Time 2, students in Manitoba, Quebec, and Northern Ireland completed the addition, subtraction, and multiplication subtests. Students' performance on these subtests was highly correlated at Time 2 for all regions, *ps* < .001. Students in Chile completed the Math Fluency subtest of the Batería III Woodcock-Muñoz Pruebas de Aprovechamiento (Woodcock-Muñoz Battery III: Achievement Tests; Muñoz-Sandoval et al., 2005) at both time points. Students were given three minutes to solve single-digit addition (sum less than or equal to 17), subtraction (the inverse of the previously-mentioned addition questions), and multiplication problems (up to the five times table). Because the reliabilities for the Spanish version of the Math Fluency subset were unavailable from the technical report (Schrank et al., 2005), internal reliability was calculated based the accuracy of individual trials where 75% of the students attempted to respond based on the current sample (Cronbach's α = .79 and .81 at Time 1 and Time 2, respectively).

4. Results

4.1. Research question 1: Are the patterns of development of number line estimation similar across regions?

4.1.1. Quantitative performance on the number line across regions

Mean PAE on the number line task was analyzed in a 2(time: Time 1, Time 2) x 4(region: Manitoba, Quebec, Northern Ireland, Chile) mixed ANOVA with repeated measures on time. Because ANOVA uses listwise deletion, only students with data at both time points were included in the analysis. Performance varied with region, $F(3, 410) = 10.63$, $MSE = 87.05$, $p < .001$, $\eta_p^2 = .07$. Post hoc tests using the Bonferroni adjustment revealed that, averaged across years, students from Chile ($M = 12.5\%$) were more accurate on the number line task than students from Manitoba ($M = 17.2\%$), Northern Ireland ($M = 17.5\%$), and Quebec ($M = 16.1\%$); no significant differences were found among the latter three regions, *ps* > .05. Students improved from Time 1 to Time 2 (18.1% vs. 13.5%), $F(1, 410) = 126.15$, $MSE = 26.57$, $p < .001$, $\eta_p^2 = .24$. These main effects were qualified by a significant region by time interaction, $F(3,$

410) = 4.50, *MSE* = 26.57, $p = .004$, $\eta_p^2 = .03$ (see Fig. 2).

Fig. 2 shows that, as expected, performance at Time 1 was consistent with group differences in age and schooling. Specifically, the Chilean students had better performance than children in Manitoba and Northern Ireland ($ps < .001$); no other group differences were significant ($ps > .05$). At Time 2, the Chilean students maintained their advantage ($ps < .01$), but the other three groups had similar levels of performance ($ps > .05$), presumably because the curricular demands were similar for these groups by this point. Students in Northern Ireland improved the most. Quantitatively, the students from Chile appeared to have better understanding of the number line task sooner than students in the other three regions.

4.1.2. Qualitative patterns of change in number line performance

We used latent profile analysis (LPA) to investigate changes in patterns of performance in the number line task for students in the four regions. Classified as a latent variable modeling approach, LPA is used to identify latent groups within a population based on a set of variables or observations (Collins & Lanza, 2009; Howard & Hoffman, 2018; Wang & Hanges, 2011). In the present study, we use LPA to assume that students can be grouped with varying degrees of probability into profiles that differ based on patterns of estimation.[3]

The cross-sectional LPA was conducted using data from all students who completed the number line task (496 at Time 1 and 502 at Time 2, see Table 1 for numbers in each region). Longitudinal analyses were conducted using the data from all students, based on full information maximum likelihood estimation (FIML),[4] in which model parameters are estimated based on all the information in the variance-covariance matrix, resulting in unbiased parameter estimates under even a high level of missing data for longitudinal analyses under missing at random assumptions (Enders, 2010).

4.1.2.1. Latent profile analysis at Time 1 and Time 2. There were a few outlier scores, defined as |z-scores| > 3.29 from the mean of our sample (Field, 2013), on the number line task ($n = 1$ at Time 1; $n = 3$ at Time 2). Analyses to evaluate the stability of the profile classification (presented in subsequent sections) based on the full sample and a sample excluding these four outliers indicated the same number and type of profile solutions. Thus, the results based on the full sample are reported in the subsequent analyses.

To assess patterns of change in performance on the number line task, we first used cross-sectional LPA at Time 1 and Time 2 to determine whether similar estimation profiles emerged. Each initial model included a single profile. Subsequent models were then compared against the previous models to determine the number of latent profiles that best fit the data. Specifically, we considered three indices to select the best fitting model: Scaled log-likelihood value (LL; higher values indicate better fit); Bayesian Information Criterion (BIC; examining the "elbow plot" to locate the n profile with largest decrease compared to the $n - 1$ profiles; Nylund-Gibson et al., 2014); and Lo-Mendell-Rubin Likelihood ratio test (LMR-LRT; $p < .05$ suggests n is better than $n - 1$ profiles; Nylund et al., 2007). Each model was tested with multiple sets of random start values exceeding 1000, with 50 initial stage iterations (Geiser, 2013). The best log-likelihood value was replicated, suggesting that the optimal set of parameter estimates is trustworthy. The first nonsignificant p value for the LMR-LRT occurred with the three-profile model at Time 1 and four-profile model at Time 2, indicating that adding more classes to the model would not statistically improve model fit. Thus, we stopped testing additional models at the four-class model at both time points.

Fit indices for each model are reported in Table 2. Although the fit statistics did not mutually identify a single LPA model as the best fitting model, the biggest improvement in fit (i.e., largest decrease in BIC) appeared at the two-profile model at both time points. Moreover, LMR-LRT values suggest that the two-profile model was significantly better than the one-profile model, whereas the three-profile model did not significantly improve the model fit at Time 1. Although the three-profile model at Time 2 significantly improved the model fit, further inspection revealed one profile with a small number of students ($n = 23$) with no consistent estimation patterns. When an additional profile does not improve the model, parsimony is favoured (Nylund-Gibson & Choi, 2018). Thus, considering the fit indices and parsimony together, the two-profile solution was retained. Once the final models were selected, we examined the entropy values for overall classification of the model. The entropy values for the two-profile models at both time points were above .80, suggesting that the latent profiles were highly discriminable (Nagin, 2005).

4.1.2.2. Transition from Time 1 to Time 2. Next, we conducted an unconditional latent transition analysis (LTA) based on the selected LPA models. We expected that the structure and number of profiles extracted from the LPA models at Time 1 and Time 2 would be highly similar. As such, we tested a full measurement invariance LTA model, with all measurement parameters held equal between both time points. The results of the unconditional LTA model replicated the results from the LPA (*LL* = 2.02, *BIC* = 189567, Entropy = .87). We labeled the two profiles as *uniform* and *variable* in accord with the patterns shown in Fig. 3 (see also swarm plots [Fig. S1] and spaghetti plots [Fig. S2], which show the location of each estimate for each target on the number line). Specifically, students in the uniform profile made accurate estimates for all target numbers. In contrast, students in the variable profile consistently placed estimates for numbers larger than 100 above the midpoint (i.e., 500) of the number line.

[3] We also computed the statistical fit of each student's responses for linear, logarithmic, and power models and compared those fits to the profiles revealed in the LPA. These statistical models did not provide additional insights into patterns of number line estimation across development. Accordingly, we do not discuss those results further, although the relevant information is included in the Supplementary Material.

[4] Sensitivity analyses of the latent transition analysis showed similar results (i.e., percentages of students who stayed or transitioned from one profile to another) for both FIML and list-wise deletion (i.e., only on students who completed the number line tasks at both time points are included) estimation strategies. Thus, data from all students were included in the analyses.

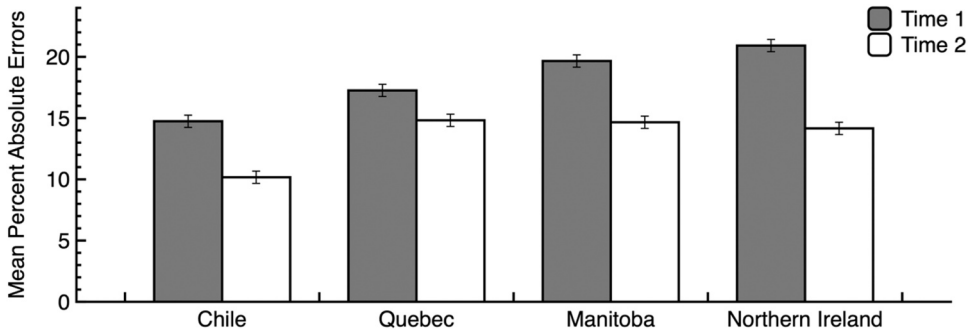

Fig. 2. *Number Line Performance (Percent Absolute Error) for the Interaction of Region by Time.* Error bars are 95% inferential confidence intervals (Jarmasz & Hollands, 2009).

Table 2
Fit Statistics for LPA Models with 1–4 Profiles at Time 1 (N = 496) and Time 2 (N = 502).

Profile-solutions	LL[a]	BIC[b]	ΔBIC	LMR-LRT[c]	Entropy
Time 1					
1	1.52	98396		-	-
2	1.57	**95688**	**2708**	***p* <.001**	.946
3	**1.72**	94898	790	*p*=.108	.961
4	1.68	94159	739	*p*=.056	.946
Time 2					
1	**2.02**	96244		-	-
2	1.97	**93808**	**2436**	*p* < .001	.969
3	1.96	92255	1553	***p* = .030**	.982
4	1.91	91681	574	*p* = .361	.934

Note. The bolded values for LL, BIC, and ΔBIC indicate the "better" model for each index. The bolded value for LMR-LRT indicates the last (*n*) model that shows significant improvement from the *n*-1 model.
[a] LL (Scaling correction factor for MLR);
[b] BIC (Bayesian information criterion);
[c] LMR-LRT (Lo-Mendell-Rubin likelihood ratio test).

We next examined the percentage of students who transitioned from one profile to another over time. At Time 1, 41% of students were classified in the variable profile and 59% of the students were classified in the uniform profile. By Time 2, the percentage of students in the uniform profile increased to 82%, indicating that most of the students demonstrated understanding of number relations by Time 2. However, 18% of students remained in the variable profile at Time 2, indicating that their understanding of the magnitude of three-digit numbers in relation to the number line was weak. From Time 1 to Time 2, 74% of students showed stable performance, either staying in the uniform profile or staying in the variable profile; 24% of students improved in that they transitioned from the variable to uniform profile, and only 2% of students "regressed" from the uniform profile to the variable profile.

As shown in Fig. 3, regardless of region, most of the students remained in the uniform profile or transitioned from the variable to the uniform profile. A chi-square test of independence showed that there was a significant association between region and profile

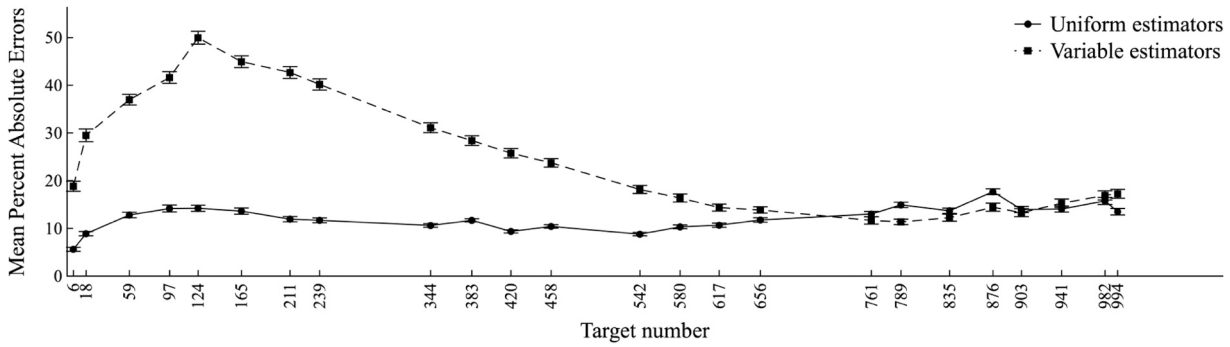

Fig. 3. *Mean Percent Absolute Error for Each Target Number for the Two Profiles in the Latent Transition Analysis Model.* Percent absolute errors were estimated based on the measurement invariant LTA model where the thresholds for qualifying as the variable and uniform profiles hold equal over time. Error bars are the standard errors of the means for each estimated target.

transition, $\chi^2(9) = 36.04$, $p < .001$, Cramer's $V = .14$. We then compared adjusted residuals among the transition profiles, adjusting for multiple comparisons using the Bonferroni method ($p < .004$). A smaller percentage of students in Northern Ireland (43%, $p < .001$) and a larger percentage of students in Chile (73%, $p = .002$) started in the uniform profile and remained in the uniform profile compared to other regions; Quebec (68%) and Manitoba (60%) did not differ significantly. A similar percentage of students transitioned from the variable to the uniform profile in Quebec (22%), Chile (20%), Manitoba (18%), and Northern Ireland (32%), $ps > .004$. Furthermore, a smaller percentage of students in Chile (4%, $p = .001$) started in the variable profile and remained in the variable profile compared to Manitoba (18%), Northern Ireland (22%) and Quebec (11%); no other comparison was significant. In summary, most students were in the uniform profile at Time 2 regardless of region, with Chile continuing to have the strongest performance.

4.1.3. What factors differentiated students across profiles?

To determine the factors that were related to students' performance profile, we explored which demographic variables (i.e., age in months, gender, school SES, and immersion status), and mathematical skills (i.e., number comparison, transcoding, arithmetic fluency, and word-problem solving) were different for students in the two profiles using the estimates obtained from the LTA. Because different measures were used at each region, analyses were conducted by region.

Only three students from Chile and eight students from Quebec stayed in the variable group. Moreover, only 19 students from Chile and 15 students from Quebec transitioned from the variable to uniform profile. By Time 2, most students from these two regions were in the uniform profile (see Fig. 4). Thus, we performed analyses to examine whether there were any differences in demographic or mathematical factors between students who were in the uniform and variable profiles at Time 1 (see Table 3). In contrast, for students in Manitoba and Northern Ireland, there was more room for improvement from Time 1 to Time 2 (see Fig. 4) than for those in the other two regions. Thus, in addition to examining whether there were any differences in demographic or mathematical factors between students who were in the uniform and variable profiles at Time 1, we examined differences for students who moved from or stayed in the variable profile from Time 1 to Time 2 (see Tables 4 and 5).

For students in Chile, the percentage of students from high-SES schools who were in the uniform profile was significantly higher than the percentage of students from low-SES schools (88.2% vs. 66.7%). Boys and girls were equally likely to be in the uniform profile. Within grade, students' age was not related to their profile membership. With respect to mathematics outcomes, students in the uniform profile had better performance on all tasks at both time points, although the difference between profiles was not significant for transcoding naming at Time 2 (see Table 3).

For students in Quebec, there were no significant associations between students' profile allocation (uniform vs. variable) and their school SES (high vs. low). Boys were more likely to be in the uniform profile than girls (84.4% vs. 59.1%). Students' age was not related to their profile membership. With respect to mathematics outcomes, students in the uniform profile at Time 1 had better performance on mathematical word-problem solving at Time 1, and on number comparison and transcoding at Time 2, than students in the variable profile at Time 1 (see Table 2).

As shown in Table 4, for students in Manitoba, boys were more likely to be in the uniform profile than girls (73.8% vs. 51.1%). Older students were more likely to be in the uniform profile than younger students. There were no associations between students' profile allocation and their immersion status. With respect to mathematics outcomes, students in the uniform profile at Time 1 had better performance on number comparison, transcoding, arithmetic fluency, and word-problem solving at both Time 1 and Time 2 than students in the variable profile at Time 1. Moreover, as shown in Table 5, students who transitioned out of the variable profile had

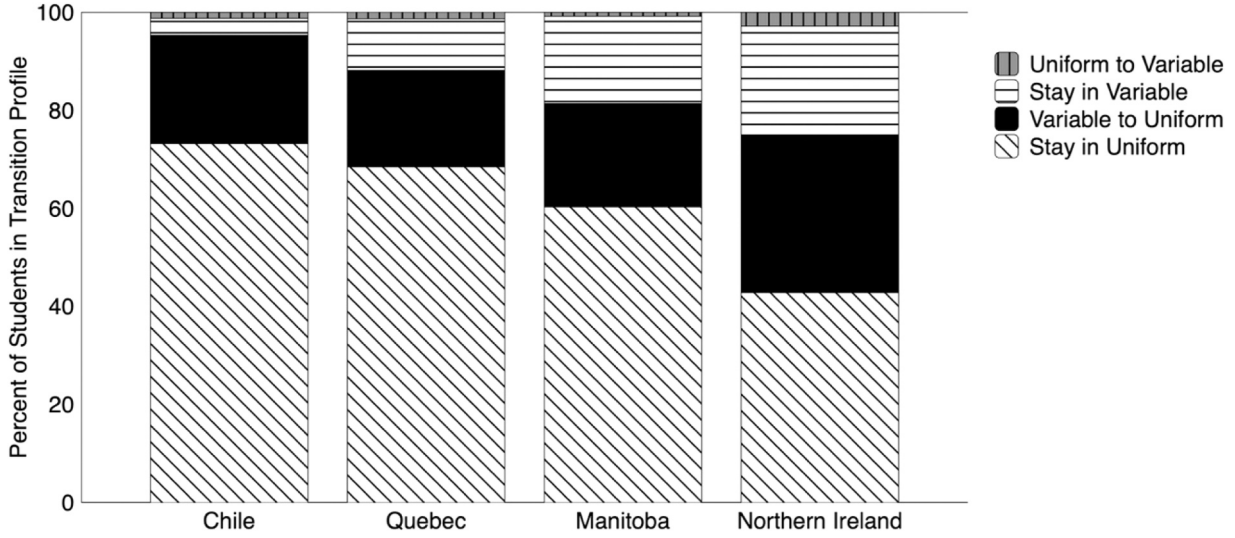

Fig. 4. *Transition Patterns from Time 1 to Time 2 Based on the Unconditional LTA model for Students from Each Region.*

Table 3
Test Statistics for Profile Allocation at Time 1 and Time 2 for Students from Chile and Quebec.

	Chile			Quebec		
	Test statistic	*p*	*Effect size*	*Test statistic*	*p*	*Effect size*
Demographic Variables						
Socioeconomic status	$\chi^2(1, N = 85) = 5.10$	**.039**	.245	$\chi^2(1, N = 81) = 0.55$	.777	.068
Gender	$\chi^2(1, N = 86) < .001$	.999	.001	$\chi^2(1, N = 76) = 5.61$	**.023**	.272
Age (in months)	$t(82) = -0.04$	.972	.009	$t(72) = 0.85$	.396	.214
Time 1						
Number Comparison	$t(83) = -2.58$	**.012**	.638	$t(74) = -1.05$	.297	.262
Transcoding (naming)	$t(84) = 4.36$	**< .001**	1.078	–	–	–
Arithmetic Fluency	$t(84) = 3.77$	**< .001**	.933	$t(70) = 1.11$	.272	.222
Word Problems	$t(84) = 2.76$	**.007**	.681	$t(70) = 2.43$	**.018**	.621
Time 2						
Number Comparison	$t(73) = -2.32$	**.023**	.640	$t(43) = -2.13$	**.039**	.654
Transcoding (naming)	$t(18.02) = 2.06$[a]	.054	.723	–	–	–
Transcoding (writing)	–	–	–	$t(47) = 2.26$	**.029**	.662
Arithmetic Fluency	$t(74) = 4.22$	**< .001**	1.138	$t(47) = 0.66$	.512	.194
Word Problems	$t(74) = 3.93$	**< .001**	1.059	$t(47) = 0.82$	.427	.242

Note. Bolded values were significant at $p < .05$. Effect sizes are phi coefficients for chi-square tests, *r* for Mann-Whitney tests, and *d* for *t*-tests.
[a] Adjusted *df* was used to correct for unequal variance

Table 4
Test Statistics for Profile Allocation at Time 1 and Time 2 for Students from Manitoba and Northern Ireland.

	Manitoba			Northern Ireland		
	Test statistic	*p*	*Effect size*	*Test statistic*	*p*	*Effect size*
Demographic Variables						
Immersion status	$\chi^2(1, N = 242) = 2.90$	.103	.109	$\chi^2(1, N = 179) = 2.10$	.175	.108
Gender	$\chi^2(1, N = 242) = 12.97$	**< .001**	.232	$\chi^2(1, N = 178) = 14.21$	**< .001**	.283
Age (in months)	$t(178) = 3.50$	**< .001**	.525	$t(175) = 1.55$	.123	.234
Time 1						
Number Comparison	$t(175) = -4.14$	**< .001**	.626	$t(173.55) = -4.24$[b]	**< .001**	.629
Transcoding (naming)	$z = -4.69$[a]	**< .001**	.349	$z = -6.22$[a]	**< .001**	.466
Transcoding (writing)	$z = -4.56$[a]	**< .001**	.340	$z = -4.34$[a]	**< .001**	.326
Arithmetic Fluency	$t(178) = 4.33$	**< .001**	.649	–	–	–
Word Problems	$t(178) = 2.63$	**.009**	.394	–	–	–
Time 2						
Number Comparison	$t(113.41) = -4.18$[b]	**< .001**	.669	$t(171) = -4.00$	**< .001**	.612
Transcoding (naming)	–	–	–	$z = -3.00$[a]	**.003**	.227
Transcoding (writing)	$t(207) = 5.99$	**< .001**	.859	$z = -2.16$[a]	**.031**	.163
Arithmetic Fluency	$t(206) = 4.31$	**< .001**	.619	$t(173) = 5.75$	**< .001**	.872
Word Problems	$t(207) = 4.14$	**< .001**	.594	$t(173) = 4.79$	**< .001**	.727

Note. Bolded values were significant at $p < .05$. Effect sizes are phi coefficients for chi-square tests, *r* for Mann-Whitney tests, and *d* for *t*-tests.
[a] Mann Whitney *U* tests were used given the skewed distributions;
[b] Adjusted *df* was used to correct for unequal variance

better performance on transcoding (naming) at Time 1, and on transcoding (writing), number comparison, arithmetic fluency, and mathematical word-problem solving at Time 2 than students who stayed in the variable profile. There were no significant differences between the two profiles for any other variables.

In Northern Ireland, boys were more likely to be in the uniform profile than girls (61.0% vs. 32.7%). There were no significant associations between students' profile allocation and their age or immersion status. With respect to mathematics outcomes, students in the uniform profile at Time 1 had better performance on all the mathematical skills assessed at Time 1 and Time 2 compared to students in the variable profile at Time 1 (see Table 4). Moreover, as shown in Table 5, students who transitioned out of the variable profile had better performance on transcoding (naming) at both time points, and on transcoding (writing) and arithmetic fluency at Time 2 than students who stayed in the variable profile. There were no significant differences between the two profiles for any other variables.

4.1.4. Summary: Research question 1

Patterns of development of number line performance showed that students in all regions improved their number line performance between Time 1 and Time 2, and growth was consistent with age and curricular differences between the regions. Two profiles based on number line performance emerged at each time point: variable and uniform. The quantitative and qualitative analyses revealed that most of the students across the sample either remained in or transitioned to the uniform profile by Time 2, with the students from Chile showing the smallest percentage in the variable profile at Time 2 relative to the other regions. Finally, gender differentiated the

Table 5
Test Statistics for Students in Manitoba and Northern Ireland who Stayed in Versus Moved from the Variable Profile.

	Manitoba			Northern Ireland		
	Test statistic	*p*	*Effect size*	*Test statistic*	*p*	*Effect size*
Demographic Variables						
Immersion status	$\chi^2(1, N = 94) = 0.01$	.999	.009	$\chi^2(1, N = 98) = 0.21$	.679	.046
Gender	$\chi^2(1, N = 94) = 0.67$	.499	.232	$\chi^2(1, N = 98) = 1.00$	.376	.101
Age (in months)	$t(78) = 0.31$	.756	.073	$t(96) = 1.58$	.125	.324
Time 1						
Number Comparison	$t(76) = -1.66$	.102	.391	$t(94) = -1.34$	.183	.279
Transcoding (naming)	$z = -2.83$[a]	**.005**	.317	$z = -2.88$[a]	**.004**	.292
Transcoding (writing)	$z = -1.59$[a]	.112	.178	$z = -1.57$[a]	.116	.159
Arithmetic Fluency	$t(78) = 1.89$	.063	.439	–	–	–
Word Problems	$t(78) = 1.78$	.079	.413	–	–	–
Time 2						
Number Comparison	$t(74.77) = -3.46$[b]	**< .001**	.769	$t(94) = -1.82$	.073	.376
Transcoding (naming)	–	–	–	$z = -2.70$[a]	**.007**	.277
Transcoding (writing)	$t(75) = 2.09$	**.040**	.480	$z = -2.66$[a]	**.008**	.273
Arithmetic Fluency	$t(48.92) = 4.47$[b]	**< .001**	1.093	$t(93) = 2.13$	**.036**	.444
Word Problems	$t(75) = 3.35$	**.001**	.796	$t(93) = 1.92$	.058	.400

Note. Bolded values were significant at $p < .05$.
[a] Mann Whitney *U* tests were used given the skewed distributions;
[b] Adjusted *df* was used to correct for unequal variance

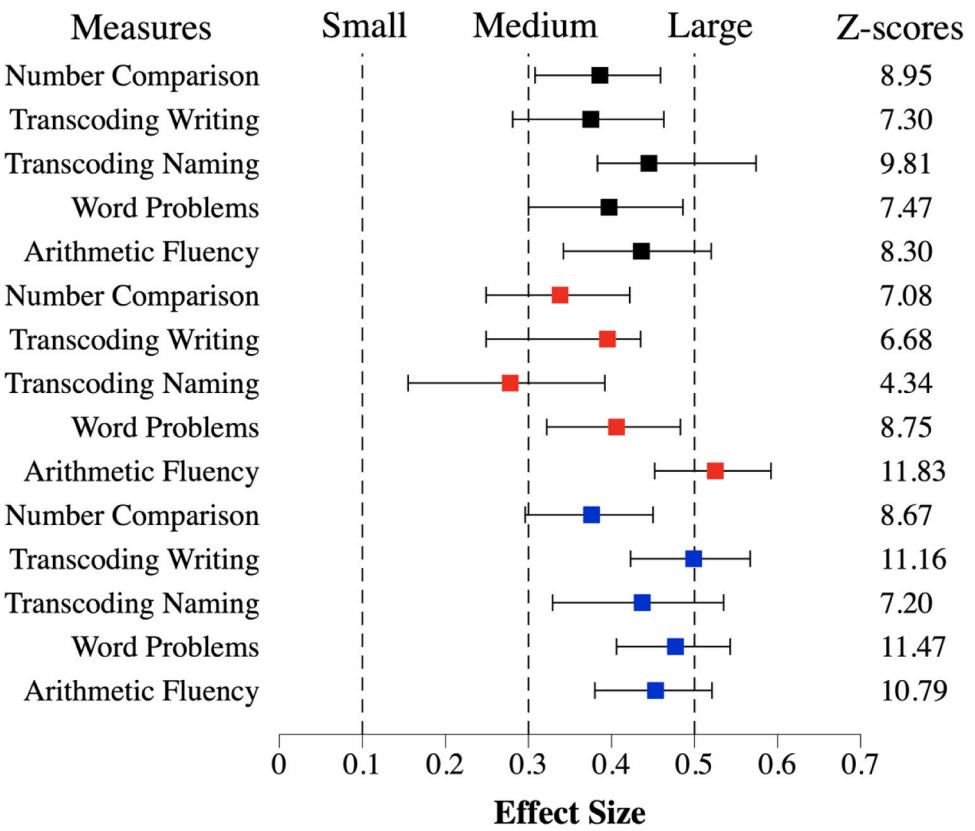

Fig. 5. *Effect Sizes from the Meta-Analysis for the Concurrent and Predictive Associations Between Number Line Estimation and Other Mathematics Outcomes Collapsed Across All Four Regions.* Black squares (rows 1–5) represent concurrent Time 1 correlations. Red squares (rows 6–10) represent correlations between Time 1 no. line and Time 2 mathematics outcomes. Blue squares (rows 11–15) represent concurrent Time 2 relations. All squares represent fixed effects in which the mean effect size (i.e., mean correlation) was weighted by sample size. All correlations were Fisher's z transformed for analyses and converted back to Pearson correlations for presentation. Categorization of effect sizes are based on Cohen (1992).

profiles, with boys more likely to be in the uniform profile at Time 1 than girls in all three regions but Chile, where no gender differences were found. Profile differences were observed in all regions on transcoding and word-problem solving, and number comparison and arithmetic fluency distinguished the profiles in all regions except Quebec.

4.2. Research question 2: Are the relations among number line estimation and performance on other mathematical tasks similar across regions?

Descriptive statistics and correlations among all measures can be found in Tables S2-S6 in the Supplementary Materials. Note that transcoding was negatively skewed for Manitoba (writing and naming) at Time 1 and Northern Ireland (writing and naming) at both Time 1 and Time 2, z-skew > 3.29. Kendall's τ correlations were used for the transcoding measures in Manitoba and Northern Ireland (Field, 2013). The remaining measures were normally distributed and thus Pearson r correlations were used.

Overall, students' number line performance was correlated with all other mathematics measures except for students in Quebec, where number line estimation was not significantly correlated with number comparison at Time 1 or word-problem solving at Time 2. We compared the strength of the correlations between number line estimation and each task across the four regions using univariate analyses of variance (for details and syntax see OSF). There were no significant differences in the strength of the correlations either concurrently (Time 1 number line estimation correlated with Time 1 outcomes; Time 2 number line estimation correlated with Time 2 outcomes) or over time (Time 1 number line estimation predicting Time 2 outcomes) across the four regions ($ps > .05$).

Next, we conducted meta-analyses (Goh et al., 2016) to examine the associations between number line estimation performance and each mathematical task across the four regions (see Fig. 5). Separate meta-analyses were performed for each task concurrently and over time. Fixed effects were used, with the mean effect size weighted by sample size. Across all regions, number line estimation was significantly associated with mathematics outcomes, both concurrently and over time. Except for the relation between number line estimation at Time 1 and arithmetic fluency at Time 2, which had a large effect size, and between number line estimation at Time 1 and naming transcoding at Time 2, which had a small effect size, medium effect sizes (i.e., .34–.50) were found between number line estimation and mathematics outcomes (all $ps < .001$). In summary, for students who varied in country, educational experiences, and language of instruction, number line estimation was consistently related both concurrently and longitudinally to performance on other mathematical tasks.

5. Discussion

Performance on number line tasks is consistently related to both concurrent and later mathematical competence (Schneider et al., 2018). Although researchers have compared performance on the number line task for students in different countries (e.g., Helmreich et al., 2011; Laski & Yu, 2014; Muldoon et al., 2011; Siegler & Mu, 2008; Torbeyns et al., 2015; Xu & LeFevre, 2018), these studies often only consisted of a single time point and did not consider many of the factors known to relate to number line performance, such as age, gender, socioeconomic status, language, and educational experience. In the present study, we examined patterns of number line performance over time across four regions that varied in language spoken and curricular requirements and where students varied in age, language of instruction, and years of formal education. Despite the variability in these factors, patterns of development and relations between number line performance and other mathematical measures were similar across regions, both qualitatively and quantitatively.

5.1. Research question 1: Are the patterns of development of number line estimation similar across regions?

Across regions and at both time points, two profiles of performance emerged: (1) a uniform profile in which estimates were accurately placed across the range of the line; and (2) a variable profile in which students often inaccurately placed estimates for target numbers greater than 100 beyond the midpoint. Regional differences in mean percentage absolute error (i.e., PAE; see Fig. 2) and differential membership in the uniform profile (see Fig. 3) may be a result of students' different educational experiences. At Time 1, approximately 40% of students across regions were classified in the variable profile suggesting that many students had not yet acquired an understanding of place value beyond two digits. In particular, 81% of students in the variable profile were from Manitoba and Northern Ireland, regions in which curriculum expectations did not include numbers greater than 100 at Time 1. Moreover, in Manitoba, there is no public preschool prior to age 5. Thus, those children had fewer years in educational settings than the others.

By Time 2, one year later, curricular expectations in all regions included knowledge of numbers to 1000 and most students (more than 75% in each of the four regions) were in the uniform profile. Thus, across these diverse groups, number line performance showed a convergence related to students' educational experience. Overall, these results support the view that performance on the number line task improves as students gain mathematical knowledge that is directly related to their curricular expectations and amount of schooling (Ashcraft & Moore, 2012; Barth & Paladino, 2011; Bouwmeester & Verkoeijen, 2012; Laski & Siegler, 2007; LeFevre et al., 2013; Muldoon et al., 2013; Praet & Desoete, 2014).

We found that students from Chile consistently made more accurate estimates on the number line than students from the other regions. Notably, number tapes and number lines are routinely used in Chilean schools, which may contribute to the superior performance of Chilean students. Also, school SES differentiated Chilean students in the uniform versus variable profiles. In Chile, most high SES schools are not subsidized by the government and thus parents pay for their children to attend (Allende et al., 2018; Valenzuela et al., 2014). Schools in socially advantaged areas have more resources and better-prepared teachers (Ramírez, 2006). Thus, despite equivalent curricula, students in high-SES schools may be introduced to more advanced mathematical concepts than

those in low-SES schools. In contrast, school SES did not significantly differentiate students from Quebec in profile membership. All the schools in the Quebec sample were publicly funded schools that followed the same curriculum and teachers at these schools had obtained an undergraduate degree in elementary teaching (e.g., Bachelor of Education) and a teaching diploma. In Quebec, school SES designation is not based on school resources, but rather on family income and maternal education. In summary, relations between mathematical performance and school SES may vary within and across countries, depending on how school SES is defined and operationalized.

Of the four regions, students in Northern Ireland showed the greatest improvement in number line estimation from Time 1 to Time 2. Although students in Northern Ireland made the least accurate estimates at Time 1, they were also approximately six months younger than students from the other three regions. These results support the view that estimates on the number line task become more accurate as students get older (Friso-van den Bos et al., 2015; Gunderson et al., 2012; Muldoon et al., 2013; Praet & Desoete, 2014; Schneider et al., 2018; Xu, Di Lonardo Burr et al., 2021). However, in addition to general development, curriculum expectations and subsequent instruction play an important role in the development of mathematics skills. Indeed, after having received formal instruction on numbers to 1000 (i.e., by Time 2), students from Northern Ireland had similar number line performance to students from other regions, indicating that they had acquired knowledge of the relative quantities of three-digit numbers and the ability to place those numbers reasonably accurately. These results suggest that early differences across regions reflect variability related to age and educational experience.

In some previous studies, researchers found relations between language spoken and number line performance (Helmreich et al., 2011; Laski & Yu, 2014; Siegler & Mu, 2009; Xu & LeFevre, 2018). Note, however, that Muldoon et al. (2011) did not find differences between Chinese and Scottish children on number line performance when they were matched on other mathematical skills, showing that cultural differences may not always reflect language differences. In the present research, because students from the four regions did not speak the same first or second languages, we cannot make direct comparisons about the role of first language in any cross-cultural differences we found. However, students from two regions, Manitoba and Northern Ireland, were enrolled in either immersion (French or Irish) or English-instruction programs. Thus, we had the unique opportunity to investigate whether learning mathematics in a second language would relate to patterns of number line performance. For these two groups, there were no significant differences in profile membership by immersion status.

Consistent with prior studies (Gunderson et al., 2012; LeFevre et al., 2013; Rivers et al., 2021; Tian et al., 2022), boys made more accurate number line estimates than girls across all four regions and were more likely to be in the uniform profile than girls in Manitoba, Quebec, and Northern Ireland. There were no gender differences in profile membership in Chile, presumably because most students were in the uniform profile. Additional research is needed to determine whether these differences are related to differential use of spatial strategies for number line estimation for boys versus girls.

In summary, several factors contributed to differences in patterns of number line estimation, both within and between regions. Consistent with previous studies, we found that number line estimation performance was related to age, SES, educational experience, and gender. Thus, when comparing students across regions, it is important to consider variations in the demographic factors that are related to the development of number line estimation skills and to mathematics learning more generally. Despite these variations, however, the pattern of number line performance was similar for these groups of students after all of them had received formal instruction for numbers to 1000.

5.2. Research question 2: Are the relations among number line estimation and performance on other mathematics tasks similar across regions?

Across regions, correlations of similar strength were found between number line estimation and diverse mathematical tasks (i.e., number comparison, number writing and naming, mathematical word-problem solving, and arithmetic fluency). The mini meta-analyses of the associations between number line estimation and these tasks across four datasets revealed medium effect sizes for all measures. This pattern is consistent with the findings from Schneider et al.'s (2018) meta-analysis, which showed a medium effect size for the relations between number line estimation and mathematical competence. In summary, regardless of where students were educated, students who made accurate estimates were also more likely to perform well on other mathematical tasks.

5.3. Implications and future research

The similar patterns of estimates and the similar correlations between number line performance and other mathematical tasks for students from all four regions suggests that the number line task is a useful index of mathematical skill across diverse groups. Importantly, however, educators and researchers must consider the range of the line in relation to curricular expectations. At Time 1, the 0–1000 number line was beyond the curricular requirements in Manitoba and Northern Ireland. A relatively large number of students in Manitoba and Northern Ireland samples were members of the variable profile, consistent with the curricular requirements. By Time 2, based on the curricula for the four regions, all students had experience with representing, comparing, and ordering three-digit numbers to 1000. Accordingly, the data revealed that most students were members of the uniform profile at Time 2. Thus, to obtain uniformly accurate estimates, the range of the task needs to be within students' number-system knowledge.

Acquiring number-system knowledge, however, is not sufficient for accurate performance on the number line estimation task. At Time 2, the mean PAE was above 10% in all regions, suggesting that learning the relations between numerals and magnitudes in the range specified by the task is only the first step in acquiring proficiency. Students also need to apply their proportional reasoning and spatial skills to precisely estimate the location of target numbers (Barth & Paladino, 2011; Gunderson et al., 2012; LeFevre et al., 2013;

Simms et al., 2016; Slusser & Barth, 2017; Tian et al., 2022). For example, knowledge of magnitudes should help students to decide that 503 is only slightly larger than 500 and so it should be placed very close to the midpoint on the 0–1000 number line. In the future, researchers should explore both number-system knowledge and spatial knowledge as crucial factors in number-line development.

In the present study, we used latent transition analysis to investigate the development of number line estimation for students in four regions. Our focus was on empirical patterns of performance and therefore we did not make inferences about underlying representations or strategy use. Other researchers have fit statistical models to similar data to make inferences about theories and strategy use in estimation, for example, using linear and logarithmic functions to infer how children's magnitude representations change (e.g., Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003) or using cyclical power models to infer strategy use (e.g., Barth & Paladino, 2011; Luwel et al., 2018; Slusser et al., 2013). However, as in the present study (see Supplementary Material), patterns of estimates may not show a best fit to one statistical model, nor do the strategies that are inferred from statistical models necessarily reflect the strategies that students use to make estimates (Xu, 2019). The accuracy data in the present study was not sufficient to provide appropriate inferences beyond patterns of performance. Instead, to test theoretical models of strategy use and their relations to mental representations and processes, researchers need to compare results inferred from statistical models with those from other dependent measures such as detailed error patterns for specific targets (Ashcraft & Moore, 2012; Berteletti et al., 2010), verbal strategy reports (Xu, 2019), and eye-tracking data (Di Lonardo et al., 2020; Di Lonardo Burr & LeFevre, 2021). Such detailed investigations may provide further insights into cross-cultural differences that result from varying educational experiences and contribute to a better understanding of how representations and strategies develop in number line estimation.

Beyond accuracy and profile membership, we examined the relations between number line performance and other mathematical measures. We found similar relations across diverse groups, suggesting that the number line is a useful tool for comparisons across countries. However, because the administration, scoring, and stimuli were not consistent across all four regions for the other numeracy and mathematical tasks, we could not use LTA to determine which tasks predicted the transition from uniform to variable profile membership. Ideally, researchers should try and make tasks as consistent as possible across sites (e.g., items, scoring, timing) when examining the relations between number line estimation and mathematical competence across diverse groups of students.

Finally, although we were able to compare number line estimation for students' learning in immersion versus English-instruction programs, we could not investigate the role of differences in language structure with these data because differences in the language of instruction across regions were confounded by other differences (e.g., age, years of education, educational experiences, and SES). Additional processing demands may be required for students to process three-digit numbers if they are learning mathematics in a second language (Barrouillet et al., 2004). Moreover, the number-naming system is more systematic in some languages than others (LeFevre et al., 2018), which may also influence number line performance (Helmreich et al., 2011). For example, two-digit number structure is more complex in French than in English or Spanish (e.g., in French, 70 is sixty-ten; 80 is four-twenty). Additional research is needed to determine how differences in language of instruction may influence number line development.

6. Conclusion

Students from different countries and regions showed similar patterns of number line performance. However, accuracy of number line estimates was related to various factors (i.e., age, school SES, educational experiences, and gender), both within and between regions. Across regions, the relations between number line estimation and other mathematics tasks were consistent. We conclude that, across diverse groups of students who vary with respect to number of years of formal education, language of instruction, and curriculum, the number line task can provide insights into both concurrent and longitudinal mathematics achievement.

Acknowledgement

Support for project was provided by the Social Sciences and Humanities Research Council (SSHRC) of Canada through an Insight Grant 435-2018-1463 to Jo-Anne LeFevre, Erin A. Maloney, Helena P. Osana, and Sheri-Lynn Skwarchuk, by the Chilean National Fund for Scientific and Technology Development (ANID/CONICYT FONDECYT) through Grant FONDECYT 1180675 and ANID – MILENIO – NCS2021_014 to María Inés Susperreguy, and by BA/Leverhulme Small Research Grant SG170445 to Judith Wylie and Victoria Simms. We would like to thank all of the participating schools, school divisions, children, and research assistants for their invaluable contributions to this research.

Appendix A

Tables A1-A5.

Table A1
Stimuli Used in the Transcoding Task for Grade 2 (Version A/Version B) in Manitoba and Northern Ireland.

1-digit	2-digit	3-digit		4-digit
6/7	28/26	113/114	300/304	1132/1145
	40/50	131/141	490 * /380	2360/3420

(*continued on next page*)

Table A1 (*continued*)

1-digit	2-digit	3-digit		4-digit
	79 * /78 *	184 * /186 *	507/503	4302/5031
	96 * /95 *	217/218	673 * /679 *	
		250/270 *	769/722	
		263/279 *	924/984 *	

Table A2
Stimuli Used in the Transcoding Task for Grade 3 in Manitoba and Quebec.

3-digit	4-digit	5-digit	6-digit	7-digit
101	1545	42,000	246,000	6002,000
392 *	2398 *	16,070 *	581,000 *	4000,070 *
210	4063	14,030	603,100	1400,000
688 *	3072 *	82,067 *	400,678 *	5080,000 *
834	5302	20,137	574,321 *	3000,000
976 *	6183 *	93,284 *	297,783 *	2090,080 *

Table A3
Stimuli Used in the Transcoding Task for Grade 3 (Version A/Version B) in Northern Ireland.

1-digit	2-digit	3-digit		4-digit	5-digit
6/7	15/16	113/114	300/304	1214/1215	12,415/14,817
	40/50	131/141	490 * /380	2360/3420	32,619/33,518
	79/78	184/186	512/513	4302/5031	56,214/54,311
	96/95	217/218	673/679	5816/6711	78,510/76,420
		250/270	769/722	6519/7317	
		263/279	918/914	8211/8312	
				9677/9582	

Table A4
Stimuli Used in the Transcoding Naming Task for Grade 2 in Chile.

3-digit	4-digit	5-digit	6-digit
769	1500	14030	246000
834	4302	42000	603100
967	7145	20737	874321
	2063		
	5398		
	8699		

Table A5
Stimuli Used in the Transcoding Naming Task for Grade 3 in Chile.

4-digit	5-digit	6-digit	7-digit	10-digit
4302	14030	246000	6000200	3000000000
2063	42000	603100	9000000	5345772183
8699	20737			
	88750			
	76055			
	98808			

Symbolic Number Comparison Stimuli Set.
Number pairs – small distance size: 5–2; 4–7; 5–3; 5–8; 3–2; 5–7; 8–6; 7–9; 5–4; 5–6; 7–8; 7–8; 9–8.
Number pairs – large distance size: 1–8; 7–1; 1–6; 5–1; 2–9; 8–2; 2–7; 9–3; 3–8; 3–7; 9–4; 4–8; 9–5.
Transcoding Stimuli Set.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cogdev.2023.101355.

References

Agencia de Calidad de la Educación. (2015). Metodología de construcción de grupos socioeconómicos Simce 2013 [Methodology for constructing SIMCE 2013 socieconomic groups]. ⟨http://archivos.agenciaeducacion.cl/Metodologia_de_Construccion_de_Grupos_Socioeconomicos_Simce_2013.pdf⟩.
Allende, C., Diaz, R., & Valenzuela, J. P. (2018). *School segregation in Chile. Global encyclopedia of public administration, public policy, and governance*. Springer International Publishing,. https://doi.org/10.1007/978-3-319-31816-5

Ashcraft, M. H., & Moore, A. M. (2012). Cognitive processes of numerical estimation in children. *Journal of Experimental Child Psychology, 111*(2), 246–267. https://doi.org/10.1016/j.jecp.2011.08.005

Bakker, M., Torbeyns, J., Wijns, N., Verschaffel, L., & De Smedt, B. (2019). Gender equality in 4- to 5- year-old preschoolers' early numerical competencies. *Developmental Science, 22*(1), Article e12718. https://doi.org/10.1111/desc.12718

Barrouillet, P., Camos, V., Perruchet, P., & Seron, X. (2004). ADAPT: A developmental, asemantic, and procedural model for transcoding from verbal to Arabic numerals. *Psychological Review, 111*(2), 368–394. https://doi.org/10.1037/0033-295X.111.2.368

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift: Development of numerical estimation. *Developmental Science, 14*(1), 125–135. https://doi.org/10.1111/j.1467-7687.2010.00962.x

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology, 46*, 545–551. https://doi.org/10.1037/a0017887

Booth, J. L., & Newton, K. J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology, 37*(4), 247–253. https://doi.org/10.1016/j.cedpsych.2012.07.001

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2012). Multiple representations in number line estimation: A developmental shift or classes of representations? *Cognition and Instruction, 30*(3), 246–260. https://doi.org/10.1080/07370008.2012.689384

Bull, R., Cleland, A. A., & Mitchell, T. (2013). Gender differences in the spatial representation of number. *Journal of Experimental Psychology: General, 142*, 181–192. https://doi.org/10.1037/a0028387

Chan, W. W. L., & Wong, T. T.-Y. (2019). Subtypes of mathematical difficulties and their stability. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000383

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Connolly, A. J. (2014). Keymath-3 Diagnostic Assessment. In C. R. Reynolds, K. J. Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of Special Education* (p. 1341). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118660584.ese1341.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, JK: Wiley. https://doi.org/10.1002/9780470567333

Di Lonardo, S. M., Huebner, M. G., Newman, K., & LeFevre, J.-A. (2020). Fixated in unfamiliar territory: Mapping estimates across typical & atypical number lines. *Quarterly Journal of Experimental Psychology, 73*(2), 279–294. https://doi.org/10.1177/1747021819881631

Di Lonardo Burr, S. M., & LeFevre, J.-A. (2021). Fixated in more familiar territory: Providing an explicit midpoint for typical and atypical number lines. *Quarterly Journal of Experimental Psychology, 74*(3), 523–535. https://doi.org/10.1177/1747021820967618

Dietrich, J. F., Huber, S., Dackermann, T., Moeller, K., & Fischer, U. (2016). Place-value understanding in number line estimation predicts future arithmetic performance. *British Journal of Developmental Psychology, 34*(4), 502–517. https://doi.org/10.1111/bjdp.12146

Ellis, A., Susperreguy, M. I., Purpura, D. J., & Davis-Kean, P. E. (2021). Conceptual replication and extension of the relation between the number line estimation task and mathematical competence across seven studies. *Journal of Numerical Cognition, 7*(3), 435–452. https://doi.org/10.5964/jnc.7033

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press,.

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics*. Sage,.

Friso-van den Bos, I., Kroesbergen, E. H., Van Luit, J. E. H., Xenidou-Dervou, I., Jonkman, L. M., Van der Schoot, M., & Van Lieshout, E. C. D. M. (2015). Longitudinal development of number line estimation and mathematics performance in primary school children. *Journal of Experimental Child Psychology, 134*, 12–29. https://doi.org/10.1016/j.jecp.2015.02.002

Geary, D. C., Hoard, M. K., Nugent, L., & Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Developmental Neuropsychology, 33*(3), 277–299. https://doi.org/10.1080/87565640801982361

Geiser, C. (2013). *Data analysis with Mplus*. The Guilford Press,.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164–1165. https://doi.org/10.1126/science.1154094

Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: The role of the linear number line. *Developmental Psychology, 48*(5), 1229–1241. https://doi.org/10.1037/a0027433

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of gender differences in science and mathematics. *Psychological Science in the Public Interest, 8*, 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Helmreich, I., Zuber, J., Pixner, S., Kaufmann, L., Nuerk, H.-C., & Moeller, K. (2011). Language effects on children's nonverbal number line estimations. *Journal of Cross-Cultural Psychology, 42*(4), 598–613. https://doi.org/10.1177/0022022111406026

Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods, 21*(4), 846–876. https://doi.org/10.1177/1094428117744021

Huber, S., Moeller, K., & Nuerk, H.-C. (2014). Dissociating number line estimations from underlying numerical representations. *Quarterly Journal of Experimental Psychology, 67*(5), 991–1003. https://doi.org/10.1080/17470218.2013.838974

Hutchison, J. E., Lyons, I. M., & Ansari, D. (2019). More similar than different: Gender differences in children's basic numerical skills are the exception not the rule. *Child Development, 90*(1), e66–e79. https://doi.org/10.1111/cdev.13044

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences, 106*(22), 8801–8807. https://doi.org/10.1073/pnas.0901265106

IBM Corp. (2021). *IBM SPSS Statistics for Windows Version 28.0*. Armonk, NY: IBM Corp,.

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology, 63*(2), 124–138. https://doi.org/10.1037/a0014164

Jung, S., Roesch, S., Klein, E., Dackermann, T., Heller, J., & Moeller, K. (2020). The strategy matters: Bounded and unbounded number line estimation in secondary school children. *Cognitive Development, 53*, Article 100839. https://doi.org/10.1016/j.cogdev.2019.100839

Lanza, S. T., & Cooper, B. R. (2016). Latent class analysis for developmental research. *Child Development Perspectives, 10*, 59–64. https://doi.org/10.1111/cdep.12163

Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development, 78*(6), 1723–1743. https://doi.org/10.1111/j.1467-8624.2007.01087.x

Laski, E. V., & Yu, Q. (2014). Number line estimation and mental addition: Examining the potential roles of language and education. *Journal of Experimental Child Psychology, 117*, 29–44. https://doi.org/10.1016/j.jecp.2013.08.007

LeFevre, J.-A., Cankaya, O., Xu, C., & Jimenez Lira, C. (2018). Linguistic and experiential factors as predictors of young children's early numeracy skills. In D. B. Berch, D. C. Geary, & K. Mann Koepke (Eds.), *Mathematical Cognition and Learning* (Vol. 4). Elsevier.

LeFevre, J.-A., Jimenez Lira, C., Sowinski, C., Cankaya, O., Kamawar, D., & Skwarchuk, S.-L. (2013). Charting the role of the number line in mathematical development. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00641

Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Gender differences in spatial cognition: Advancing the conversation. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*, 127–155. https://doi.org/10.1002/wcs.1380

Link, T., Nuerk, H.-C., & Moeller, K. (2014). On the Relation between the Mental Number Line and Arithmetic Competencies. *Quarterly Journal of Experimental Psychology, 67*(8), 1597–1613. https://doi.org/10.1080/17470218.2014.892517

Luwel, K., Peeters, D., Dierckx, G., Sekeris, E., & Verschaffel, L. (2018). Benchmark-based strategy use in atypical number lines. *Canadian Journal of Experimental Psychology Revue canadienne Délelőtt psychologie expérimentale, 72*(4), 253. https://doi.org/10.1037/cep0000153

Mather, N., & Woodcock, R.W. (2005). Manual del examinador (L. Wolfson, Trans.). Woodcock-Johnson III Pruebas de aprovechamiento [Examiner's manual. Woodcock- Johnson III Tests of Achievement]. Riverside.
McCrimmon, A.W., & Climie, E.A. (2011). Test Review: D. Wechsler Wechsler Individual Achievement Test—Third Edition. San Antonio, TX: NCS Pearson, 2009. Canadian Journal of School Psychology, 26(2), 148–156. https://doi.org/10.1177/0829573511406643.
Mizala, A., & Torche, F. (2012). Bringing the schools back in: the stratification of educational achievement in the Chilean voucher system. *International Journal of Educational Development, 32*(1), 132–144. https://doi.org/10.1016/j.ijedudev.2010.09.004
Muldoon, K., Simms, V., Towse, J., Menzies, V., & Guoan, Yue (2011). Cross-cultural comparisons of 5-year-olds' estimating and mathematical ability. *Journal of Cross Cultural Psychology, 42*(4), 669–681. https://doi.org/10.1177/0022022111406035
Muldoon, K., Towse, J., Simms, V., Perra, O., & Menzies, V. (2013). A longitudinal analysis of estimation, counting skills, and mathematical ability across the first school year. *Developmental Psychology, 49*(2), 250–257. https://doi.org/10.1037/a0028240
Mullis, I.V. S., Martin, M.O., Foy, P., Kelly, D.L., & Fishbein, B. (2020). TIMSS 2019 International Results in Mathematics and Science. https://timssandpirls.bc.edu/timss2019/international-results/.
Muñoz-Sandoval, A.F., Woodcock, R.W., McGrew, K.S., & Mather, N. (2005). Batería III Woodcock- Muñoz: Pruebas de Aprovechamiento [Woodcock-Muñoz Battery III: Achievement Tests]. Riverside.
Muthén, B., & Muthén, L. K. (1998). *Mplus User's Guide* (7th ed..,). Muthén & Muthén,.
Nagin, D. (2005). *Group-Based Modeling of Development*. Harvard University Press,. https://doi.org/10.4159/9780674041318
Nuraydin, S., Stricker, J., Ugen, S., Martin, R., & Schneider, M. (2023). The number line estimation task is a valid tool for assessing mathematical achievement: A population-level study with 6484 Luxembourgish ninth-graders. *Journal of Experimental Child Psychology, 225*, Article 105521. https://doi.org/10.1016/j.jecp.2022.105521
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396
Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440–461. https://doi.org/10.1037/tps0000176
Nylund-Gibson, K., Grimm, R., Quirk, M., & Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 439–454. https://doi.org/10.1080/10705511.2014.915375
Oberski, D. L. (2016). A review of latent variable modeling with R. *Journal of Educational and Behavioral Statistics, 41*, 226–233. https://doi.org/10.3102/1076998615621305
OECD. (2018). Programme for International Student Assessment. http://www.oecd.org/pisa/publications/pisa-2018-results.htm.
Peeters, D., Degrande, T., Ebersbach, M., Verschaffel, L., & Luwel, K. (2016). Children's use of number line estimation strategies. *European Journal of Psychology of Education, 31*(2), 117–134. https://doi.org/10.1007/s10212-015-0251-z
Praet, M., & Desoete, A. (2014). Number line estimation from kindergarten to grade 2: A longitudinal study. *Learning and Instruction, 33*, 19–28. https://doi.org/10.1016/j.learninstruc.2014.02.003
Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development, 79*(2), 375–394. https://doi.org/10.1111/j.1467-8624.2007.01131.x
Ramírez, M.-J. (2006). Understanding the low mathematics achievement of Chilean students: A cross-national analysis using TIMSS data. *International Journal of Educational Research, 45*(3), 102–116. https://doi.org/10.1016/j.ijer.2006.11.005
Reinert, R. M., Huber, S., Nuerk, H.-C., & Moeller, K. (2017). Sex differences in number line estimation: The role of numerical estimation. *British Journal of Psychology, 108*(2), 334–350. https://doi.org/10.1111/bjop.12203
Rivers, M. L., Fitzsimmons, C. J., Fisk, S. R., Dunlosky, J., & Thompson, C. A. (2021). Gender differences in confidence during number-line estimation. *Metacognition and Learning, 16*(1), 157–178. https://doi.org/10.1007/s11409-020-09243-7
Rosli, R. (2011). Test Review: A. J. Connolly KeyMath-3 Diagnostic Assessment: Manual Forms A and B. Minneapolis, MN: Pearson, 2007. *Journal of Psychoeducational Assessment, 29*(1), 94–97. https://doi.org/10.1177/0734282910370138
Rosselli, M., Ardila, A., Matute, E., & Inozemtseva, O. (2009). Gender differences and cognitive correlates of mathematical skills in school-aged children. *Child Neuropsychology, 15*(3), 216–231. https://doi.org/10.1080/09297040802195205
Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: a meta-analysis. *Child Development, 89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068
Schrank, F.A., McGrew, K.S., Ruef, M.L., Alvarado, C.G., Muñoz-Sandoval, A.F., & Woodcock, R.W. (2005). Overview and technical supplement (Batería III Woodcock-Muñoz Assessment Service Bulletin No. 1).
McCrimmon, A.W., & Climie, E.A. (2011). Test Review: D. Wechsler Wechsler Individual Achievement Test—Third Edition. San Antonio, TX: NCS Pearson, 2009. Canadian Journal of School Psychology, 26(2), 148–156. https://doi.org/10.1177/0829573511406643.
Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x
Siegler, R. S., & Mu, Y. (2008). Chinese children excel on novel mathematics problems even before elementary school. *Psychological Science, 19*(8), 759–763. https://doi.org/10.1111/j.1467-9280.2008.02153.x
Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237–250. https://doi.org/10.1111/1467-9280.02438
Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001
Simms, V., Clayton, S., Cragg, L., Gilmore, C., & Johnson, S. (2016). Explaining the relationship between number line estimation and mathematical achievement: The role of visuomotor integration and visuospatial skills. *Journal of experimental child psychology, 145*, 22–33. https://doi.org/10.1016/j.jecp.2015.12.004
Slusser, E. B., & Barth, H. C. (2017). Intuitive proportion judgment in number-line estimation: Converging evidence from multiple tasks. *Journal of Experimental Child Psychology, 162*, 181–198. https://doi.org/10.1016/j.jecp.2017.04.010
Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology General, 142*(1), 193–208. https://doi.org/10.1037/a0028560
Song, S., Xu, C., Maloney, E., Skwarchuk, S-L., Di Lonardo Burr, … LeFevre, J-A. (2021). Longitudinal relations between young students' feelings about mathematics and arithmetic performance. *Cognitive Development, 59*, Article 101078. https://doi.org/10.1016/j.cogdev.2021.101078
Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and gender differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology, 101*, 20–51. https://doi.org/10.1016/j.jecp.2008.02.003
Tian, J., Dam, S., & Gunderson, E. A. (2022). Spatial skills, but not spatial anxiety, mediate the gender difference in number line estimation. *Developmental Psychology, 58*(1), 138–151. https://doi.org/10.1037/dev0001265
Tikhomirova, T., Malykh, A., Lysenkova, I., Kuzmina, Y., & Malykh, S. (2022). The development of number line accuracy in elementary school children: A cross-country longitudinal study. *British Journal of Educational Psychology*. https://doi.org/10.1111/bjep.12566
Torbeyns, J., Schneider, M., Xin, Z., & Siegler, R. S. (2015). Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents. *Learning and Instruction, 37*, 5–13. https://doi.org/10.1016/j.learninstruc.2014.03.002
Valenzuela, J. P., Bellei, C., & Ríos, D. D. L. (2014). Socioeconomic school segregation in a market-oriented educational system. The case of Chile. *Journal of Education Policy, 29*(2), 217–241. https://doi.org/10.1080/02680939.2013.806995
Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *Journal of Cognition, 1*(1), 8. https://doi.org/10.5334/joc.6

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of gender differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*, 250. https://doi.org/10.1037/0033-2909.117.2.250

Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods, 14*(1), 24–31. https://doi.org/10.1177/1094428110383988

Wechsler, D. (2009). *Wechsler Individual Achievement Test* (Third edition.,). Pearson,.

Whyte, J. C., & Bull, R. (2008). Number games, magnitude representation, and basic number skills in preschoolers. *Developmental Psychology, 44*(2), 588–596. https://doi.org/10.1037/0012-1649.44.2.588

Xu, C. (2019). Ordinal skills influence the transition in number line strategies for children in Grades 1 and 2. *Journal of Experimental Child Psychology, 185*, 109–127. https://doi.org/10.1016/j.jecp.2019.04.020

Xu, C., Di Lonardo Burr, S., Douglas, H., Susperreguy, M. I., & LeFevre, J.-A. (2021). Number line development of Chilean children from preschool to the end of kindergarten. *Journal of Experimental Child Psychology, 208*, Article 105144. https://doi.org/10.1016/j.jecp.2021.105144

Xu, C., Di Lonardo Burr, Skwarchuk, S-L., Douglas, H., Lafay, A., … LeFevre, J-A. (2022). Pathways to learning mathematics for students in French-immersion and English-instruction programs. *Journal of Educational Psychology, 114*(6), 1321–1342. https://doi.org/10.1037/edu0000722

Xu, C., & LeFevre, J.-A. (2016). Training young children on sequential relations among numbers and spatial decomposition: Differential transfer to number line and mental transformation tasks. *Developmental Psychology, 52*(6), 854–866. https://doi.org/10.1037/dev0000124

Xu, C., & LeFevre, J.-A. (2018). Cross-cultural comparisons of young children's early numeracy performance: Effects of an explicit midpoint on number line performance for Canadian and Chinese-Canadian children. *Bordón Revista Dėlelőtt Pedagogía, 70*(3), 131. https://doi.org/10.13042/Bordon.2018.60966

Xu, C., LeFevre, J-A., Skwarchuk, S-L., Di Lonardo Burr, Lafay, A., … Simms, V. (2021). Individual differences in the development of children's arithmetic fluency from grades 2 to 3. *Developmental Psychology, 57*(7), 1067–1079. https://doi.org/10.1037/dev0001220

Xu, C., Lafay, A., Douglas, H., Di Lonardo Burr, S. M., LeFevre, J-A., Osana, H. P., Skwarchuk, S-L., Wylie, J., Simms, V., & Maloney, E. A. (2022). The role of mathematical language skills in arithmetic fluency and word-problem solving for first- and second-language learners. *Journal of Educational Psychology, 114*(3), 513–539. https://doi.org/10.1037/edu0000673

Xu, X., Chen, C., Pan, M., & Li, N. (2013). Development of numerical estimation in Chinese preschool children. *Journal of Experimental Child Psychology, 116*(2), 351–366. https://doi.org/10.1016/j.jecp.2013.06.009

Zhang, T., Chen, C., Chen, C., & Wei, W. (2020). Gender differences in the development of semantic and spatial processing of numbers. *British Journal of Developmental Psychology, 38*(3), 391–414. https://doi.org/10.1111/bjdp.12329