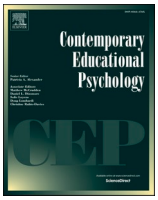
Contemporary Educational Psychology 75 (2023) 102224

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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



Disentangling magnitude processing, natural number biases, and benchmarking in fraction comparison tasks: A person-centered Bayesian classification approach

Frank Reinhold [a,*], Timo Leuders [a], Katharina Loibl [b]

[a] Institute for Mathematics Education, University of Education Freiburg, Freiburg, Germany
[b] Institute for Education, University of Education Freiburg, Freiburg, Germany

ARTICLE INFO

Keywords:
Transitive strategies
Typical errors
Natural number bias
Magnitude estimation
Size comparison

ABSTRACT

Research on fraction comparison shows that students often follow biased comparison strategies, in particular such strategies that build on their knowledge of natural numbers. On the other hand they also apply successful comparison strategies such as benchmarking or fraction magnitude processing. Which strategies are applied or even combined depends on the students' knowledge and on the task type. To investigate these complex relationships, we developed a balanced 2 × 2-dimensional itemset (congruent vs. incongruent items; benchmarking vs. non-benchmarking items) and a Bayesian classification of individual students' performance (solution patters, response time, and individual distance effect), which we applied to an assessment of $N = 350$ sixth graders. We could show that the classification of the students with respect to possible solution strategies matched our hypotheses: We could replicate existing patterns *and* found additional composite strategies such as 'benchmarking or bias' with a bias only in solution rates of non-benchmark items. In further analyses we found 'benchmarking or suppressed bias-strategies (i.e., a bias in problem solving time of non-benchmarking items). Our study extends previous knowledge on individual strategies in fraction comparison and proposes a new person-centered approach to classify individual student profiles even with small profile sizes.

1. Introduction

Research during the last decades has shown that understanding fractions is a challenge for most students (e.g., Behr et al., 1983; Lortie-Forgues et al., 2015; Siegler et al., 2011). One of the main reasons for this phenomenon is the high complexity of cognitive processes students have to perform when *learning* fractions (e.g., Loibl & Leuders, 2019; Reinhold, Hoch, et al., 2020; Van Hoof et al., 2018) and when *operating* with fractions (e.g., Reinhold, Obersteiner, et al., 2020; Vamvakoussi & Vosniadou, 2004; Van Hoof et al., 2013). One commonly used task type to investigate cognitive processes which lie at the heart of fraction understanding is the symbolic fraction comparison task. If students engage in the problem whether 3/7 or 2/3 is the larger fraction, different cognitive processes may be activated (and also combined): (1) *fraction magnitude processing*, (2) *symbolic fraction manipulation*, that is, finding a common denominator via expanding and reducing both fractions to compare the resulting numerators afterwards, or (3) *utilizing benchmarks*, that is, comparing both given fractions to a salient third one—which may utilize (1) and/or (2). Apart from these normatively correct processes, many students apply erroneous processes, such as (4) different types of *biased reasoning* by processing only part of the given information (i.e., numerators or denominators) in isolation.

During the last years, a lot of effort has been put into the empirical identification of these cognitive processes, among them think-aloud studies (e.g., González-Forte et al., 2019), eye tracking studies (e.g., Obersteiner & Tumpek, 2016), or fMRI studies (e.g., Obersteiner, Dresler, et al., 2019). A frequently applied empirical approach relies on assessments that are specifically designed to trigger observable answer patterns, following the argument that the underlying cognitive processes *map* to systematic behavior in certain comparison tasks. In the present study, we follow this approach, but unlike most studies, we aim to include distinct cognitive processes that may be activated when comparing two fractions into one cognitive model and, thus, to shed light on how the processes may interact on an individual level. To this end, we show how a Bayesian classification approach (updating probabilities of hypotheses on students' comparison strategies) may serve as

* Corresponding author at: University of Education Freiburg, Institute for Mathematics Education, Kunzenweg 21, 79117 Freiburg, Germany.
E-mail address: frank.reinhold@ph-freiburg.de (F. Reinhold).

https://doi.org/10.1016/j.cedpsych.2023.102224

Available online 6 September 2023

an effective and non-explorative statistical approach to identify distinct cognitive processes and group students according to their behavior in assessment tasks. In the following we briefly summarize the state of research with respect to the cognitive processes "fraction magnitude processing" (1.1), "benchmarking" (1.2) and "biased reasoning (i.e., natural number bias, 1.3). We then describe how research tried to identify these processes and in how far the common approach of averaging and clustering patterns of groups of students is limited (1.4). Finally we present our approach by explicating a model for the cognitive processes and by proposing a person-centered approach of classifying individual solution patterns (1.5).

1.1. The role of fraction magnitude processing in comparison tasks: The distance effect

Fraction understanding requires knowledge about very different meanings of fractions, such as ratio, part-whole relationship, division, measurement—and magnitude (Behr et al., 1983). Fraction magnitude is assumed to be a mental representation of a continuous part-whole ratio (especially visual represented numerosities, lengths, areas) as a *single* quantity—representing *one* numerical value. Research has found ample evidence for such a mental magnitude representation in schoolchildren, infants, and even primates (Bonn & Cantlon, 2017; Matthews & Chesney, 2015). It is assumed that after appropriate instruction symbolically represented fractions can be associated with such an approximate magnitude (Reinhold, Obersteiner, et al., 2020). In formal instruction of fractions this ability is usually supported (and assessed) by representing the fraction on a number line (Schneider & Siegler, 2010). If two to-be-compared fractions can be successfully represented by construction of a mental model of the two fraction magnitudes, the larger of both fractions can be identified by directly inspecting the mental model.

A commonly-used index to relate fraction magnitude processing to observable behavioral data is the *distance effect*—i.e., the smaller the numerical distance between the two to-be-compared fractions, the more difficult the item (e.g., Schneider et al., 2017, for an overview). There is empirical evidence that such a distance effect may be present both regarding accuracy (Reinhold, Obersteiner, et al., 2020; Sprute & Temple, 2011) and response time (e.g., Meert et al., 2010).

1.2. The role of benchmarking in fraction comparison tasks: Salient straddling benchmarks

The use of *straddling benchmarks* (Obersteiner et al., 2020) in fraction comparisons can serve as a viable strategy in specific tasks with appropriate reference points, such as 1/2. Students who apply this strategy can be expected to argue, for example, that 4/9 is smaller than 2/3, because $4/9 < 1/2$ and $1/2 < 2/3$. Such benchmarking strategies (Clarke & Roche, 2009; also referred to as transitive comparison, Post et al., 1986; or correct reference point comparison Behr et al., 1986) are well reported and occur, even if they are not explicitly taught to students (Clarke & Roche, 2009; Fazio et al., 2016). Although some empirical findings indicate that students use benchmarks (González-Forte et al., 2019; Liu, 2018), current studies are rather inconclusive on the specific role that the use of benchmarks plays for individuals when comparing the size of two fractions (DeWolf & Vosniadou, 2011, 2015; Obersteiner et al., 2020). The benchmark magnitudes used by young adults seem to be restricted to 1/2, 1/3, 1/4, and 3/4 with 1/2 being most efficient with respect to response times and error rates (Liu, 2018).

From a cognitive perspective, there are two plausible processes that may underly such a straddling benchmarking strategy for the reference point 1/2: Students may apply a magnitude estimation and locate the magnitudes of the two fraction as 'lower' or 'higher' compared to 1/2 and then infer by inspecting the resulting mental model. Another process for comparing one fraction to the benchmark 1/2 relies on a basic reasoning by performing an arithmetic doubling (or halving) procedure: 4/9 is smaller than 1/2 because the double of 4 is smaller than 9 (or 4 is smaller than half of 9).

Both processes, magnitude comparison or arithmetic reasoning, are very plausible since they rely on very basic mental models and procedures that are familiar even before formal schooling, since 1/2 is the most fundamental magnitude, and since doubling is a very fundamental arithmetic process (Liu, 2018; Meert et al., 2010). Research has not yet been able to distinguish these two processes (i.e., magnitude comparison and arithmetic reasoning) during the utilization of benchmarking strategies.

In both processes, the numerical *distance* between the two fractions to be compared is of no relevance when utilizing such benchmarking strategies, as the two fractions are not compared quantitatively to each other via their magnitudes, but only categorically to one half as a third fraction. Following this argument, it can be questioned if the commonly used approach to assess magnitude processing in fraction comparison tasks (i.e., the distance effect) serves as an adequate operationalization when salient benchmarks like 1/2 or 1 are involved (Reinhold, Obersteiner, et al., 2020). However, Obersteiner et al. (2020) reported that—contrary to their hypothesis—benchmarks did not facilitate the use of magnitude estimation in complex fraction comparison tasks for the whole sample, as no difference in participants' error patterns were found between benchmark and no-benchmark items in the sample-based analysis of their study.

A core idea for the present study, resulting from this state of research is: In order to distinguish cognitive processes of benchmarking, one cannot rely on distance effects, but needs to base the classification of a students' strategy on the answer patterns (solutions rates and/or response times) when systematically varying items with and without an opportunity of a salient straddling benchmark strategy.

1.3. The role of biases in fraction comparison tasks: Congruent and incongruent items

An erroneous but common cognitive process that students engage in when comparing two fractions is to fall back to simple comparisons of the numerators and the denominators as natural numbers: $3 > 2$ and $7 > 3$ and therefore $3/7 > 2/3$. This phenomenon is referred to as the so-called *natural number bias* (Alibali & Sidney, 2015; Ni & Zhou, 2005)—which has attracted a lot of attention within the last years (DeWolf & Vosniadou, 2011; Gómez & Dartnell, 2019; González-Forte et al., 2019; Obersteiner et al., 2020; Obersteiner, Marupudi, et al., 2019; Prediger, 2008; Reinhold, Obersteiner, et al., 2020; Rinne et al., 2017; Vamvakoussi & Vosniadou, 2004; Van Hoof et al., 2013, 2018).

When students reason in line with a natural number bias they show systematic and observable error patterns, i.e., to incorrect responses in so-called *incongruent* tasks (i.e., the larger fraction is composed of the smaller natural numbers, e.g., 3/7 vs. 2/3), and to correct responses in so-called *congruent* tasks (i.e., the larger fraction is composed of the larger natural numbers, e.g., 3/8 vs. 5/9). Besides such natural number bias patterns in solution rates (e.g., Meert et al., 2010; Reinhold, Obersteiner, et al., 2020; Vamvakoussi & Vosniadou, 2004), it could also be found in response times, i.e., faster response times in congruent items than in incongruent items, given a comparably high solution rate (e.g., Vamvakoussi, Van Dooren, & Verschaffel, 2012; Van Hoof, Lijnen, Verschaffel, & Van Dooren, 2013; Obersteiner, Van Hoof, & Verschaffel, 2013). In addition, some studies also have found systematic reverse bias patterns—i.e., correct responses or faster response times in incongruent items than in congruent items—in students (e.g., Gómez & Dartnell, 2019; Reinhold, Obersteiner, et al., 2020; Rinne et al., 2017) and adults (Obersteiner et al., 2020). Given this reliable empirical evidence for a variety of different 'natural number biases' as phenomena (in terms of observable behavior during the assessment of fraction comparison tasks), the underlying cognitive processes are still a matter of debate (Alibali & Sidney, 2015; Vamvakoussi et al., 2012):

(1) If students have no actual fraction knowledge, it is reasonable

that they will engage in natural number reasoning: 'the smaller the numbers the smaller the fraction'—which will result in a systematic *typical bias* pattern. This is in line with a conceptual change view on the development of rational number concepts (Vamvakoussi & Vosniadou, 2004), suggesting that students engaging in the above-mentioned cognitive processes are unaware of their erroneous reasoning. (2) If students have made their first conceptual change from natural number concepts to rational number concepts, they may have a plausible so-called synthetic concept, and follow the argument that 'for fractions, it is the other way around' (Reinhold, Obersteiner, et al., 2020; Rinne et al., 2017), leading to the erroneous reasoning: 'the smaller the numbers the larger the fraction' (González-Forte et al., 2019; 2023). Again, the underlying cognitive processes focus on the processing of natural numbers—yet, resulting in a systematic *reverse bias* pattern exactly the other way around as the typical bias pattern. As described, and in contrast to the strategy mentioned before, this strategy would require an awareness for (synthetic) fraction concepts. (3) An alternative explanation for the above mentioned *reverse bias* pattern is that students consistently use a *gap thinking* strategy in fraction comparison tasks (Gómez & Dartnell, 2019) which can also be considered a synthetic concept in terms of conceptual change. Following this strategy students would argue that the larger the difference between the numerator and the denominator, the smaller the fraction (González-Forte et al., 2019; 2023). The underlying cognitive process again relies on natural number concepts, explicitly applying subtraction. Consistent application of gap thinking in items with non-common components and proper fractions would result in the *reverse bias* pattern—because it always leads to correct solutions in incongruent items (e.g., $2/3 > 4/9$, because $3 - 2 = 1$ and $9 - 4 = 5$), but it *may* lead to incorrect solutions in congruent items (e.g., $1/3 > 5/9$, because $3 - 1 = 2$ and $9 - 5 = 4$; see Gómez & Dartnell, 2019).

Students applying the three strategies above mentioned will show a *strong* natural number bias—i.e., a typical or a reverse bias pattern in the *solution rates* of congruent and incongruent fraction comparison items. Here, conceptual change theory gives a plausible rationale for student thinking.

Besides such strong natural number biases, *suppressed* natural number biases are also well-reported—i.e., a typical or a reverse bias pattern in the *response times* of congruent and incongruent fraction comparison items combined with high solution rates in all items. Here, we use the term 'suppressed' to indicate that a *dual process* account (cf. Van Dooren & Inglis, 2015, for a discussion of dual processes in mathematics thinking) can describe student thinking that would lead to such systematic patterns in response time in the following sense: Students with a further elaborated fraction knowledge (i.e., a complete conceptual change from natural to rational number concepts) may still struggle with incongruent and congruent fraction comparison tasks. Here, at least two different dual processes seem plausible: (1) natural number concepts are quickly and automatically activated in the intuitive system, and in some tasks (depending on item congruency) have to be intervened with rational number concepts by the analytic system (Vamvakoussi et al., 2012)—resulting in longer response time in incongruent (i.e., *suppressed typical bias*) or in congruent items (i.e., *suppressed reverse bias*). Another plausible explanation for the underlying cognitive processes is (2) that in fraction comparison tasks both natural and rational number concepts are activated automatically and intuitively, and the analytical system has to negotiate between the two lines of reasoning, resulting in longer response times when natural and rational number reasoning leads to different decisions. This would lead to the same patterns of different response times in congruent and incongruent items as described above.

However, following up on this discussion about the underlying cognitive processes that lead to such observable bias patterns in the solution rate, it is rather unlikely that students demonstrating any of the described natural number biases actually process *fraction* magnitudes when comparing fractions. Instead, it is plausible to assume that their underlying cognitive processes are different types of *natural number* magnitude estimations (e.g., comparison of the magnitude of natural numbers, or estimating the difference between two natural numbers, i.e., subtraction)—which are straightforward considering the base-ten system. In fact, Reinhold, Obersteiner, et al. (2020) could show that low-performing six-graders demonstrating a persistent natural number bias did *not* show a significant distance effect in fraction comparison tasks, while their classmates demonstrating no natural number bias did show a distance effect—which empirically underpins the above mentioned assumption of no fraction magnitude processing in biased students.

1.4. The role of interindividual differences in research on fraction comparison tasks

In earlier studies, the natural number bias was usually investigated as a group phenomenon, i.e., utilizing whole-sample approaches to study average differences in performance or response time in congruent tasks when compared to incongruent tasks—which led to meaningful insights, but also to inconclusive results. A generally agreed upon explanation for those inconclusive results is that whole-sample approaches may mask individual students' development of fraction understanding that may lead to answer patterns not in line with the performance found on the group level. Instead, different stages of students' knowledge on a certain topic, such as fractions, should be included.

Consequently, this argument recently led to the demand to substitute sample-based statistical approaches by more person-centered statistical approaches, which aim at unmasking individual differences in the strategies used by the students (e.g., Flunger et al., 2017). In fact, these studies could reveal individual differences in bias patterns—resulting from differences in the underlying cognitive processes (Gómez & Dartnell, 2019; González-Forte et al., 2019; Reinhold, Obersteiner, et al., 2020; Rinne et al., 2017): Besides students demonstrating a *typical bias* pattern (i.e., better performance on congruent than incongruent tasks), students showing a *no bias* pattern or even a *reversed bias* pattern (i.e., higher solution rates in incongruent than congruent tasks) were identified.

Against these findings, we suggest that the question whether students show a strong typical bias or a strong reverse bias may not be adequately answered in whole samples; instead it seems more adequate to ask which part of the sample systematically shows one or the other cognitive process. To elaborate on that idea, Reinhold, Obersteiner, et al. (2020) did not find a statistically relevant effect of item congruency with a whole sample approach; in fact, the sample could be clustered into three roughly equally-sized groups, showing (a) a strong typical bias, (b) a strong reverse bias, and (c) no bias at all. As the typical and the reverse biased students showed comparable item congruency effects into the opposite direction, this would statistically result in a vanishing effect in the whole sample. Given the actual results, we argue that the interpretation of an absence of an item congruency effect in that specific sample (which would be a coherent interpretation of a whole sample statistical approach) does not describe the strategies actually used by students. Reviewing present studies from this point of view may raise plausible explanations for contradictory or non-expected findings, such as the ones referred to above. Methodologically, this situation may be better illuminated by a person-centered analysis.

1.5. The present study

When focusing on fraction comparison tasks and considering congruent and incongruent items separately, former studies did show that students typically follow clear and replicable patterns. However, the role of benchmarking, the role of fraction magnitude processing beyond a distance effect, and whether there are students who rely on biased patterns only in tasks that do not allow for salient benchmarking strategies, remains unclear. The present study attempts to address these

research gaps.

We argue that in order to disentangle intraindividual differences in the underlying cognitive processes that students engage in when comparing two fractions, it is necessary to (1) consider congruent and incongruent items, as well as benchmark and no-benchmark items, (2) to recognize that different cognitive processes affect different behavioral data, such as solution rates, reaction times, or a distance effect, (3) to consider that a students' systematically used strategy may consist of more than one unique cognitive process, and (4) to apply person-centered statistical approaches that allow to unmask individual differences. Drawing on these four points, we developed a balanced fraction comparison itemset with a 2 × 2-dimensional structure: congruent vs. incongruent items, and items that can (benchmark item) vs. cannot (no-benchmark item) be solved using 1/2 as a reference point. Our hypothesis was that we would be able to replicate patterns for familiar comparison strategies—*and* to find additional mixed-pattern profiles, by sequentially considering solution rates firstly, response times secondly, and distance effects for more differentiated comparison strategies. Our hypothesized strategies are given in Table 1.

To illustrate this further, we argue that, e.g., for students who are able to apply benchmarking at 1/2 as a strategy to compare fractions it is reasonable to assume that they may "fall back" to other, non-sophisticated, or even erroneous strategies when their focused benchmarking strategy cannot be applied in a specific item. Moreover, consider a student not having completed the conceptual change from natural to rational number concepts; we argue that neither their problem-solving time, nor a potential distance effect would be informative about their underlying cognitive processes, but only their solution pattern. Furthermore, we argue that students who show a suppressed typical bias and a suppressed reverse bias cannot be expected to differ in terms of solution rates, but only in terms of reaction times—which is why we chose a sequential analysis of different data (see Table 1). In step 1 of the sequential analysis, we classified students based on their solution rates; in step 2, we classified those students who solved (nearly) all items—i.e., the "Capable Solving" row in Table 1—based on their response times; in a third and final step, we classified those students who solved (nearly) all items and did not show relevant differences in problem solving time—i.e., the "Proficient Solving" row in Table 1—based on their individual distance effects.

In contrast to recent studies, we did not apply an explorative person-centered statistical approach (such as a cluster analysis, or a latent profile analysis), but developed a hypothesis-driven method to group students with regard to our hypothesized strategies, using a Bayesian updating procedure of probabilities of classification hypotheses. Our hypotheses for the different strategies given in Table 1 transform into the anchor points for our sequential Bayesian classification approach, which represent our predictions of solution patterns and which are given in Fig. 1 for classification step 1 (solution rates) and in Fig. 2 for classification step 2 (response times). We give a more differentiated argument for these predictions in the next section.

2. Method

2.1. Sample

Our sample consist of $N = 350$ German six-grade students from the states of Bavaria and Baden-Württemberg. Fractions are taught in grade six with comparable curricular frameworks in both states. The study was conducted at the end of the school year after the formal instruction of fractions. It is relevant to mention that the most-prominent way to teach fraction comparison in the German curriculum is via finding a common denominator (Heck Ribeiras et al., 2022; Reinhold & Reiss, 2020). Therefore, it is assumed that (a) some students still are affected by a natural number bias (Reinhold, Obersteiner, et al., 2020) and (b) more elaborated ways of comparing fractions are found by students themselves and not due formal instruction (Clarke & Roche, 2009).

2.2. Instrument

We developed a 2 × 2 structured itemset, consisting of congruent and incongruent items, as well as benchmark and no-benchmark items with respect to 1/2. We used single-digit numerators and denominators, and left out fractions with common components and fractions with the numerator 1. For each of the four cells, we selected six fraction comparison items in a way that the final itemset showed no statistically significant difference in distance between the two to be compared fractions between the four cells, $F(3, 20) = 2.70$, $p = .07$; a Levene's test showed homogeneity of variance in distance between the four cells, $F(3, 20) = 0.87$, $p = 0.47$. The complete 24 items in the order as presented to the students can be reviewed in Fig. 3.

To underpin our assumption that our balanced 2 × 2 dimensional itemset is appropriate to investigate our hypothesis of different correct

Table 1
Hypothesized strategies to compare two fractions that students can systematically apply, consisting of various cognitive processes uniquely found in previous studies.

	Solution rate				*Response time differences*		*Distance effect*
	Benchmark item		*No-benchmark item*		*Benchmark item*	*No-benchmark item*	
Strategies visible in solution rates	*Congruent*	*Incongruent*	*Congruent*	*Incongruent*	*Congruent to incongruent*	*Congruent to incongruent*	*Overall*
Strong typical bias	high	low	high	low	—	—	—
Strong reverse bias	low	high	low	high	—	—	—
Guess	by chance	by chance	by chance	by chance	—	—	—
Benchmark, or guess	high	high	by chance	by chance	—	—	—
Benchmark, or strong typical bias	high	high	high	low	—	—	—
Benchmark, or strong reverse bias	high	high	low	high	—	—	—
Capable solving	high	high	high	high	—	—	—
Strategies visible in problem-solving time							
Suppressed typical bias	high	high	high	high	slower	slower	—
Suppressed reverse bias	high	high	high	high	faster	faster	—
Benchmark, or suppressed typical bias	high	high	high	high	equally fast	slower	—
Benchmark, or suppressed reverse bias	high	high	high	high	equally fast	faster	—
Proficient solving	high	high	high	high	equally fast	equally fast	—
Strategies visible in distance effect							
Rapid procedural arithmetic	high	high	high	high	equally fast	equally fast	no effect
Pure magnitude estimation	high	high	high	high	equally fast	equally fast	high effect

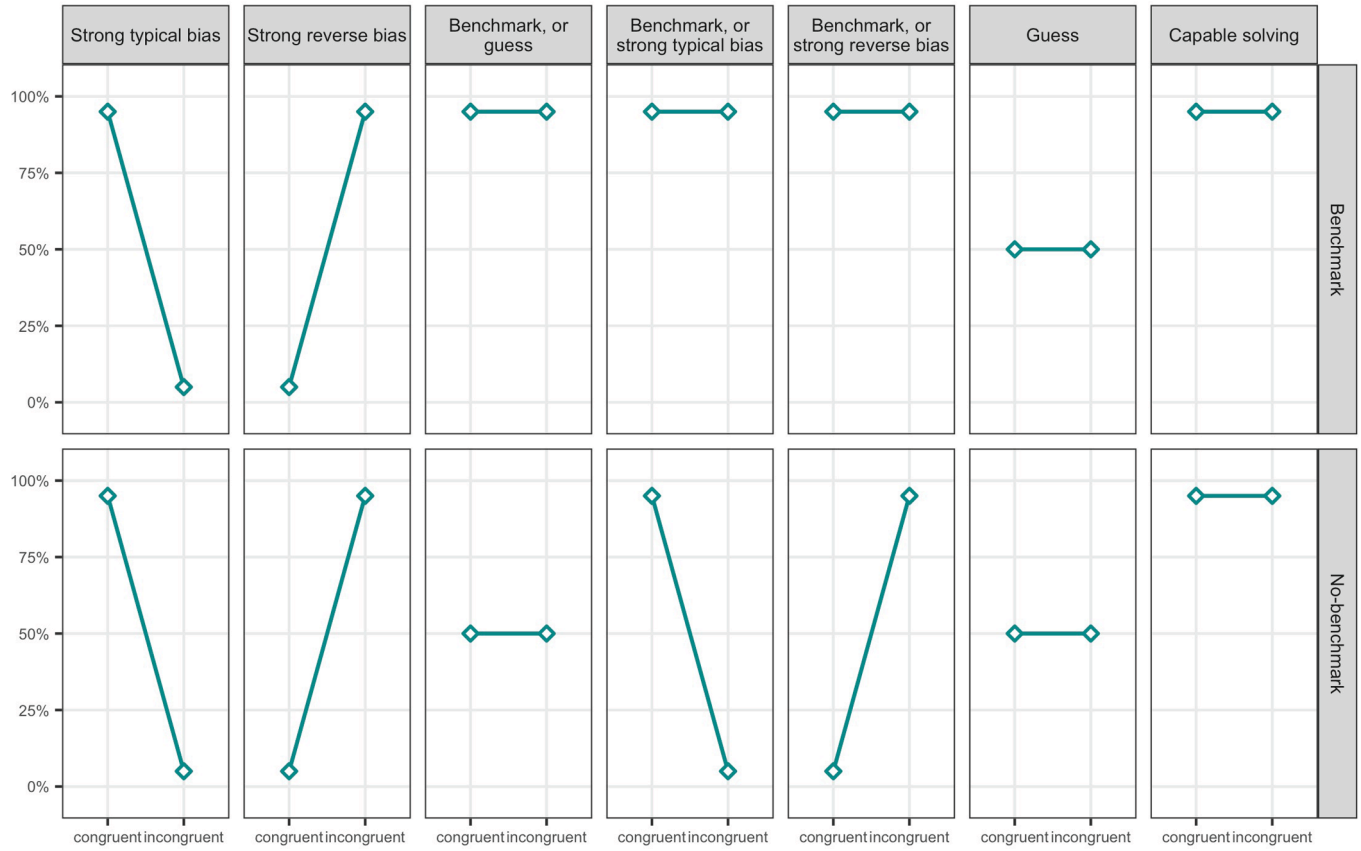


Fig. 1. Hypotheses for patterns in the solution rate, when different strategies are used to compare two fractions, used in step 1 of our Bayesian updating procedure of probabilities of classification hypotheses.

or erroneous strategies when engaging in items from these four cells, we conducted a confirmatory factor analysis on the solution rates of the 24 items—with diagonally weighted least squares (DWLS) to estimate the model parameters (as the data is binary for each item, i.e., 0 = incorrect answer, 1 = correct answer)—comparing different theoretically plausible item structures. Overall, the integrated 2 × 2 factor model yielded a very good model-data fit, $RMSEA = 0.034$, $CFI = 0.976$, $TLI = 0.973$ (see Table 2 for the fit indices of all four compared models). Theoretically, focusing only on natural number biases would lead to a two factor model (congruent vs. incongruent items). The 2 × 2 factor model yielded a better fit to the data than this two factor "congruence-only" model, scaled chi-square differences between models $\chi^2(5) = 14.4$, $p < .05$. Theoretically, focusing only on the utilization of benchmarks would lead to another two factor model (benchmark vs. no-benchmark items). The 2 × 2 factor model yielded a better fit to the data than this two factor "benchmark-only" model, $\chi^2(5) = 970.3$, $p < .001$. Theoretically, focusing only on magnitude estimation would lead to a one factor model (fraction comparison problems). The 2 × 2 factor model yielded a better fit to the data than this one factor "magnitude-only" model, $\chi^2(6) = 70.6$, $p < .001$. Given these results, we consider our itemset appropriate to be used to investigate our hypothesis.

2.3. Procedure and conditions

Students took part in the assessment anonymously, on a voluntary basis, and with the informed consent given by themselves as well as by their parents. Both the Bavarian State Ministry of Education (IV.7-BO4106.2019/52/9) and the Freiburg District Council Department of Schools and Education (7–6499.2) approved the study. The assessments took place during mathematics lessons in schools and were conducted and/or supervised by the first author of this paper.

The study was conducted in a digital environment with one fraction comparison task per page. The items were given in randomized groups of four items (one from each cell of the 2 × 2 itemset), with a fixed sequence for all participants (Fig. 3). For each fraction comparison task, two fractions were presented on the touch screen side-by-side and students had to select the larger of the two fractions by clicking on it.

To increase variation in the data without changing the reasoning processes in principle but only varying their prevalence, we randomly assigned the participants to one of four variants of the digital environment (see Fig. 2) that differed gradually in how salient magnitude and benchmarking with one half is triggered. In each variant, students were asked to select the larger fraction.

Version A represents a standard, non-prompted design of the fraction comparison task where both fractions appeared at the screen at the same time. Version B and version C aimed at stimulating magnitude processing by (i) presenting the fraction on the right first for 5 s, (ii) presenting a fraction bar below both fractions, and (iii) prompting students to imagine the size of both fractions on the fraction bar in version B or to imagine the size of the second fraction given the size of the first fraction in version C. Furthermore, version D aimed at stimulating a benchmarking strategy by highlighting 1/2 in the fraction bars in addition to version C (see Fig. 4).

We measured response times in each of the four conditions starting when both fractions were present on the screen and ending when students chose one fraction by finger typing on it.

2.4. Data and statistical analyses

In order to analyze the student data, we applied a Bayesian classification approach, by which each individual can be classified as pertaining to a class of individuals with a typical strategy. Since the behavior of

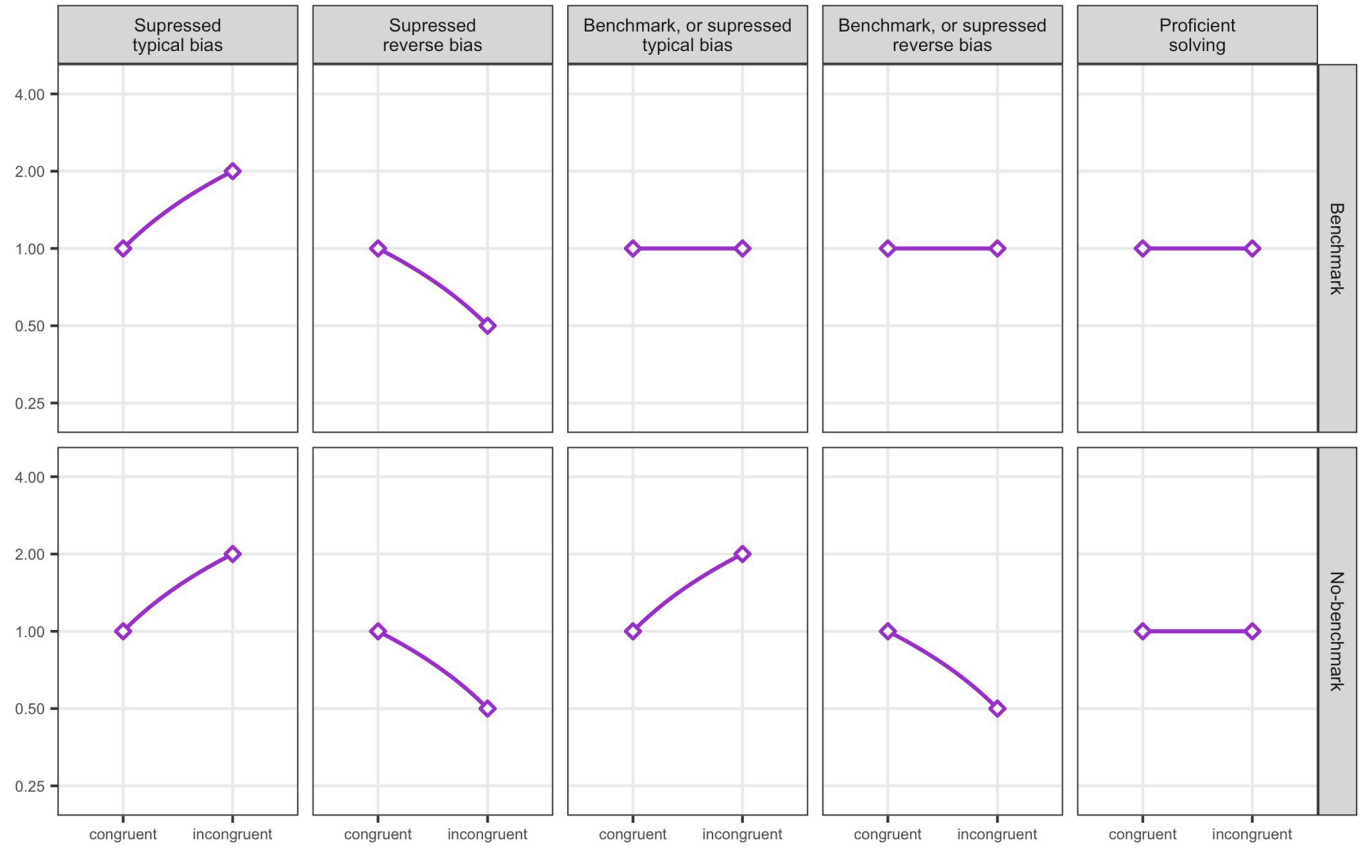


Fig. 2. Hypotheses for patterns in the response time, when different strategies are used to compare two fractions, used in step 2 of our Bayesian updating procedure of probabilities of classification hypotheses.

each student in each task is neither unique nor deterministic, it is reasonable to assume a probability distribution for each student, to belong to each of the possible classes, and a probabilistic answer pattern depending on the strategies used.

In order to account for this uncertainty in interpreting a student's reasoning methodologically, in our Bayesian classification approach we assume that each student has a consistent comparison strategy which can be modeled by a set of probabilities:$p_i(H^k)$ represents the probability of the hypothesis that student i has the comparison strategy k. This probability has an (unknown) prior value and is updated by each the pieces of evidence collected in the experiment. More specifically, for the strategies visible in solution rates (cf. Table 1), we account for the fact that students only approximately respond according to their strategy by attributing to the evidence E (i.e., the students' actual behavior of giving a correct or false response) the likelihood $p_i\left(E|H^k\right)$. This likelihood expresses the conditional probability of the evidence E (correct or wrong answer) under the condition of the student having a strategy k. For example, the probability for answering correctly to an incongruent task without a benchmark, is low for a student with a natural-number bias, and high for a student who applies a magnitude comparison.

When a student responds consistently by applying one strategy in all (or most) tasks, the cumulative evidence ($E_{i,j}$ for each student i and task j) should lead to a considerable increase in the respective posterior probability for this strategy via Bayesian updating[1]:

$$p_i^{\text{posterior}}(H^k) \propto p\left(E_{i,1}|H^k\right) \cdot \ldots \cdot p\left(E_{i,24}|H^k\right) \cdot p_i^{\text{prior}}(H^k)$$

The exact values for the likelihoods are not known, however we can assign a low (0.1), medium (0.5 = chance), and high (0.9) solution probability for each task and strategy, as indicated in Table 1. Fig. 5 illustrates the change of probability during a typical sequence of solutions in one student.

The procedure does not aim at estimating exact probabilities, but at distinguishing between the strategies by increasing or decreasing their relative probabilities based on each new piece of evidence (i.e., each response to a new task). These assumptions define a Naïve Bayesian Classification procedure (Duda et al., 2012). This approach belongs to the class of Bayesian networks, which have become frequently used in educational assessment (Culbertson, 2016). The simplifying assumption of independence of the steps has proven adequate in many applications (e.g., in disease prediction, Langarizadeh & Moghbeli, 2016), even with dependencies between the likelihoods (Domingos & Pazzani, 1997).

The classification of the student i as having strategy k is then supported by the amount of change in the probability ratios. These changes of probability based on evidence are typically expressed by Bayes factors. In the present analysis, there are pairwise ratios of any two hypotheses (i.e. strategies). $BF_{1:2}(i)$, for example, is defined by

$$\frac{p_i^{\text{posterior}}(H^1)}{p_i^{\text{posterior}}(H^2)} = \underbrace{\prod_{j=1\ldots24} \frac{p\left(E_{i,j}|H^1\right)}{p\left(E_{i,j}|H^2\right)}}_{=:BF_{1:2}(i)} \cdot \frac{p_i^{\text{prior}}(H^1)}{p_i^{\text{prior}}(H^2)}$$

To substantiate the classification decision for each student we recur to (a) the ratio of the dominant hypothesis to the subsequent one, e.g. $BF_{1:2}(i) = 100:1$ and (b) the highest posterior probability, when assuming equally distributed priors, e.g. $p_i^{\text{posterior}}(H^1) = 99.9\%$. This ratio of the dominant hypothesis to the subsequent one produces a standardized measure (i.e., the Bayes factor) for the accuracy of each

[1] The calculations were programmed by Timo Leuders in CindyScript (https://www.cinderella.de/); the code can be made available by request.

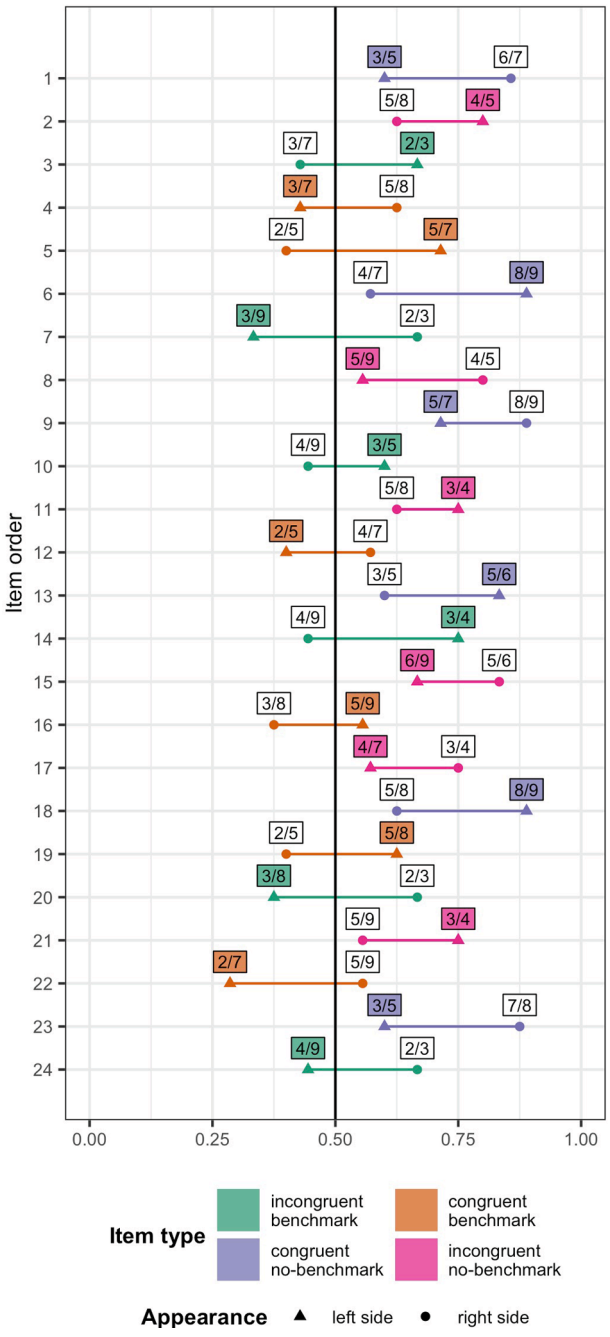


Fig. 3. Complete itemset used in the study, showing the 2 × 2 structure in different colors. The fraction that was given on the left side of the screen is indicated by a triangle and highlighted.

students' classification decision: For example, if a student's actual answer pattern would result in two equally plausible patterns of the hypothesized prior probabilities (Table 1), this would result in a *BF* of 1; if, on the other hand, a student's actual answer pattern would result in a dominant hypothesis three times as plausible as the second most plausible hypothesis, this would result in $BF = 3$. We only consider students with an accuracy of $BF > 3$ as appropriately classified.

To further validate the results of the classification in step 1, we investigated whether the classification can actually be considered as a result of a systematic answer pattern and not as a random effect. To this purpose we performed the following bootstrapping test: We randomized all 350 per-item answers for each of the 24 items which resulted in 350 *plausible* student patterns per bootstrapping draw—eliminating the hypothesized *systematic* student patterns and recognizing only item difficulty. For each of these bootstrapping draws, the Bayesian classification approach described above was conducted—resulting in 95% confidence intervals for the cluster sizes based on the assumption of all students answering completely at random (given the empirical item difficulty). We considered a hypothesized solution pattern occurring "not at random" in our dataset, if the actual cluster size is outside of the 95% CI.

For a further classification, one can take into account that students who fall into the category of using the "proficient" strategy, i.e., who solve all tasks correctly, can do so by different strategies, which can be distinguished by the problem solving time (cf. Table 1). We modelled this behavior by assuming shorter or longer response times in incongruent items than in congruent items. More specifically, we assumed for each incongruent item and each proficient strategy (see again Table 1) a Gaussian distribution of the student's problem solving time T_i, normalized with the mean time for congruent items. Throughout we applied a logarithmized time scale and therefore assumed a time ratio of 2 as typical for a longer and 0.5 as typical for a shorter response time. The likelihood of the student i having (relative) response time T_i in task j, when applying strategy k is thus

$$p\left(T_i|H^k\right) = \frac{1}{N} \cdot \exp\left(-\frac{1}{d} \cdot \left|\log(T_i) - \log(T_{j,k})\right|^2\right)$$

with $T_{j,k}$ being the assumed relative response time for each item and strategy (e.g., 0.5 for the half relative response time). The Bayesian classification proceeds as described above with continuous likelihoods. For more details on the choice of parameters, see Leuders and Loibl (2020). Again, we only consider students with an accuracy of $BF > 3$ as appropriately classified in the second step.

The bootstrapping approach described to validate the results of the classification in step 1 is from our point of view only applicable for solution rates, not response time differences—as the chosen time ratio of 2 and 0.5 do not represent the actual time ratios that have to be expected but act as reasonable cutoffs for time ratios to distinguish 'faster' or 'slower' answers. That is why we chose another validation approach to test the resulting problem-solving time patterns against chance: We varied the width of the Gaussian distribution for the normalized problem solving time to check the robustness of our Bayesian classification approach in step 2.

For the third step, we estimated individual distance effects in problem-solving time as the η^2 in per-student regression models—predicting the problem-solving time in all 24 items *only* by the distance between the two to-be-compared fractions. We then estimated plausible breakpoints in all the η^2 values using Jenks' natural breaks optimization (Rabosky et al., 2014, which can be considered a type of one-dimensional *k*-means clustering).

Table 2
Model fit indices for all four compared and theoretically applicable models for grouping the items.

Model	RMSEA	90% CI	CFI	TLI	χ^2	df	*p*
Full 2 × 2-factor model	0.034	[0.017, 0.047]	0.976	0.973	253.57	246	0.357
2-factor congruence-only model	0.035	[0.019, 0.048]	0.974	0.971	267.34	251	0.229
2-factor benchmark-only model	0.089	[0.080, 0.097]	0.832	0.815	1238.16	251	< 0.001
1-factor magnitude-only model	0.089	[0.080, 0.098]	0.830	0.814	1249.88	252	< 0.001

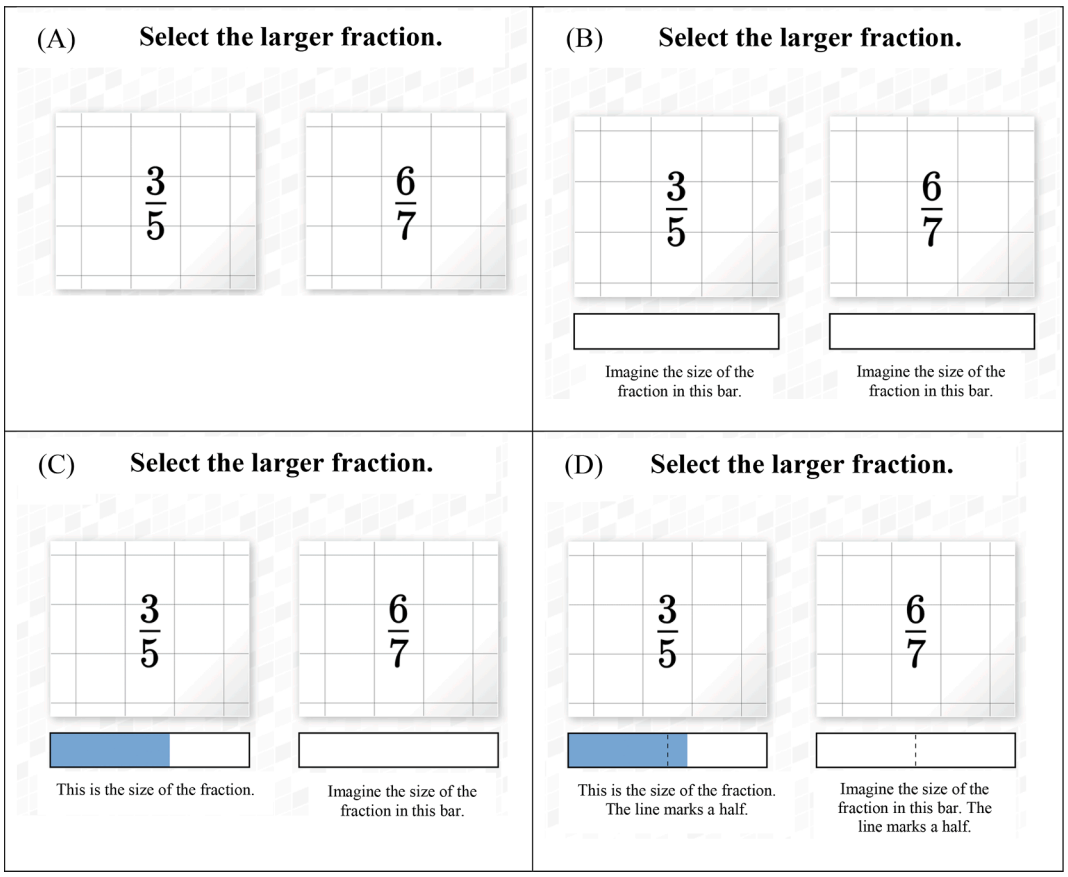


Fig. 4. Different versions of the digital environment.

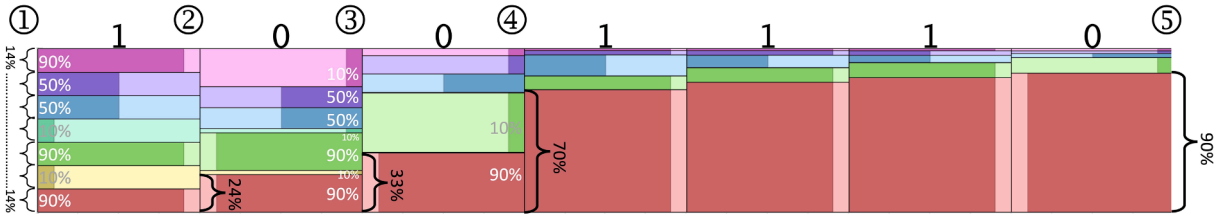


Fig. 5. Classification of example student: ① The seven bars represent the seven strategies (each colored bar representing one strategy) with equal priors (14%), represented by the equal heights of the bars. The evidence "1 = correct" for the solution of the first task is in favor of strategies 1, 5, and 7 with a likelihood of 90% for "correct solution", represented by the darker, left part of the corresponding bars. This results in a posterior of 90% / (90% + 50% + 50% + …) = 24% for these three strategies. The evidence "1 = correct" is in disfavor of strategies 4 and 6 with likelihood of 10% for "correct solution", resulting in a posterior of 3% for these two strategies. The evidence "1 = correct" has a likelihood of 50% for strategies 2 and 3 (guessing), resulting in a posterior of 13% for these two strategies. ② The posteriors of the strategies are considered priors for the next step, represented by the different heights of the seven bars in the second column. The new evidence "0 = incorrect" for the solution of the second task is in favor of strategy 5 and 7 with a likelihood 90% for "incorrect", represented by the darker, right part of the bar. This results in a posterior of 33% for strategy 5 and 7. ③ The again updated priors of the strategies are represented by the new heights of the bars. The next evidence "0 = incorrect" for the solution of the third task is in favor of strategy 7, resulting in a posterior of 70% for strategy 7. ④ The new priors of the strategies are again represented by the different heights of the bars. ⑤ The cumulative evidence leads to a strong classification of the student (posterior 90% for strategy 7 after seven tasks). Final posteriors after 24 tasks often accumulate to > 99.9%.

3. Results

3.1. Descriptive results and preliminary validation

As described above, we implemented four versions of the assessment tool which were randomly distributed throughout the sample to increase overall variance in the data via increasing saliency of magnitude or benchmarking strategies. Following our idea of a random increase in variance, we did not expect to find systematic effects of the four different versions in terms of solution rate, problem solving time, and the effects of item congruency as well as applicability of a benchmarking-at-1/2 strategy. We checked these assumption by means of two (generalized) linear mixed models for the estimated solution probability and the problem solving time.

Regarding the results presented in Table 3, we consider this assumption represented in the data. The confidence intervals presented in the table underpin that there were no considerable differences in the above mentioned effects when students who were assessed with version A (i.e., standard, non-prompted design) were compared to any of the other three versions; we consider the one significant interaction effect regarding differences in the effect of item congruency on the solution probability when comparing students from version A and version D

Table 3
Results of the (generalized) linear mixed models for the estimated solution probability and the problem solving time. The two to-be-compared models differ in the grouping predictor (see Fig. 2) and the corresponding interaction effects—with Version A as baseline. *OR* = Odds ratio describing the change in solution probability, CI = 95% Confidence interval (for the ORs, statistical significance is indicated by confidence intervals excluding 1). Est = Estimated marginal mean describing the difference in problem solving time in seconds (for estimates of time differences, statistical significance is indicated by confidence intervals excluding 0). Marginal and conditional R^2—as proposed by Nakagawa and Schielzeth (2013)—are estimates for variance explained by the fixed effects only (marginal) and the entire model including random effects (conditional).

	Estimated solution probability				Problem solving time [seconds]			
	Without grouping		With grouping		Without grouping		With grouping	
Fixed effects	*OR*	*CI*	*OR*	*CI*	*Est.*	*CI*	*Est.*	*CI*
(Intercept)	3.31	2.44–4.48	2.63	1.65–4.19	4.51	3.68–5.34	4.48	3.48–5.47
Congruency	2.62	1.82–3.78	3.51	2.18–5.65	−0.33	−1.45–0.78	−0.30	−1.49–0.90
Benchmarking	1.07	0.74–1.53	1.08	0.68–1.69	−0.59	−1.70–0.53	−0.83	−2.03–0.36
Congruency * Benchmarking	0.82	0.49–1.38	0.77	0.40–1.51	1.17	−0.41–2.75	1.45	−0.24–3.15
VersionB			1.25	0.71–2.21			0.52	−0.36–1.40
VersionC			1.29	0.75–2.22			0.01	−0.83–0.85
VersionD			1.56	0.88–2.77			−0.42	−1.31–0.47
Congruency * VersionB			0.72	0.44–1.17			−0.28	−0.97–0.41
Congruency * VersionC			0.86	0.54–1.37			0.04	−0.62–0.70
Congruency * VersionD			0.49	0.30–0.80			0.10	−0.59–0.80
Benchmarking * VersionB			0.99	0.64–1.55			0.43	−0.26–1.12
Benchmarking * VersionC			0.99	0.64–1.51			0.23	−0.43–0.89
Benchmarking * VersionD			0.99	0.63–1.55			0.33	−0.37–1.02
Congruency * Benchmarking * VersionB			1.09	0.56–2.14			−0.25	−1.22–0.73
Congruency * Benchmarking * VersionC			0.90	0.47–1.71			−0.47	−1.40–0.46
Congruency * Benchmarking * VersionD			1.33	0.68–2.63			−0.37	−1.36–0.61
Random effects	Var.		Var.		Var.		Var.	
Student (N = 350)	2.33		2.33		5.80		5.77	
Item (k = 24)	0.08		0.08		0.93		0.93	
Model fit indices								
Marginal R2	0.032		0.035		0.005		0.009	
Conditional R2	0.441		0.444		0.313		0.315	
AIC	7554.66		7565.25		46239.79		46260.65	

to be of less importance than the following results. There was neither a considerable change in variance on the student random intercepts from the models without the grouping factor (representing association with the four conditions) to the models with grouping added as a fixed effect, nor in the marginal R^2 values, nor the AICs—which was true for both solution probability and problem solving time (Table 3). In line with that, the models containing grouping—and all relevant interactions—did not result in significant better model fits than the models without the grouping factor, which was true for both solution probability, $\chi^2(12) = 13.42$, $p = .34$, and problem solving time, $\chi^2(12) = 10.93$, $p = .53$. Considering these results, we concluded that adding four different assessment versions did not result in systematic variation but rather a random increase in the variability of the data—and consequently we did not consider the grouping variable relevant for further analyses. It should be mentioned that we did find an overall significant item congruency effect, but no significant effect of benchmarking in the present sample (Table 3) which illustrates that we may expect more students demonstrating a natural number bias than benchmarking strategies when disentangling individual cognitive processes.

3.2. Disentangling individual cognitive processes in the fraction comparison task

We aimed at disentangling individual differences in the underlying cognitive processes of students when comparing two fractions. Based on the itemset developed (Fig. 1) systematic application of one of the hypothesized strategies (Table 1) would lead to a consistent behavior of each student and therefore a predictable pattern, regarding (1) solution rate, (2) problem solving time, or (3) individual distance effects.

In the first step of the classification by means of the Bayesian classification procedure described above, we considered strong effects, assessable via patterns in the solution rate. This led to a valid classification of 319 out of 350 students (Table 4, left)—with the remaining 31 students (showing a classification accuracy below $BF = 3$) not considered in all further analyses. All resulting group sizes are outside of the respective 95% CIs (see section 2.4 for a detailed

Table 4
Number of students classified into the hypothesized strategies to compare two fractions based on our Bayesian classification approach. Step 1 considered solution rates only. Step 2 considered problem solving time for the 185 students classified as capable based on their solution rates. *Md*(*BF*) = Median of the Bayes factors representing classification quality calculated as the quotient of the probabilities of the most probable classification and the second most probable classification.

	Step 1: Solution Rate		Step 2: Problem Solving Time	
Classification	*N*	*BF (Md)*	*N*	*BF (Md)*
Not classifiable	*31*	*< 3*		
Strong typical bias	67	16,522		
Strong reverse bias	18	3104		
Guess	44	51		
Benchmark, or guess	3	6		
Benchmark, or strong typical bias	2	14		
Benchmark, or strong reverse bias	0	—		
Capable solving	185	129		
Not classifiable			57	< 3
Suppressed typical bias			9	10
Suppressed reverse bias			17	10
Benchmark, or suppressed typical bias			34	11
Benchmark, or suppressed reverse bias			5	31
Proficient solving			63	28

explanation of the bootstrapping procedure; see Supplemental Material S1 for the specific results)—and thus are considered differing significantly from chance. In addition, group sizes in step 1 did not vary significantly between the four versions of the digital environment, $\chi^2(18) = 15.98$, $p = .59$ (see Supplemental Material S2 for the specific results), underpinning the assumption that the different assessment versions did not result in systematic variation.

We could confirm patterns of strong typical and reverse biases with students not differing between benchmark and no-benchmark items, but with regard to item congruency (Fig. 6, first two profiles). In addition, a very small group of 5 students demonstrated patterns with no differences in the high solution rate in benchmark items, but showing one of the three expected patterns in no-benchmark items—arguably falling back to guessing or strong (typical) bias patterns in items when benchmarking at 1/2 cannot be applied (Fig. 6, benchmarking profiles). As expected, a group of students who showed a guessing pattern could be confirmed (Fig. 6, fifth profile). Furthermore, a larger group of 185 students were capable of solving items from all four item categories to a very high extent (Fig. 6, last profile). We argue that these 185 students may be further distinguished in their underlying strategy to compare fractions in our study—yet not based on their overall high solution rates, but their response times.

In a second classification step, we considered differences in problem solving time to yield further insights into individual differences of the underlying cognitive processes while comparing fractions. We utilized the individual ratio of problem solving time between congruent and incongruent items. To classify students based on the same assumptions, we standardized problem solving time in congruent tasks per students—distinguishing between benchmark and no-benchmark items. This led to a valid classification of 128 out of 185 students (Table 3, right)—with the remaining 57 students showing a classification accuracy below $BF = 3$ and thus not considered in all further analyses. The results were robust when varying the width of the Gaussian distributions (see section 2.4 for a detailed explanation of how the distributions for normalized problem solving time were modelled; see Supplemental Material S3 for the specific results). Again, group sizes in step 2 did not vary significantly between the four versions of the digital environment, $\chi^2(15) = 9.53$, $p = .85$ (see Supplemental Material S4 for the specific results).

We could confirm patterns of suppressed typical and reverse biases with students ratio of problem-solving time of congruent and incongruent items not differing between benchmark and no-benchmark items (Fig. 7, first two profiles). Here, students repeatedly answered slower (suppressed typical bias) or faster (suppressed reverse bias) in incongruent items than in their student-centered geometric mean problem solving time in congruent items. We assumed to find patterns corresponding to benchmarking when possible, but falling back to suppressed bias patterns in no-benchmark items—i.e., no difference in problem solving time in benchmark items, but in no-benchmark items. We considered the resulting patterns of students classified as 'benchmarking or falling back to a suppressed typical bias' and students classified as 'benchmarking or falling back to a suppressed reverse bias' in line with our hypothesis (Fig. 7, third and fourth profile). Again, a larger group of 63 students were proficient in solving items from all four item categories to a very high extent and with no relevant *systematic* difference in the response times in congruent or incongruent items, both in benchmark and no-benchmark items (Fig. 7, last profile). As argued before, we think that these 63 students may be further distinguished based on the individual students' distance effects.

In a third classification step, we considered differences in individual

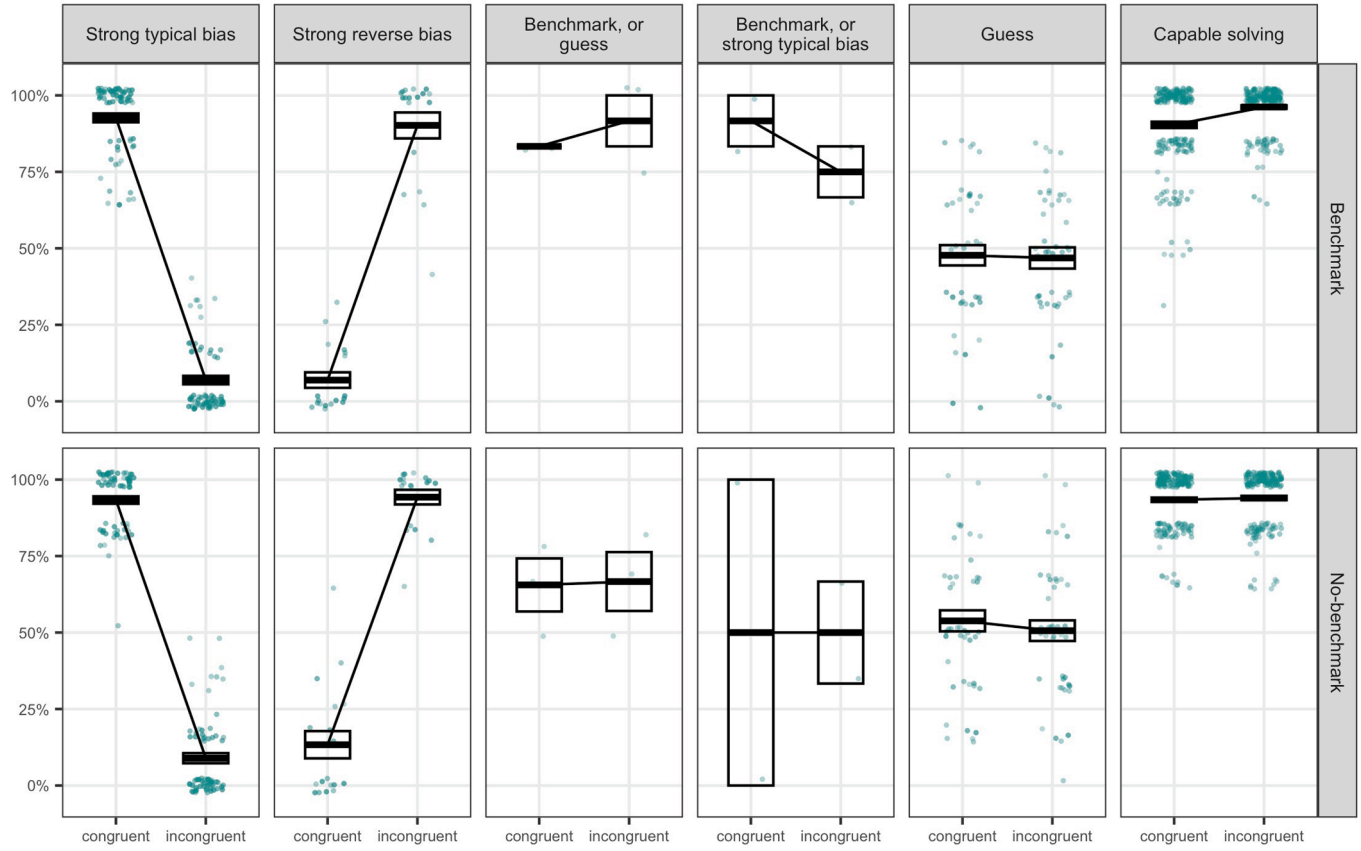


Fig. 6. Scatter plot of the solution rates in incongruent and congruent tasks in no-benchmark and benchmark contexts for students classified on the hypothesized solution rate patterns. Crossbars represent Means and ± 1 SE of all students in the profile. Note that individual scores are slightly jittered in *x*- and *y*-direction to avoid overplotting. The opacity values of the individual points represent the goodness of the classification—the darker, the better.

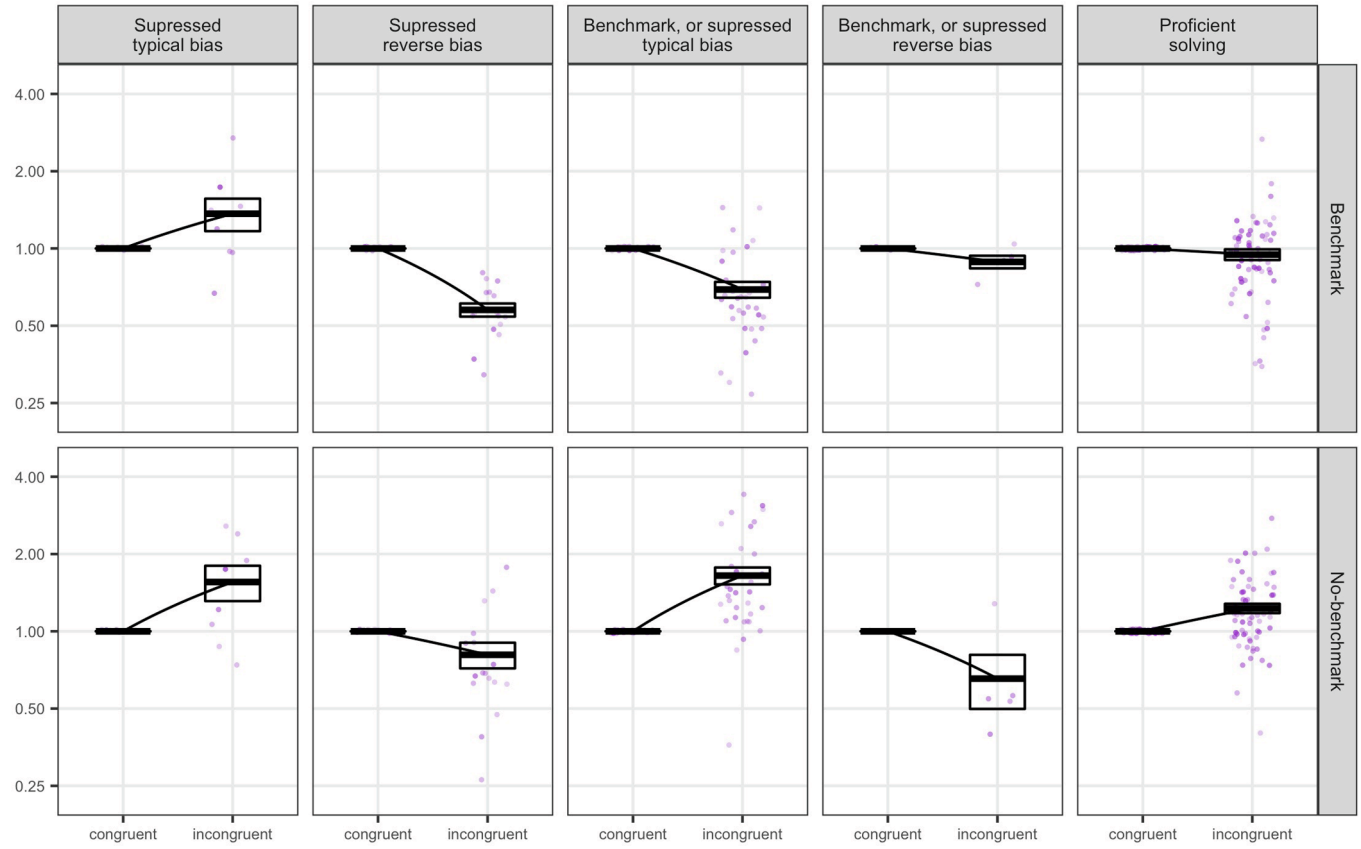


Fig. 7. Scatter plot of the standardized problem-solving time in incongruent fraction comparison items—in no-benchmark and benchmark contexts—for students classified on the hypothesized response time patterns. Values higher than 1 represent responses to the specific incongruent item given slower than to the student-centered geometric mean response time in congruent items within the item type (no-benchmark or benchmark); values between 0 and 1 represent faster responses. The time ratio (*y*-axis) is logarithmized to base 2, i.e. 'twice as fast' and 'half as slow' is given in equal distance from a ratio of 1. Crossbars represent Means and ± 1 SE of all students in the profile. Note that individual scores are slightly jittered in *x*- and *y*-direction to avoid overplotting. The opacity value of the individual points represent the goodness of the classification—the darker, the better.

distance effects on problem solving time regardless of item specifications to divide those 63 proficient students into students who systematically applied magnitude estimation (which should result in large individual distance effects) and students who solved the problems by rapid procedural arithmetic (which should result in a negligible individual distance effect). Fig. 8 shows that both groups of students could be found in our sample. Indeed, students showing a "capable solving" pattern in solution rates, and a "proficient solving" pattern in response time *did* still differ in their individual distance effect, *Mdn* = 0.018, *IQR* = 0.006–0.053 (Fig. 8). Estimating plausible breakpoints given the data lead to three plausible 'centers' at $\eta_1^2 < 0.001$, $\eta_2^2 = 0.075$, and $\eta_3^2 = 0.216$—in line with the assumption of groups of students showing a negligible, or a (medium to) large distance effect; we consider the center for the middle cluster $\eta_2^2 = 0.075$ the point that divides students demonstrating a 'pure and rapid arithmetic'-strategy versus students demonstrating a 'magnitude'-strategy.

4. Discussion

4.1. Replication and broadening the perspective on individual fraction comparison profiles

Based on the data from a broad sample of students and the systematically constructed collection of tasks, we could replicate the patterns in

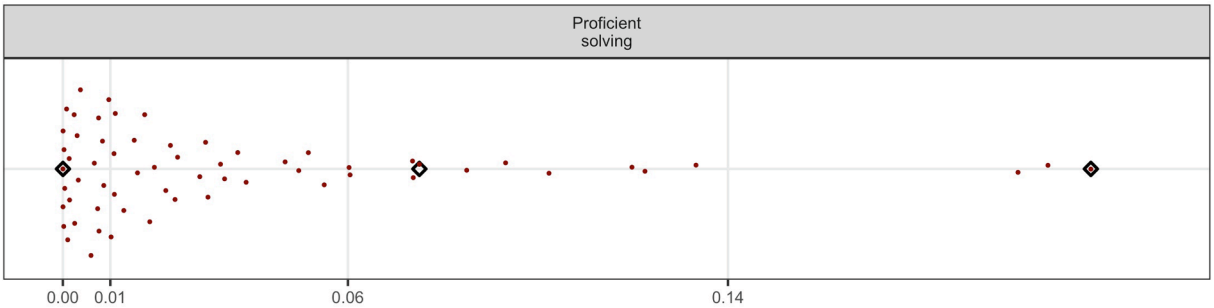


Fig. 8. Beeswarm plot of the individual effect of distance on problem solving time, given as the η^2 of a linear model of response time predicted *only* by item distance for each student. The subdivision of the effect size axis (*y*-axis) is inspired by Cohens 'rules of thumb' for interpreting the size of η^2. Black diamonds represent the results of an estimation of the optimum breakpoints given the data using Jenks natural breaks optimization.

solution rate and problem-solving times which are reported in literature—such as those connected to natural number bias (Alibali & Sidney, 2015; DeWolf & Vosniadou, 2011; Gómez & Dartnell, 2019; Ni & Zhou, 2005; Obersteiner et al., 2020; Obersteiner, Marupudi, et al., 2019; Prediger, 2008; Reinhold, Obersteiner, et al., 2020; Rinne et al., 2017; Vamvakoussi & Vosniadou, 2004; Van Hoof et al., 2013, 2018) and to benchmarking (Clarke & Roche, 2009; Fazio et al., 2016; González-Forte et al., 2019; Liu, 2018; Obersteiner et al., 2020)—in a single sample and with a non-explorative statistical procedure. Furthermore, we were able to identify further theoretically expected "mixed" strategies, such as suppressed bias strategies with or without benchmarking. Here, we argue that a distinction between *strong* bias strategies visible in solution patterns and *suppressed* bias strategies visible in response times allows to satisfy the two guiding theories in research on erroneous patterns in fraction comparison tasks, the conceptual change account (Vamvakoussi & Vosniadou, 2004) and the dual process account (Van Dooren & Inglis, 2015). Overall, we consider our findings as a support and differentiation of previous research resulting in a more integrative picture of fraction comparison strategies in particular, and rational number concepts in general.

These results draw on two major ingredients of our approach: Firstly, we based our classification of comparison strategies on a theoretical description of cognitive processes summarized from previous findings and a prediction of respective solution patterns (behavior). This approach of cognitive (diagnostic) modelling aims at a valid assessment of individual latent constructs (Leighton & Gierl, 2007). Secondly, by applying a Bayesian classification procedure, we used a person-centered approach, which allowed for a simultaneous identification of a larger number of individual strategies. Thus, our approach combines the advantages of the very commonly used approach of clustering and comparing group averages to date, that allows the identification of a variety of strategies with a confirmatory and less explorative approach by predicting and validating individual solution patterns. We believe this combination is a necessary step for advancing theory on fraction comparison.

4.2. Magnitude processing and benchmarking

Magnitude processing is a core concept in research on numerical cognition. Although it is generally agreed upon that whole number magnitude processing and fraction magnitude processing refer to comparable cognitive processes (Schneider & Siegler, 2010; Siegler et al., 2011), it is not completely clear how fraction magnitude processing occurs in specific situations. We argue that the often used operationalization of a distance effect between two to-be-compared numbers may not be completely suitable to assess fraction magnitude processing for a large group of students. The ability to utilize straddling benchmarks (Obersteiner et al., 2020) requires the ability to process fraction magnitudes—yet, not the direct comparison of the magnitudes of the two to-be-compared fractions, but, e.g., the comparisons with the straddling benchmark. Following that, we argue that the use of a distance effect as a measure of fraction magnitude processing should be applied only to those subgroups of students, for whom clear hypothesis about the specific underlying occurrence of a distance effect can be derived: In our study, we refrain from evaluating a distance effect for students demonstrating, e.g., a strong bias profile, as the most plausible underlying cognitive process resulting in such a strong bias profile neglects the comparison of the numerical distance between the two to-be-compared fractions completely—which is in line with existing empirical analyses (cf., Reinhold, Obersteiner, et al., 2020, who explicitly showed that students with a strong typical bias profile, or a strong reverse bias profile did not show a distance effect, while students with no strong bias profile did). Therefore, a distance effect—if present—may not validly inform about the cognitive processes of strongly biased students. However, picturing students with no systematic differences in solution rates and response times in congruent and incongruent fraction comparison items, it is reasonable that a distance effect on an individual level—if present—would be in line with a comparison of two fractions based on their numerical distance, rather than rapid procedural solving. Following this argument, utilizing a distance effect for the latter mentioned proficient solvers should be considered suitable to inform about the individual underlying cognitive process.

4.3. Educational and practical implications

Our work sheds light on the relevance of person-centered statistical methods in addition to sample-based quantitative methods in the assessment of students' skill in fraction comparison. In fact, students can be clustered into profiles of shared systematic errors or shared utilized strategies. Gaining empirical insights into the existence of such profiles is not only relevant for foundational research on the psychology of rational numbers, but also can serve as a basis for the research-informed development of individualized approaches to support learning: Besides 'being wrong' in fraction comparison tasks, knowledge about the underlying systematic misconceptions or the already-mastered strategies—i.e., the belonging of a specific student to one of the theoretically derived profiles—can inform about more- or less-promising learning support: While a student demonstrating a strong typical bias pattern may profit most from a conceptual change intervention deconstructing the application of natural number strategies first (e.g., by comparing fractions with the same numerator—guided by iconic representations), another student demonstrating a strong reverse bias pattern may, e.g., profit more from a reflection of the applied strategy when comparing fractions with the same denominator; for a third student demonstrating a suppressed typical bias, a conceptual change intervention may not be beneficial, but prompts pointing to a broad variety of different strategies (to promote a strategy repertoire) may aid his or her individual needs (Reinhold & Reiss, 2020).

While the identification of such profiles may also be achieved using commonly-used explorative statistical approaches (i.e., *k*-means or hierarchical clustering, or latent profile analysis), the confirmatory approach via Bayesian classification (as proposed in the present study) has several practical implications itself. Since the approach is not explorative in nature, strong theoretical hypothesis about the different cognitive processes (that lead to different answer patterns) are necessary; only after these cognitive processes are identified, one can look for indicators in student behavior that may result in identifying the respective profile as result of a specific type of misconception or erroneous strategy. This process (in contrast to the application of an explorative statistical method) is structurally similar to a teachers' diagnostic judgments (Loibl & Leuders, 2021)—and may therefore be used to foster teachers' diagnostic competences regarding fractions.

4.4. Limitations and future research

The specific construction of items and the variance among the students lead to a broad spectrum of identifiable strategies. However, these strategies may still be specific for a certain level of fraction knowledge and for a specific curriculum. In order to largen the support for our findings, students from other cultural and curricular contexts, different grades, and from various school types should be investigated. Especially, among those who are not proficient and rely on various misconceptions, there may be more solution patterns which might not be captured by our model. Another approach to strengthen our assumptions on comparison strategies could be to test students before and after interventions which differentially foster certain strategies. Following this idea, interventions that seek to strengthen the use of straddling benchmarks to compare two fractions could be developed to increase the number of students who systematically apply this strategy—which were only present in rather small subgroups in our sample, even though we derived stimuli that aimed at trigger the use of the straddling benchmark 1/2 (see Fig. 4).

We have stressed the confirmatory characteristic of our approach

and showed that it lead to highly certain classifications. However, it also has some drawbacks: Modelling the prediction of student behavior relies on the choice of certain likelihoods (the conditional error probabilities and the gaussian distributions of problem solving time). When these are not known precisely (which would require an alternative reliable procedure of determining the latent constructs), one can only apply sensible approximations. One can still test for the dependance of the results on parameter values, which we did (e.g., 0.3 and 0.7 instead of 0.1 and 0.9 as low and high likelihoods) and only found small displacements in the number of non-classifications. Furthermore, with our approach, one can only identify strategies which have been modelled theoretically before. Only when results deviate from a strong classification certainty one can hypothesize that the student does not apply one of the modelled strategies consistently. Also, our design does not suffice to distinguish certain alternative theories on the interplay of magnitude estimation and proficient symbolic processing, as indicated in the theory section. This distinction would require even more elaborate designs. In order to better connect our results to existing research, one should strive to combine different measurement approaches (e.g., think-aloud interviews, ore eye-tracking) and thus cross-validate the overlapping assumptions on students' comparison strategies. However, since the cognitive processes are rather implicit and on a short timescale one has to carefully design such a validation approach.

Moreover, our approach to analyze the given data does allow for systematic application of a certain strategy *within* one category of items and a change in strategies *between* categories of items, but not for a (sophisticated or random) application of different strategies *within* a specific category of items. This seems relevant to note, since this may be a valid explanation for the unclassified students in step 2 of our analysis—and a recent work of Obersteiner et al. (2022) showed that such changes in systematic strategies between items of one category occur in adults and thus could also be expected to some degree in individual students. Here, retrospective interview studies, eye-tracking, or other process assessment may shed light on the still missing pieces of information about the cognitive processes relevant for fraction comparison tasks.

Such (in)consistent uses of strategy within specific item categories may be related with—not only—the number of unclassified students when looking at response times in classification step 2, but also with the arguably lower classification quality when focusing on response times, than solution rates, in general. It should be noted that this may also be due to our choice of relatively easy, non-demanding fraction comparison items. Here, more complex items with two digit numerators or denominators might be used and integrated into our framework.

4.5. Conclusions

All in all, our effort to cognitively conceptualize and empirically identify students' strategies in fraction comparison within one sample in our view contributes to a more integrated picture of the phenomena encountered in many previous studies. Furthermore, the methodological approach, when combined with data from more intricate test designs, may give even more insights into the cognitive substrate underlying fraction comparison.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Frank Reinhold: Project administration, Visualization, Writing – review & editing, Investigation (Data collection), Resources (Material), Writing – original draft, Validation, Formal analysis, Software (Assessment), Conceptualization, Methodology, Visualization. **Timo Leuders:** Validation, Formal analysis, Conceptualization, Methodology, Software (Statistical Analysis), Writing – review & editing. **Katharina Loibl:** Conceptualization, Validation, Formal analysis, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The software used for the fraction comparison task used in this study was developed in the ALICE:fractions project (Stefan Hoch, Frank Reinhold, Bernhard Werner, Kristina Reiss, & Jürgen Richter-Gebert) and customized by Stefan Hoch for the purpose of the present study.

The Bayesian classification algorithm is based on CindyScript (Richter-Gebert & Kortenkamp, 2012), wich is made available and supported by Jürgen Richter-Gebert and Ulrich Kortenkamp, see https://www.cinderella.de. We want to thank the developers for their ongoing support.

This study was approved by both the Bavarian State Ministry of Education (IV.7-BO4106.2019/52/9) and the Freiburg District Council Department of Schools and Education (7-6499.2). We want to thank all students for taking part in the study.

We acknowledge support by the Publication Fund of the University of Education Freiburg.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2023.102224.

References

Alibali, M. W., & Sidney, P. G. (2015). Variability in the natural number bias: Who, when, how, and why. *Learning and Instruction, 37*, 56–61. https://doi.org/10.1016/j.learninstruc.2015.01.003

Behr, M. J., Lesh, R. A., Post, T. R., & Silver, E. (1983). Rational Number Concepts. In R. Lesh, & M. Landau (Eds.), *Acquisition of Mathematics Concepts and Processes* (pp. 91–125). Academic Press.

Behr, M. J., Post, T. R., & Wachsmuth, I. (1986). Estimation and children's concept of rational number size. In H. Schoen, & M. Zweng (Eds.), *Estimation and mental computation* (pp. 101–111). NCTM.

Bonn, C. D., & Cantlon, J. F. (2017). Spontaneous, modality-general abstraction of a ratio scale. *Cognition, 169*, 36–45. https://doi.org/10.1016/j.cognition.2017.07.012

Clarke, D. M., & Roche, A. (2009). Students' fraction comparison strategies as a window into robust understanding and possible pointers for instruction. *Educational Studies in Mathematics, 72*(1), 127–138. https://doi.org/10.1007/s10649-009-9198-9

Culbertson, M. J. (2016). Bayesian Networks in Educational Assessment: The State of the Field. *Applied Psychological Measurement, 40*(1), 3–21. https://doi.org/10.1177/0146621615590401

DeWolf, M., & Vosniadou, S. (2011). The whole number bias in fraction magnitude comparisons with adults. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1751–1756). Cognitive Science Society.

DeWolf, M., & Vosniadou, S. (2015). The representation of fraction magnitudes and the whole number bias reconsidered. *Learning and Instruction, 37*, 39–49. https://doi.org/10.1016/j.learninstruc.2014.07.002

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning, 29*(2/3), 103–130. https://doi.org/10.1023/A:1007413511361

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

Fazio, L. K., DeWolf, M., & Siegler, R. S. (2016). Strategy use and strategy choice in fraction magnitude comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(1), 1–16. https://doi.org/10.1037/xlm0000153

Flunger, B., Trautwein, U., Nagengast, B., Lüdtke, O., Niggli, A., & Schnyder, I. (2017). A person-centered approach to homework behavior: Students' characteristics predict

their homework learning type. *Contemporary Educational Psychology, 48*, 1–15. https://doi.org/10.1016/j.cedpsych.2016.07.002

Gómez, D., & Dartnell, P. (2019). Middle Schoolers' Biases and Strategies in a Fraction Comparison Task. *International Journal of Science and Mathematics Education, 17*(6), 1233–1250. https://doi.org/10.1007/s10763-018-9913-z

González-Forte, J. M., Fernández, C., Van Hoof, J., & Van Dooren, W. (2019). Exploring Students' Reasoning About Fraction Magnitude. In M. Graven, H. Venkat, A. Essien, & P. Vale (Eds.), *Proceedings of the 43rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 272–279). PME.

González-Forte, J. M., Fernández, C., Van Hoof, J., & Van Dooren, W. (2023). Incorrect Ways of Thinking About the Size of Fractions. *International Journal of Science and Mathematics Education, 21*(7), 2005–2025. https://doi.org/10.1007/s10763-022-10338-7

Heck Ribeiras, P., Wittmann, G., & Obersteiner, A. (2022). In welcher Weise unterstützen Schulbücher Vorstellungsumbrüche beim Lernen von Bruchzahlen? Sine Schulbuchanalyse [In what ways do textbooks support conceptual change in learning fractions? A textbook analysis]. *mathematica didactica*, *45*. https://doi.org/10.18716/OJS/MD/2022.1595.

Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: A Systematic Review. *Acta Informatica Medica, 24*(5), 364. https://doi.org/10.5455/aim.2016.24.364-369

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications* ((1st ed.).). Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.

Leuders, T., & Loibl, K. (2020). Processing Probability Information in Nonnumerical Settings – Teachers' Bayesian and Non-bayesian Strategies During Diagnostic Judgment. *Frontiers in Psychology, 11*, 678. https://doi.org/10.3389/fpsyg.2020.00678

Liu, F. (2018). Mental representation of fractions: It all depends on whether they are common or uncommon. *Quarterly Journal of Experimental Psychology, 71*(9), 1873–1886. https://doi.org/10.1080/17470218.2017.1366532

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction, 62*, 1–10. https://doi.org/10.1016/j.learninstruc.2019.03.002

Loibl, K., & Leuders, T. (2021). Modeling Teachers' Diagnostic Judgments by Bayesian Reasoning and Approximative Heuristics. *RISTAL, 4*, 88–108. https://doi.org/10.23770/rt1844

Lortie-Forgues, H., Tian, J., & Siegler, R. S. (2015). Why is learning fraction and decimal arithmetic so difficult? *Developmental Review, 38*, 201–221. https://doi.org/10.1016/j.dr.2015.07.008

Matthews, P. G., & Chesney, D. L. (2015). Fractions as percepts? Exploring cross-format distance effects for fractional magnitudes. *Cognitive Psychology, 78*, 28–56. https://doi.org/10.1016/j.cogpsych.2015.01.006

Meert, G., Grégoire, J., & Noël, M.-P. (2010). Comparing the magnitude of two fractions with common components: Which representations are used by 10- and 12-year-olds? *Journal of Experimental Child Psychology, 107*(3), 244–259. https://doi.org/10.1016/j.jecp.2010.04.008

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Ni, Y., & Zhou, Y.-D. (2005). Teaching and Learning Fraction and Rational Numbers: The Origins and Implications of Whole Number Bias. *Educational Psychologist, 40*(1), 27–52. https://doi.org/10.1207/s15326985ep4001_3

Obersteiner, A., Alibali, M. W., & Marupudi, V. (2020). Complex fraction comparisons and the natural number bias: The role of benchmarks. *Learning and Instruction, 67*, Article 101307. https://doi.org/10.1016/j.learninstruc.2020.101307

Obersteiner, A., Alibali, M. W., & Marupudi, V. (2022). Comparing fraction magnitudes: Adults' verbal reports reveal strategy flexibility and adaptivity, but also bias. *Journal of Numerical Cognition, 8*(3), 398–413. https://doi.org/10.5964/jnc.7577

Obersteiner, A., Dresler, T., Bieck, S. M., & Moeller, K. (2019). Understanding Fractions: Integrating Results from Mathematics Education, Cognitive Psychology, and Neuroscience. In A. Norton, & M. W. Alibali (Eds.), *Constructing Number* (pp. 135–162). Springer International Publishing. https://doi.org/10.1007/978-3-030-00491-0_7.

Obersteiner, A., Marupudi, V., & Alibali, M. W. (2019). Adults' strategy use in fraction comparison. In M. Graven, H. Venkat, A. Essien, & P. Vale (Eds.), *Proceedings of the 43rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 153–160). PME.

Obersteiner, A., & Tumpek, C. (2016). Measuring fraction comparison strategies with eye-tracking. *ZDM, 48*(3), 255–266. https://doi.org/10.1007/s11858-015-0742-z

Obersteiner, A., Van Hoof, J., & Verschaffel, L. (2013). Expert mathematicians' natural number bias in fraction comparison. In A. M. Lindmeier, & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education, 3* pp. 393–400). PME.

Post, T., Behr, M., & Lesh, R. (1986). Research-Based Observations About Children's Learning of Rational Number Concepts. *Focus on Learning Problems in Mathematics, 8*(1), 39–48.

Prediger, S. (2008). The relevance of didactic categories for analysing obstacles in conceptual change: Revisiting the case of multiplication of fractions. *Learning and Instruction, 18*(1), 3–17. https://doi.org/10.1016/j.learninstruc.2006.08.001

Rabosky, D. L., Grundler, M. C., Anderson, C. J., Title, P. O., Shi, J. J., Brown, J. W., … Larson, J. G. (2014). BAMMtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution, 5*, 701–707.

Reinhold, F., Hoch, S., Werner, B., Richter-Gebert, J., & Reiss, K. (2020). Learning Fractions with and without Educational Technology: What Matters for High-Achieving and Low-Achieving Students? *Learning and Instruction, 65*, Article 101264. https://doi.org/10.1016/j.learninstruc.2019.101264

Reinhold, F., Obersteiner, A., Hoch, S., Hofer, S. I., & Reiss, K. (2020). The Interplay Between the Natural Number Bias and Fraction Magnitude Processing in Low-Achieving Students. *Frontiers in Education, 5*, 29. https://doi.org/10.3389/feduc.2020.00029

Reinhold, F., & Reiss, K. (2020). Anschauliche Wege zum Größenvergleich von Brüchen [Descriptive methods for comparing the size of fractions]. *Zeitschrift für Mathematikdidaktik in Forschung und Praxis*, 1. https://doi.org/10.48648/vp5k-6360.

Richter-Gebert, J., & Kortenkamp, U. H. (2012). The Cinderella.2 Manual – Working with the Interactive Geometry Software. Springer. https://doi.org/10.1007/978-3-540-34926-6.

Rinne, L. F., Ye, A., & Jordan, N. C. (2017). Development of fraction comparison strategies: A latent transition analysis. *Developmental Psychology, 53*(4), 713–730. https://doi.org/10.1037/dev0000275

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science, 20*(3), e12372. https://doi.org/10.1111/desc.12372

Schneider, M., & Siegler, R. S. (2010). Representations of the magnitudes of fractions. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1227–1238. https://doi.org/10.1037/a0018170

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

Sprute, L., & Temple, E. (2011). Representations of Fractions: Evidence for Accessing the Whole Magnitude in Adults. *Mind, Brain, and Education, 5*(1), 42–47. https://doi.org/10.1111/j.1751-228X.2011.01109.x

Vamvakoussi, X., Van Dooren, W., & Verschaffel, L. (2012). Naturally biased? In search for reaction time evidence for a natural number bias in adults. *The Journal of Mathematical Behavior, 31*(3), 344–355. https://doi.org/10.1016/j.jmathb.2012.02.001

Vamvakoussi, X., & Vosniadou, S. (2004). Understanding the structure of the set of rational numbers: A conceptual change approach. *Learning and Instruction, 14*(5), 453–467. https://doi.org/10.1016/j.learninstruc.2004.06.013

Van Dooren, W., & Inglis, M. (2015). Inhibitory control in mathematical thinking, learning and problem solving: A survey. *ZDM, 47*(5), 713–721. https://doi.org/10.1007/s11858-015-0715-2

Van Hoof, J., Degrande, T., Ceulemans, E., Verschaffel, L., & Van Dooren, W. (2018). Towards a mathematically more correct understanding of rational numbers: A longitudinal study with upper elementary school learners. *Learning and Individual Differences, 61*, 99–108. https://doi.org/10.1016/j.lindif.2017.11.010

Van Hoof, J., Lijnen, T., Verschaffel, L., & Van Dooren, W. (2013). Are secondary school students still hampered by the natural number bias? A reaction time study on fraction comparison tasks. *Research in Mathematics Education, 15*(2), 154–164. https://doi.org/10.1080/14794802.2013.797747