International Electronic Journal of Mathematics Education
2022, 17(1), em0667
e-ISSN: 1306-3030
https://www.iejme.com

Research Article

OPEN ACCESS

The Integer Test of Primary Operations: A Practical and Validated Assessment of Middle School Students' Calculations with Negative Numbers

Julie Nurnberger-Haag [1]*, Joseph Kratky [1], Aryn C. Karpinski [1]

[1] Kent State University, Kent, OH, USA
***Corresponding Author:** jnurnber@kent.edu**

Citation: Nurnberger-Haag, J., Kratky, J., & Karpinski, A. C. (2022). The Integer Test of Primary Operations: A Practical and Validated Assessment of Middle School Students' Calculations with Negative Numbers. *International Electronic Journal of Mathematics Education, 17*(1), em0667. https://doi.org/10.29333/iejme/11471

ARTICLE INFO

Received: 3 Sep. 2021

Accepted: 31 Oct. 2021

ABSTRACT

Skills and understanding of operations with negative numbers, which are typically taught in middle school, are crucial aspects of numerical competence necessary for all subsequent mathematics. To more swiftly and coherently develop the field's understanding of how to foster this critical competence, we need shared measures that allow us to compare results across studies with diverse populations and theoretical perspectives. Yet, to date no validated instrument exists to assess all four primary operations (addition, subtraction, multiplication and division). Thus, we conducted a Rasch analysis of the Integer Test of Primary Operations (ITPO) with 187 middle school students to provide a valid and reliable assessment with good person and item fit. The implications of this study are numerous for multiple stakeholders including scholars, test and textbook developers, as well as teachers. First, we validated three forms of the ITPO to foster future longitudinal studies of how integer arithmetic knowledge is maintained or decays as well as how such knowledge might be related to success in STEM disciplines. Second, our analysis provides trustworthy insights about relative difficulty of integer problem structures because regardless of test form similar problem structures loaded together. For instance, sums of additive inverses were the easiest structure, whereas division by -1 was more difficult than multiplying or dividing by any other integer. We discuss each of these and other findings that have practical implications for learning and teaching integers. Third, for broader mathematics assessments in which minimal items can be included to measure integer knowledge, this study informs which items would serve the intended assessment purpose. Finally, we provide the three forms as an appendix in printable formats to ensure these validated tests are practical to implement for teachers as well as scholars.

Keywords: integer arithmetic, negative numbers, assessment, Rasch analysis, addition, subtraction, multiplication, division, middle school

INTRODUCTION

Integer operations, which are typically taught in middle school, are crucial aspects of numerical competence necessary for all subsequent mathematics. Moreover, competence with negative numbers is necessary for all subsequent science, technology, engineering, and mathematics (STEM) courses or fields. Consider just a few examples, such as chemical reactions of positive and negative charges and vectors in physics. Calculations with negative numbers are counterintuitive, particularly subtracting a negative number and multiplying or dividing two negative integers, because each of these could or do result in a positive solution (Fischbein, 1987; French, 2001). Numerous studies have focused on the operations of addition and subtraction with negative numbers (Bishop et al., 2014; Pettis & Glancy, 2015; Stephan & Akyuz, 2012; Thompson & Dreyfus, 1988; Tsang et al., 2015). In fact, an entire edited book was published that focused exclusively on studies of these two operations (Bofferding & Wessman-Enzinger, 2018).

With whole number operations, children experience a protracted period from preschool through second grade focused on addition and subtraction before beginning foundations of multiplication and division (National Governors Association Center for Best Practices [NGA] & Council of Chief State School Officers [CCSSO], 2010). However, this is not the case for integer arithmetic learning. In the U.S. Standards students are to learn and master all primary operations with negative numbers in middle school, and specifically by the end of seventh grade (NGA & CCSSO, 2010). In practice, this is usually done within a single unit early in the beginning of the school year. Thus, future research needs to understand integer learning more broadly than addition and subtraction studies.

Constructs such as the meaning of negative numbers, ordering numbers, primary operations on two integers (addition, subtraction, multiplication and division), operating with more than two integers (e.g., $^{-}4 - 3 + ^{-}5$), meaning of negative signs as opposite operations, graphing on number lines, understanding and applying negative numbers to real-life contexts such as temperature and debt are just a few of the constructs to include and better specify (NGA & CCSSO, 2010; Vlassis, 2008). As a field, we have not yet determined the full canon of what constitutes integer arithmetic knowledge, although some work was begun in a discussion group session at the conference of the International Group for the Psychology of Mathematics Education (Bofferding et al., 2014). This single paper does not intend to solve this quandary about which multiple dimensions are necessary to claim mastery of integer knowledge. The purpose of this study is to ensure the field has at least one measure with good psychometric properties (e.g., reliability and validity) for each of the four primary operations using binary numbers. Rasch analysis is uniquely suited for this purpose (Callingham & Bond, 2006; Shaw, 1991).

Measures Used in Prior Integer Studies

Studies of integer learning have often been conducted using qualitative methods such as task-based interviews, because mathematics education currently favors these approaches (Callingham & Bond, 2006). Such research methods provide crucial insights about nuances, depth, and processes of student thinking. To supplement these more in-depth analyses and to make generalizable interpretations about integer knowledge, reliable and valid instruments that are practical for scholars as well as teachers are needed.

In terms of large-scale assessments, Ryan and Williams (2007) included three integer operations items on their broader assessment of mathematics with 15,000 four- to fifteen-year-olds. Including only three items was appropriate for an assessment in which integer constructs were just a small aspect of the study purpose. However, when researchers, large-scale test designers or textbook developers make decisions about what items to include, it is crucial to have psychometric information about the kinds of items representing each construct. A measure that assesses integer calculation with the four primary operations of addition, subtraction, multiplication, and division satisfies this purpose.

Given the dearth of integer knowledge measures, researchers often develop their own measures in order to conduct the study they intend to do. Liebeck (1999) was the only study prior to Nurnberger-Haag (2015, 2018, 2020) that compared student learning with a chip model to a number line model, so it has been extensively cited. Liebeck's (1999) posttest only study conclusions are baseless, however, due to multiple issues with the study design that introduced many threats to validity (e.g., no pretest was used, teacher was conflated with instructional model, lessons were not parallel, no random assignment of class or student to method) as well as potential problems with the measure used to compare student knowledge after learning with a particular model. Liebeck's (1999) measure consisted of just 10 addition and subtraction problems with two or three terms. Moreover, only the digits 2 and 3 were used (e.g., $2 - 3$; $^{-}3 - ^{-}2$; $3 - 2 + ^{-}3$) with the faulty assumption that this could represent students' integer operation skill with any number. No psychometric analyses were reported for this measure.

The only existing measure on integer primary operations with some psychometric information that we have found is a 24-item test of integer subtraction conducted in Malaysia (Periasamy & Zaman, 2009). This test was thorough in the sense that the face validity was addressed (Crocker & Algina, 2008), because a) they consulted teachers that these problems were of the types taught in school and b) every possible permutation of subtraction of two integers was accounted for in terms of the structure of a smaller absolute value being first or second and the use of single digit as well as double-digit integers (Periasamy & Zaman, 2009). On the other hand, similar to Liebeck (1999), a limitation of their items was that all these items used consistent absolute values such that students could use a previous item to determine another item. For instance, for an item such as $2 - 5$ on Periasamy and Zaman's (2009) test, a test-taker might be more likely to reason that the answer to $2 - ^{-}5$ should be different than if the items had the same structure but different numbers, such as $4 - 8$ and $3 - ^{-}7$. The reliability of the dichotomous (i.e., incorrect and correct) open response items were analyzed using Kuder-Richardson (KR20), which was 0.92. However, this was only conducted on the pilot test data (N=35). In the main findings of the study with 124 participants, only descriptive statistics of item inaccuracy rates were reported rather than the person- and item-fit statistics and item-person map, as is customary with Rasch analysis (e.g., Bond & Fox, 2015; Callingham & Bond, 2006). A 24-item measure that only assesses the single operation of subtraction once expanded to all primary operations would be impractical for research or classrooms if each of the four operations required 24 items. Moreover, this number of items per construct would be still more unwieldy if part of a broader test of integer knowledge that included constructs other than binary operations. Nevertheless, Periasamy and Zaman (2009) provided important foundational insights for the field to design a more practical and psychometrically rigorous instrument. They were the first to develop a stand-alone measure that attended to every possible permutation of mathematical structure of for any integer operation such that patterns could begin to be discerned.

As noted previously, integer arithmetic tends to be taught as four separate operations (i.e., addition, subtraction, multiplication, division). However, disciplinary conceptions of arithmetic involve two groups of operations (i.e., addition and multiplication). That is, as one example, addition and subtraction can be considered "strands" of the same dimension (i.e., subtraction is the same as addition of the opposite of a value). Thus, any instrument designed to assess ability with integer arithmetic may have a multidimensional structure and should be evidenced prior to conducting any analyses in subsequent studies or using the instrument in the classroom.

Purpose of Study

The focus of this study was to provide evidence of the psychometric properties of the Integer Test of Primary Operations (i.e., addition, subtraction, multiplication and division of two integers). In particular, the items were designed as three parallel forms to afford future research designs about instructional methods for integer learning to assess student knowledge prior to instruction, after instruction, and at some later time point. As opposed to the common pretest with immediate posttest designs

common in psychology as well as in mathematics education, research that investigates longer term knowledge is necessary to have any meaningful understanding of how integer instruction impacts student knowledge on time scales that matter (Nurnberger-Haag, 2018). A reliable and valid instrument that is useful for researchers while also being practical for classroom teachers to also use to assess learning is crucial.

To date, no instrument exists that assesses students' achievement with negative numbers and integer arithmetic with good reliability and validity evidence across multiple studies and populations. Rasch Analysis is uniquely suited to this endeavor (Shaw, 1991). As Callingham and Bond (2006) noted it is interesting how infrequently quantitative approaches are used in our field that educates about quantitative ideas. Rasch analyses are important for the development of instruments in ways that might seem extensively quantitative to scholars just becoming acquainted with this method, but this statistical approach is amenable to being informed by and informing qualitative considerations (Callingham & Bond, 2006; Shaw, 1991).

Accordingly, this paper addressed the primary research question: Can the three forms of the Integer Test of Primary Operations be used as a valid and reliable measure of middle school students' skill with adding, subtracting, multiplying and dividing integers? To answer this, we asked: (RQ1) "What is the dimensional structure of the ITPO?" To satisfy practical purposes, if RQ1 was satisfied then additional questions were planned about the integer problem structures: (RQ2) After instruction which integer problem structures did middle school students find most difficult or easiest? (RQ2a) Is -1(X) or X divide -1 a more difficult construct than other multiplication/ division items? (RQ2b) How difficult were additive inverse items? (RQ2c) Were items involving subtracting a negative number the most difficult items?

METHODOLOGY

Measure Development— Integer Test of Primary Operations (ITPO)

There are many constructs of integer knowledge to consider. Thus, the Integer Arithmetic Test used in a prior study (Nurnberger-Haag, 2015) consisted first of mixed addition and subtraction items, then mixed multiplication and division, then ordering items, generating additive inverses, and opposite operation items. Although there are also many other constructs that compose integer knowledge, an assumption of Rasch analysis is that it measures a single construct (i.e., unidimensionality, Bond & Fox, 2015). Thus, this study focused on providing a valid and reliable assessment of the primary operations that began the larger test of 36 items (i.e., 10 integer addition items, 10 integer subtraction items, eight multiplication items, eight division items) in its initial development phase. The problem structures were identified in a table to provide the permutations of operations on negative numbers. Some items on a given form were adopted from prior research with integer arithmetic (Liebeck, 1990; Periasamy & Zaman, 2009; Ryan & Williams, 2007) and then problems were created using other integers that maintained the same problem structure. To increase the chances that the assessment measured negative number understanding rather than underestimating this knowledge due to mistakes of whole number calculations, based on practical teaching experience specific numbers were chosen (e.g., multiplying by 5 or 2 is easier than multiplying by 7). Each addition and subtraction binary item from Liebeck (1990) was used an anchor item that provided consistency across forms (Q2, Q5, Q11, and Q13). Recall that these items all have integers with an absolute value of two or three (Liebeck, 1990). Items from the Periasamy & Zaman (2009) subtraction test were used that contained at least one negative integer.

Content validity (Fowler, 2013) was independently addressed by two practitioners in the field and congruency was met regarding the operations and variation of problem structures within form and consistency of item problem structure across forms. Face validity (Crocker & Algina, 2008) was also addressed during recruitment of participating schools in which administrators and teachers confirmed that these problems were covered in their curriculum as aspects of integer knowledge.

The instrument development process was determined based on the intention to use Classical Test Theory (CTT) in a quasi-experimental study of integer instructional methods (Nurnberger-Haag, 2015), so the instrument was piloted in four phases using factor analysis at each phase to eliminate items that were not performing as expected. After the fourth and final phase of development, the instrument had been piloted with 388 students and reduced to 35 primary operation items. The current study analyzed the items that were consistent across all phases of development (n = 31). Test questions were open-ended and then dichotomously scored (i.e., 0 = Incorrect; 1 = Correct).

Settings, Participants and Procedures

School districts that used multiple methods (e.g., a chip model, number line, and real-life contexts) were recruited to ensure that the test participants had experienced these typical instructional practices. The participating public Midwestern school was selected because all grade 7 students in that school experienced multiple integer models as part of their normal instruction. Participants (N = 187) came from all grade 7 classes that were taught by two female teachers. The principal investigator administered the ITPO to each class during each teacher's regular class periods with the teacher present. Human subjects protocols were followed and for this study students remained completely anonymous to the researcher in that no identifying information was collected. Two students left all answers of the ITPO blank, so they were removed from further analysis. Additionally, extreme cases are inestimable in Rasch analysis, so these were eliminated from the final analysis sample (Linacre, 1994). Sixteen extreme cases of students performing at floor or ceiling (i.e., summary scores of zero or 31, meaning all incorrect or correct responses, respectively) were identified and removed.

Test form and Teacher Group Comparisons

The remaining participants (N = 169) were split into groups based on teacher (Teacher One, Teacher Two) and test form (A, B, and C). These groups were compared to ensure these samples could be combined into one final analysis sample and verify that the three forms could be treated as the same test. A 2 X 3 Factorial Analysis of Variance (ANOVA) with total ITPO score as the dependent variable was used to compare groups. All assumptions for ANOVA were met. There were no significant differences between teachers (F [1,163] = 2.661, p = .105) or test forms (F [2,163] = 2.764, p = .798) on total ITPO scores. Additionally, the interaction between teachers and test forms was not significant (F [5,163] = 2.764, p = .066). Thus, the groups were combined for further analysis.

Data Analysis

Descriptive analysis – including missing data and extreme values (i.e., participants with scores of zero or perfect scores on the ITPO) – was performed using the Statistical Package for the Social Sciences (SPSS) software (version 25). Dimensionality analysis was performed using Rasch Principal Components Analysis of Residuals via Winsteps (version 4.4.5). Rasch Analysis (the Rasch Dichotomous Model) via Winsteps was used to analyze the internal structure of the identified components.

Rasch Principal Components Analysis of Residuals (PCAR) was used to identify components (called contrasts) that exist within the ITPO. PCAR contrasts depict sets of items orthogonal to the Rasch dimension (Linacre, 2019). Within a contrast, items are grouped into (three) clusters and compared based on their item loading. Typically, an identified cluster of strong positive or strong negative loadings within a contrast may represent an additional dimension – to the Rasch dimension – impacting items in the measure. Eigenvalues are used to measure the "strength" of a contrast. Linacre (2019) states that contrasts with eigenvalues less than two likely represent the random "noise" expected in the Rasch Model. Thus, any contrast with an eigenvalue greater than two was analyzed for consideration as an additional dimension to the Rasch dimension. Subsequently, items were separated based on analysis of contrasts and Rasch analysis was performed on these dimensions independently.

The Rasch Dichotomous Model was used to analyze data addressing the parameters of Person Ability and Item Difficulty. Person ability and item difficulty are measured concurrently and are typically observed on an Item-Person (i.e., Wright) map. A general rule-of-thumb for reading an Item-Person map is that persons measured below a given item's difficulty are more likely to answer that item incorrectly. Items measured below a person's ability tend to be easier for persons of that ability level.

Person ability and item difficulty summary statistics were reported. Winsteps identifies the lower (i.e., "real") and upper (i.e., "model") values of summary statistics, suggesting a range of values the true statistic may have (Linacre, 2019). The Root-Mean-Square Error (RMSE) index is used to determine the amount of error in the data. Person and item separation measures indicate how well the instrument can distinguish between person ability levels (or strata) and item difficulty levels (Linacre, 2013). Person reliability is akin to traditional test reliability in Classical Test Theory, and item reliability has no comparative traditional measure (Linacre, 2019). Both are measures of reproducibility of results (i.e., position on the Item-Person map).

Infit and outfit Mean Square Fit Statistics (MNSQ) were used to determine if any item or person was misfitting in the Rasch model. MNSQ values below 0.5 or above 1.5 indicate items or persons that may be unproductive for or degrading to the measurement of the construct (Wright & Linacre, 1994). Misfitting items or persons were further scrutinized, primarily based on point-measure correlations. Point-measure correlations below 0.3 are considered problematic (Linacre, 2019). Finally, model fit (i.e., log-likelihood χ^2; Global Root-Mean-Square Residual [RMSR]) and dimension reliability (i.e., Kuder Richardson Formula 20 [KR-20]; Kuder & Richardson, 1937) were analyzed.

RESULTS

Descriptives, Missing Data, and Extreme Cases

Prior to any analyses, the data were assessed to identify missing data. Two cases with missing data were identified and removed via listwise deletion. Additionally, data were assessed to identify extreme cases of floor or ceiling performance (i.e., summary scores of zero or 31). Such extreme cases are inestimable in Rasch analysis so these are typically removed from consideration (Linacre, 1994). Sixteen extreme cases were identified and removed.

Assumptions

Remaining cases (N = 169) were split into six groups based on teacher (Proctor One, Proctor Two) and test form (A, B, and C). Homogeneity across groups was assessed via Factorial Analysis of Variance (ANOVA) with total ITPO score as the dependent variable. Levene's Test indicated homogeneity of variance across group ITPO mean scores (F [5,163] = .680, p = .639). Additionally, analysis of skewness (Skewness/SE = -2.337) and kurtosis (Kurtosis/SE = -1.995) indicated ITPO total scores were approximately normally distributed. No significant difference in mean ITPO scores was identified based on teacher (F [1,163] = 2.661, p = .105) or test form (F [2,163] = 2.764, p = .798). Furthermore, no significant mean difference was identified across teachers by test form (F [5,163] = 2.764, p = .066). The nonsignificant results justify the choice to combine data across all groups for further analysis. See **Table 1** for contrast clusters and item loadings. The item problem structure of addition and subtraction items in **Table 1** and subsequently in the text of the manuscript is denoted using capital letters to represent the larger absolute value of N or n representing negative integers and P or p for positive integers. For example, n + p structure indicates the sum of additive inverses with the negative number first, whereas n + P indicates the sum of a negative number and a positive number with a greater absolute value.

Dimensionality

Dimensionality was assessed using Rasch Principal Components Analysis of Residuals (PCAR). The total raw variance in observations had an eigenvalue of 47.657. Measures (i.e., Person Ability and Item Difficulty) explained 35% of this variance (Person Ability, 18.3%; Item Difficulty, 16.7%). That is, the Rasch dimension explained 35% of the variance in observations. The raw unexplained variance had an eigenvalue equivalent to the number of assessed items (31) and represented 65% of the total raw variance.

PCAR analyzed the unexplained variance based on standardized residual variance as contrasts (i.e., components). Five contrasts were observed. The eigenvalue of a contrast is a measure of the strength of a component related to the items. That is, the first contrast explained 8.4% of the total raw variance with an item strength of 3.997 (i.e., the eigenvalue). Proportionally, the raw variance explained by items was approximately twice the unexplained variance in Contrast 1. The second contrast explained 6.4% (item strength = 3.047) of the total raw variance. The raw variance explained by items was approximately three times larger than the unexplained variance in Contrast 2. Remaining contrasts were found with eigenvalues less than two. Linacre (2019) states that eigenvalues less than two suggest that these contrasts are representative of the random "noise" expected in the Rasch Model. Thus, PCAR suggested two additional components affecting the outcomes observed in the ITPO in addition to the Rasch dimension.

PCAR contrasts depict sets of items orthogonal to the Rasch dimension (i.e., the first component; Linacre, 2019). Items are grouped into (three) clusters and compared based on their item (i.e., factor) loading. Typically, an identified cluster of strong positive (Cluster 1, e.g., loading > .3) or strong negative loadings (Cluster 3) within a contrast may represent an additional dimension – to the Rasch dimension – affecting items in the measure. Cluster 1 and Cluster 3 of Contrast 1 had a Pearson correlation of .326 and a disattenuated correlation (i.e., Pearson correlation measured without error) of .534. Additionally, Clusters 1 and 3 of Contrast 2 also suggested an additional component (i.e., disattenuated $r = .524$). Linacre (2019) states that disattenuated correlations below .570 supports the existence of additional components affecting items. See **Table 1** for contrast clusters and item loadings.

Table 1. Principal Components Analysis of Residuals (PCAR) item loadings – contrast one and contrast two ($N = 31$)

Item*	Contrast 1 cluster	Contrast 1 item loading	Item content structure	Contrast 2 cluster	Contrast 2 item loading
28	1	.63	n x n	2	-.06
23	1	.61	n ÷ n	2	-.02
29	1	.59	n ÷ (-1)	2	-.09
27	1	.58	n ÷ n	2	-.04
26	1	.55	p x (-1)	2	-.12
31	1	.50	n÷ (-1)	2	-.17
25	1	.49	p x n	3	-.28
21	1	.47	n x n	3	-.22
30	1	.33	n ÷ p	2	-.09
24	2	.27	n x p	2	.05
22	2	.26	-1 x p	2	.06
16	2	.04	n – N	1	.62
20	2	.03	n – N	1	.65
18	2	-.03	p – N	1	.55
15[b]	3	-.11	n – P	1	.38
13[a]	3	-.13	p – N	1	.57
9[a]	3	-.14	N – p	1	.42
11[a]	3	-.19	p + N	3	-.32
6	3	-.19	p + N	3	-.29
3	3	-.22	N + p	3	-.24
17	3	-.24	p + n	3	-.52
14	3	-.24	p – P	2	-.02
8	3	-.24	n + p	3	-.37
4	3	-.26	n + P	3	-.28
12[b]	3	-.27	p – P	2	-.03
5[a]	3	-.30	N – n	2	.14
19	3	-.31	n + P	3	-.30
7[a]	3	-.34	n+N	2	-.17
10[b]	3	-.36	N -n	1	.32
1	3	-.41	n + N	3	-.19
2[a]	3	-.45	n + P	3	-.30

Note. *Items listed relative to their cluster and item loading in Contrast 1. [a]Item adopted from Liebeck (1990). [b]Item adopted from Periasamy and Zaman (2009) in test Form A

Clusters 1 and 3 are typically observed as they include the items with the largest item loadings (positively and negatively, respectively). The items in Contrast 1, Cluster 1 shared a multiplicative commonality (i.e., all items involve multiplication or division). The items in Contrast 1, Cluster 3 shared an additive commonality (i.e., all items involve addition and subtraction). The items in Contrast 2, Cluster 1 shared a subtractive only commonality whereas Contrast 2, Cluster 3 suggested an additive only commonality (except for Item 21 and Item 25). Linacre (2019) suggests consideration of removal or restructuring of items that are affected by any identified additional dimensions. However, all identified commonalities are pertinent to the measurement of student ability with integer arithmetic including negative integers.

Typically, a single contrast identifies a single dimension (i.e., the positively or negatively loading cluster within the contrast become a dimension). While this is the case for Contrast 1 (i.e., multiplicative items separated from additive items), the second contrast separated the remaining items by additive operation (i.e., addition or subtraction). Thus, the two identified contrasts with large item strength (i.e., > 2) formed three dimensions. Subsequent analyses separated items of the ITPO across three dimensions: (1) Addition (AddD), (2) Subtraction (SubD), and (3) Multiplication/Division (MDdim). **Table 2** provides the items specific to each dimension.

Table 2. Dimensions of the Integer Test of Primary Operations (ITPO)

Dimension	ITPO item
Addition (AddD)	1, 2, 3, 4, 6, 7, 8, 11, 17, 19
Subtraction (SubD)	5, 9, 10, 12, 13, 14, 15, 16, 18, 20
Multiplication/Division (MDdim)	21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31

Addition Dimension

AddD consisted of 10 total items. Prior to applying the Rasch Dichotomous Model (RDM), missingness and outlier analysis were reexamined on the original data (N = 187). Eighty extreme cases (i.e., scores of 0 (n=71) or 10 (n=9), see **Table 3**) were identified. A reduced sample (N = 107, M = 7.12, and SD = 2.37) was then analyzed using the RDM. Data were negatively skewed (*Skewness/SE* = 5.538) and mesokurtic (*Kurtosis/SE* = 1.443).

Table 3. Missing and extreme cases removed from the sample (N = 187) prior to Rasch analysis

Level	Addition construct n (%)	Subtraction construct n (%)	Multiplication/division construct n (%)
Missing	-	-	2 (1.1%)
Ceiling	9 (4.8%)	14 (7.5%)	2 (1.1%)
Floor	71 (38%)	23 (12.3%)	82 (43.9%)

Addition person summary and fit statistics

Person Ability measures (in logits; see **Table 4**) ranged from -2.35 to 2.32. Mean Person Ability (1.17) suggested a negatively skewed distribution, confirmed by observation of the Item-Person (Wright) map. Thus, items of the AddD tended to be easier for persons to endorse (i.e., Person Ability Mean > 0). Person Root-Mean-Square-Error (RMSE = .93) was close to the model RMSE (.90) indicating little error in the data. Person separation (1.02) suggested low discrimination (i.e., only one or two strata) between persons' ability levels. Low Person reliability (.51) indicated a narrow range of person ability levels (Linacre, 2019). Linacre alternatively suggests that a measure with low person reliability (i.e., < .80) may also benefit from the inclusion of additional items.

Analysis of Mean Square (MNSQ) and standardized (ZSTD) summary fit statistics indicated significant (i.e., | ZSTD | > 1.96) maximum infit and outfit statistics. The value of the maximum outfit suggested the existence of persons degrading to measurement using the Rasch Model (Wright & Linacre, 1994). Three persons (Person 61, Person 117, and Person 164) were identified with unusual fit statistics.

Table 4. Integer Test of Primary Operations (ITPO) addition dimension person statistics summary (N = 107)

Statistic	Total score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	7.10	10	1.17	.89	1.00/.12	.99/.12
P.SD	2.40	.00	1.33	.17	.17/.68	.45/.72
S.SD	2.40	.00	1.33	.17	.17/.68	.45/.72
Max	9.00	10	2.32	1.08	1.54/2.85	4.01/2.76
Min	1.00	10	-2.35	.66	.73/-1.85	.57/-1.68

Note. Real/Model RMSE = .93/.90; Real/Model True SD = .95/.97; Real/Model Separation = 1.02/1.07; Real/Model Person Reliability = .51/.54; Coefficient Alpha (KR-20) = .71, SEM = 1.26

Person 61 and Person 117 obtained large MNSQ outfit statistics (MNSQ = 2.25, MNSQ = 4.01, respectively); however, their standardized fit statistics did not indicate troubling misfit. That is, Person 61 received a nonsignificant standardized outfit (ZSTD = 1.25, p > .05) and Person 117 had an unusual outfit (ZSTD = 2.05, p < .05) but may not be considered outliers (i.e., | ZSTD | > 2.58). Person 164 (Infit MNSQ = 1.54, Outfit MNSQ = 1.64) obtained unstandardized fit statistics that are good for measurement (Wright & Linacre, 1994). However, the standardized fit (Infit ZSTD = 2.85, p < .01; Outfit ZSTD = 2.76, p < .01) suggested significant misfit.

These fit statistics did not provide a clear picture for misfit persons (i.e., that the measure may benefit from removal of persons). Additionally, Linacre (1994) suggests that Rasch modeling is robust to misfit persons (more so than misfit items). Therefore, removal of persons was not considered.

Addition item summary and fit statistics

Summary statistics for AddD are defined in **Table 5**. Item difficulty measures ranged from -1.34 logits to .77 logits. Item reliability (.80) suggests an acceptable sample size was used for analysis of item difficulty. Item separation (1.98) indicates good discrimination of levels (strata) between items (i.e., three levels; Linacre, 2013). Real RMSE (.27) was close to the Model RMSE (.26) suggesting little error in the data. Analysis of MNSQ fit statistics reveals significant (i.e., | ZSTD | > 1.96) maximum infit and outfit statistics, as well as significant minimum infit statistics. Wright and Linacre (1994) indicate that items with MNSQ statistics between 1.50 and 2.00 may be unproductive for measurement, though not degrading to a scale.

Table 5. Integer Test of Primary Operations (ITPO) addition dimension item statistics summary (N = 10)

Statistic	Total score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	76.2	107	.00	.26	.97/-.07	.99/.14
P.SD	8.3	.00	.60	.03	.20/1.36	.10/.46
S.SD	8.7	.00	.63	.03	.21/1.43	.31/1.44
Max	93.0	107	.77	.34	1.29/2.30	1.58/3.04
Min	64	107	-1.34	.23	.61/-1.98	.59/-1.80

Note. Real/Model RMSE = .27/.26; Real/Model True SD = .53/.54; Real/Model Separation = 1.98/2.05; Real/Model Item Reliability = .80/.81; Standard Error of Item Mean = .20

Individual items were scrutinized based on significant minimum and maximum fit statistics. Item 7 (Outfit MNSQ = 1.58, ZSTD = 3.04, $p < .05$) was potentially misfit to the model. The point-measure correlation for Item 7 (.31) was the lowest recorded value among AddD items. Additionally, Item 7 shared common structure with Item 1 (i.e., a negative first term summed with a larger negative second term in parentheses) that has a similar item difficulty measure (.55 and .44 for Item 7 and Item 1, respectively). However, other items also had equivalent structure (e.g., Item 6 and Item 11) but did not present as misfit or redundant. Furthermore, the item-person map (see **Figure 1**) did not suggest redundancy of items and analysis of Person Ability statistics indicated that the AddD may benefit from additional items. Thus, removal of Item 7 was not considered.

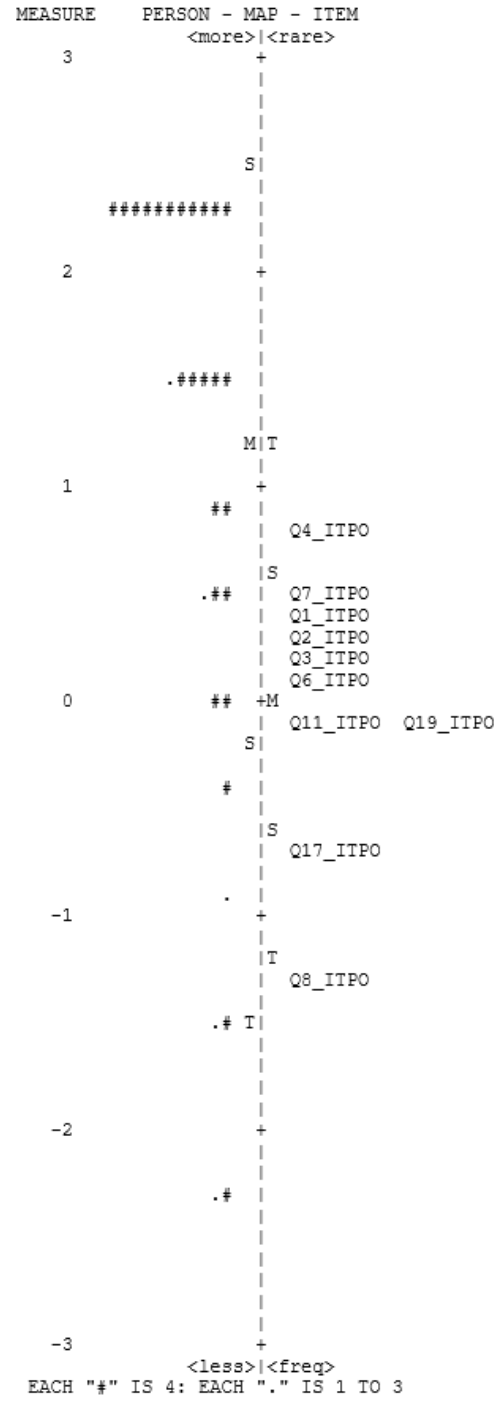

Figure 1. Item-person map for the addition dimension of the integer test of primary operations

Addition model fit and reliability

The log-likelihood chi-square statistic (χ^2 = 951.80, *df* = 959, *p* = .56) was not significant. Additionally, the Global Root-Mean-Square Residual (RMSR = .3767) was less than the expected RMSR (.3769). Both measures supported good model fit. The Kuder-Richardson (KR-20) reliability statistic for the AddD was .71.

Addition descriptives: Accuracy of full and reduced sample

Although Rasch analyses render the following descriptives moot from a measurement and evaluation perspective, from a practical mathematics classroom instruction perspective, it would be useful to consider the descriptive statistics of items inclusive of all students who took the assessment. Thus, **Table 6** includes the accuracy of each addition item of the full sample (i.e., from 72 to 88%) to compare to the accuracy of those included in the Rasch subsample (60 to 87%). To provide insights about whether some students might have better conceptual understanding of the calculation even if inaccurate, the accuracy rates are posted next to the rates of accuracy of student open responses that have the accurate sign. Note that approximately 75% or more of students in the Rasch sample answered each addition item with the correct sign (e.g., 18 who were inaccurate on the most difficult question 4, provided a positive solution).

Table 6. Addition items by level, mathematical structure, accuracy rate, and correct sign usage

Rasch level[a]	Integer structure	Digit structure	Item #	Accurate Rasch subsample n (%)	Correct sign Rasch subsample n(%)	Accurate full sample n(%)
3	n+P	DD+DD	Q4	64 (59.8%)	82 (76.6%)	135 (72.2%)
2	n+N	SD+SD	Q7	68 (63.6%)	80 (74.8%)	139 (74.3%)
	n+N	SD+SD	Q1	70 (65.4%)	82 (76.6%)	141 (75.4%)
	n+P	SD+SD	Q2	73 (68.2%)	81 (75.7%)	144 (77%)
	N+p	DD+SD	Q3	74 (69.2%)	87 (81.3%)	145 (77.5%)
	p+N	SD+SD	Q6	75 (70.1%)	89 (83.2%)	146 (78.1%)
	p+N	SD+SD	Q11	79 (73.8%)	90 (84.1%)	150 (80.2%)
	n+P	SD+SD	Q19	79 (73.8%)	88 (82.2%)	150 (80.2%)
	p+n	DD+DD	Q17	87 (81.3%)	-	158 (84.5%)
1	p+n	DD+DD	Q8	93 (86.9%)	-	164 (87.7%)

Note. Sample size of each analysis: Accurate and Correct Sign Rasch Subsamples (n = 107); Accurate Full Sample (N = 187). Capital N or P indicates these integers have the greater absolute value. SD indicates single digit and DD indicates double digit regardless of magnitude comparisons. [a]Rasch level based on separation index in item summary statistics

Subtraction Dimension

SubD consisted of ten total items. Prior to applying the RDM, missingness and outlier analysis were reexamined on the original data (N = 187). Thirty-seven extreme cases (i.e., scores of 0 (n = 23) or 10 (n = 14), see **Table 3**) were identified. A reduced sample (N = 150, M = 4.82, SD = 2.609) was then analyzed using the RDM. Data were approximately symmetric (*Skewness/SE* = 1.217) and platykurtic (*Kurtosis/SE* = 3.350).

Subtraction person summary and fit statistics

Person ability measures (M = -.07; see **Table 7**) ranged from -2.40 to 2.40 suggesting an approximately symmetric distribution (i.e., on average item difficulty was equivalent to person ability). The item-person map suggested a bimodal distribution of Person Ability scores. Person RMSE (.86) was close to the model RMSE (.82) suggesting little error in the data. Person separation (1.33) suggested moderate discrimination between strata of person abilities and Person reliability (.64) was low, like the AddD.

Analysis of person ability fit statistics suggested some misfit persons (i.e., large [MNSQ > 2] and unlikely [| ZSTD| > 1.96] fit statistics; Bond & Fox, 2015; Wright & Linacre, 1994). MNSQ fit statistics suggested only a few problematic persons (i.e., MNSQ > 2). However, misfit persons are not as problematic as misfit items (Linacre, 1994). Thus, removal of misfit persons was not considered.

Table 7. Integer Test of Primary Operations (ITPO) subtraction dimension person statistics summary (N = 150)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	4.80	10	-.07	.81	1.00/.04	1.04/.07
P.SD	2.60	0	1.42	.14	.26/.87	.61/.96
S.SD	2.60	0	1.43	.14	.26/.87	.61/.97
Max	9	10	2.40	1.08	1.77/2.70	3.79/2.70
Min	1	10	-2.40	.67	.59/-1.96	.42/-1.75

Note. Real/Model RMSE = .86/.82; Real/Model True SD = 1.14/1.16; Real/Model Separation = 1.33/1.42; Real/Model Person Reliability = .64/.67; Coefficient Alpha (KR-20) = .73, SEM = 1.36

Subtraction item summary and fit statistics

Item Difficulty measures ranged from -1.21 logits to 1.06 logits. Item Real RMSE (.21) was close to the model RMSE (.20) suggesting little error in the data. Item reliability (.92) suggested a large enough sample size was used. Item separation (3.40) indicated excellent discrimination between strata of item difficulty. High item reliability and separation provided evidence of construct validity for the dimension (Linacre, 2019). Analysis of fit statistics suggested potentially misfit items. However, all item infit and outfit MNSQ fit statistics all presented as productive for measurement (Wright & Linacre, 1994) and the item-person map (see **Figure 2**) did not suggest redundancy of items. Additionally, analysis of Person abilities suggested that the SubD may benefit from more items. Therefore, removal of items was not considered. See **Table 8** for full item summary statistics.

Table 8. Integer Test of Primary Operations (ITPO) subtraction dimension item statistics summary (N = 10)

Statistic	Total score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	72.3	150	.00	.20	.99/-.09	1.04/.08
P.SD	18.6	.00	.73	.01	.14/1.47	.29/1.58
S.SD	19.6	.00	.77	.01	.15/1.55	.30/1.67
Max	103	150	1.06	.21	1.15/1.60	1.52/2.68
Min	46.00	150	-1.21	.20	.78/-2.19	.63/-2.18

Note. Real/Model RMSE = .21/.20; Real/Model True SD = .70/.70; Real/Model Separation = 3.40/3.51; Real/Model Item Reliability = .92/.92; Standard Error of Item Mean = .24

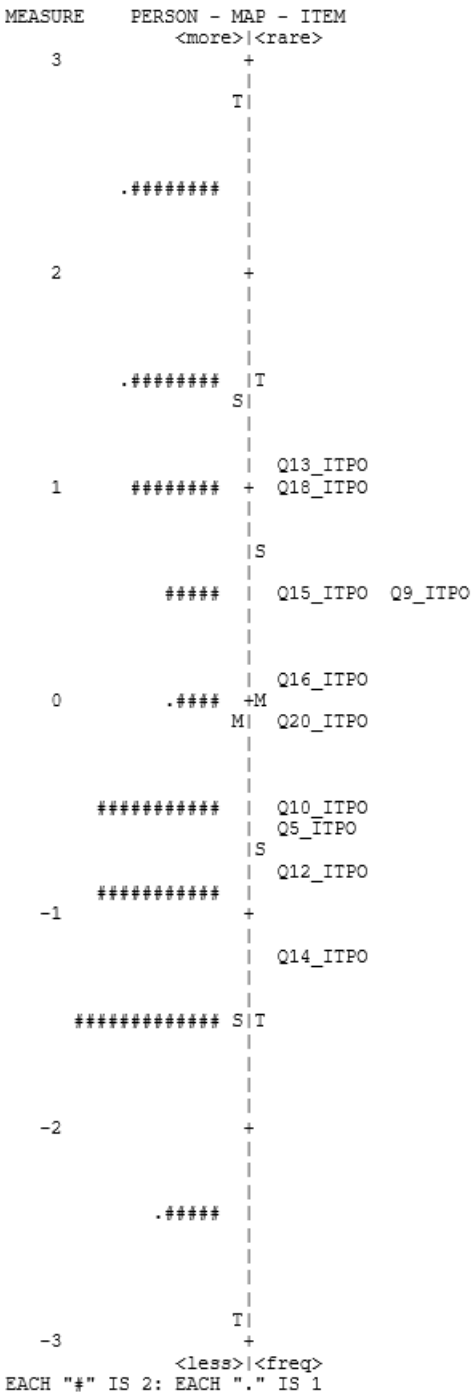

Figure 2. Item-person map for the addition dimension of the integer test of primary operations

Subtraction model fit and reliability

The log-likelihood chi-square statistic (χ^2 = 1501.13, *df* = 1494, *p* = .44) was not significant. Additionally, the Global Root-Mean-Square Residual (RMSR = .4048) was less than the expected RMSR (.4066) suggesting the model was overfit to the data. Both measures supported good model fit. The KR-20 reliability statistic for the AddD was .73.

Subtraction descriptives: Accuracy of full and reduced sample

Although Rasch analyses render the following descriptives moot from a measurement and evaluation perspective, from a practical mathematics classroom instruction perspective, it would be useful to consider the descriptive statistics of items inclusive of all students who took the assessment. Thus, **Table 9** includes the accuracy of each subtraction item of the full sample (i.e., from 37 to 67%) to compare to the accuracy of those included in the Rasch subsample (31 to 69%). To provide insights about whether some students might have better conceptual understanding of the calculation even if inaccurate, the accuracy rates are posted next to the rates of accuracy of student open responses that have the accurate sign. Note that approximately 50% or more of students in the Rasch sample answered each subtraction item with the correct sign (e.g., 44 students who were inaccurate on the most difficult question 13, correctly provided a negative solution, such that 60% of the sample knew the solution was negative).

Table 9. Subtraction items by level, mathematical structure, accuracy rate, and correct sign usage

Rasch level[a]	Integer structure	Digit structure	Item #	Accurate Rasch subsample n (%)	Correct sign Rasch subsample n(%)	Accurate full sample n (%)
4	p-N	SD-SD	Q13	46 (30.7%)	90 (60%)	69 (36.9%)
	p-N	DD-DD	Q18	47 (31.3%)	75 (50%)	70 (37.4%)
3	n-P	SD-DD	Q15	59 (39.3%)	81 (54.4%)	82 (43.9%)
	N-p	SD-SD	Q9	59 (39.3%)	138 (92.0%)	82 (43.9%)
2	n-N	SD-SD	Q16	70 (46.7%)	84 (56%)	93 (49.7%)
	n-N	SD-SD	Q20	74 (49.3%)	87 (58%)	97 (51.9%)
1	N-n	SD-SD	Q10	85 (56.7%)	120 (80.0%)	108 (57.8%)
	N-n	SD-SD	Q5	87 (58%)	123 (82.0%)	110 (58.8%)
	p-P	DD-DD	Q12	93 (62%)	105 (70.0%)	116 (62%)
	p-P	DD-DD	Q14	103 (68.7%)	109 (72.7%)	126 (67.4%)

Note. Sample size of each analysis: Accurate and Correct Sign Rasch Subsamples (n=150); Accurate Full Sample (N=187). Capital N or P indicates these integers have the greater absolute value. SD indicates single digit and DD indicates double digit regardless of magnitude comparisons. [a]Rasch level based on separation index in item summary statistics

Multiplication/Division Dimension

The MDdim consisted of 11 total items. Prior to applying the RDM, missingness and outlier analysis were reexamined on the original data (N = 187). Two cases of missing data and 84 extreme cases (i.e., scores of 0 (n = 82) or 11 (n = 2), see **Table 3**) were identified. A reduced sample (N = 101, M = 6.73, SD = 3.01) was then analyzed using the RDM. Data were approximately symmetric (*Skewness/SE* = 1.733) and platykurtic (*Skewness/SE* = 2.859).

Multiplication/division person summary and fit statistics

Person Ability (M = .64, see **Table 10**) ranged from -2.41 to 2.40 logits on the MDdim suggesting a negatively skewed Person Ability distribution (i.e., Person Ability Mean > 0; Items easier to endorse on average). The item – person map (see **Figure 3**) supported the distribution of person abilities. The real RMSE (.84) was near the model RMSE (.82) indicating little error in the data. Person separation (1.41) suggested a moderate discrimination between person ability levels (Linacre, 2013). Person reliability (.66) was low, like the AddD and SubD.

Table 10. Integer Test of Primary Operations (ITPO) multiplication/division dimension person statistics summary (N = 101)

Statistic	Total score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	6.70	11.00	.64	.81	1.00/.18	.99/.15
P.SD	3.00	.00	1.45	.16	.15/.59	.35/.69
S.SD	3.00	.00	1.46	.16	.15/.60	.36/.69
Max	10.00	11.00	2.40	1.06	1.39/2.14	2.84/2.05
Min	1.00	11.00	-2.41	.62	.75/-1.10	.45/-1.10

Note. Real/Model RMSE = .84/.82; Real/Model True SD = 1.19/1.20; Real/Model Separation = 1.41/1.46; Real/Model Person Reliability = .66/.69; Coefficient Alpha (KR-20) = .79, SEM = 1.38

Analysis of person ability fit statistics suggested some misfit persons (i.e., large [MNSQ > 2] and unlikely [| ZSTD| > 1.96] fit statistics; Bond & Fox, 2015; Wright & Linacre, 1994). MNSQ fit statistics suggested only a few problematic persons (i.e., MNSQ > 2). However, misfit persons are not as problematic as misfit items (Linacre, 1994). Thus, removal of misfit persons was not considered.

Multiplication/division item summary and fit statistics

Item Difficulty (see **Table 11**) ranged from -1.01 to .73 logits. Real RMSE (.26) was close to model RMSE (.25) indicating little error in the data. Item separation (1.64) and item reliability (.73) suggested a large enough sample size and moderate discrimination between item strata in the construct hierarchy. The item-person map (see **Figure 3**) suggested a relatively uniform distribution. Item fit statistics did not indicate any problematic items (i.e., MNSQ > 2).

Table 11. Integer Test of Primary Operations (ITPO) multiplication/division dimension item statistics summary (N = 11)

Statistic	Total score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	61.80	101.00	.00	.25	1.01/.03	.99/-.13
P.SD	7.80	.00	.50	.01	.16/1.31	.26/1.23
S.SD	8.20	.00	.53	.01	.17/1.27	.27/1.29
Max	77.00	101.00	.73	.27	1.31/2.20	1.50/2.20
Min	50.00	101.00	-1.01	.24	.81/-1.54	.65/-1.86

Note. Real/Model RMSE = .26/.25; Real/Model True SD = .43/.43; Real/Model Separation = 1.64/1.72; Real/Model Item Reliability = .73/.75; Standard Error of Item Mean = .16

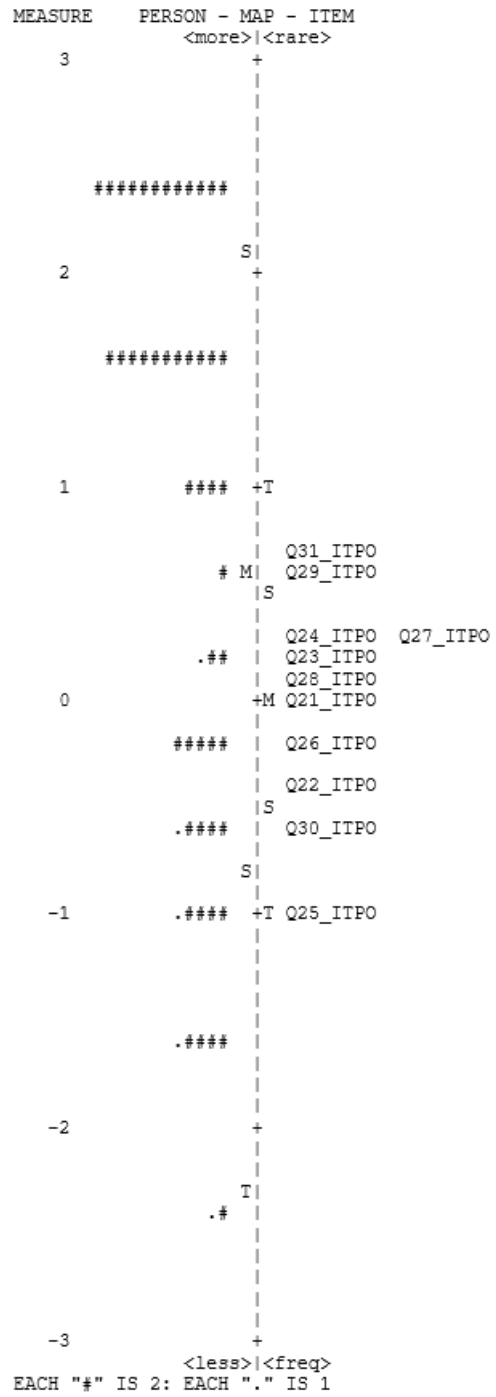

Figure 3. Item-person map of the multiplication/division dimension of the integer test of primary operations

Multiplication/division model fit and reliability

The log-likelihood chi-square statistic (χ^2 = 1067.16, *df* = 1071, *p* = .53) was not significant suggesting good model fit. However, the Global Root-Mean-Square Residual (RMSR = .3961) was greater than the expected RMSR (.3953) suggesting that the model was slightly underfit to the data. The KR-20 reliability statistic for the MDdim was .79.

Multiplication/division descriptives: Accuracy of full and reduced sample

Although Rasch analyses render the following descriptives moot from a measurement and evaluation perspective, from a practical mathematics classroom instruction perspective, it would be useful to consider the descriptive statistics of items inclusive of all students who took the assessment. Thus, **Table 12** includes the accuracy of each multiplication/division item of the full sample (i.e., from 71% to 86%) to compare to the accuracy of those included in the Rasch subsample (50% to 76%). To provide insights about whether some students might have better conceptual understanding of the calculation even if inaccurate, the accuracy rates are posted next to the rates of accuracy of student open responses that have the accurate sign. Note that approximately 62% or more of students in the Rasch sample answered each multiplication/division item with the correct sign (e.g.,

29 who were inaccurate on the most difficult question 31 that involved dividing a positive number by -1, 79% correctly provided a negative solution although their actual solution was inaccurate).

Table 12. Multiplication and division items by level, mathematical structure, accuracy rate, and correct sign usage

Rasch level[a]	Integer structure	Item #	Accurate Rasch subsample n (%)	Correct sign Rasch subsample n(%)	Accurate full sample n (%)
3	P÷n [P÷-1]	Q31	50 (49.5%)	79 (79.0%)	132 (71.4%)
	N÷n [N÷-1]	Q29	52 (51.5%)	63 (62.4%)	134 (72.4%)
2	N÷n	Q27	57 (56.4%)	64 (63.4%)	139 (75.1%)
	n×P	Q24	57 (56.4%)	67 (66.3%)	139 (75.1%)
	N÷n	Q23	59 (58.4%)	63 (62.4%)	141 (75.4%)
	n×N	Q28	60 (59.4%)	69 (68.3%)	142 (76.8%)
	n×N	Q21	62 (61.4%)	66 (65.3%)	144 (77.8%)
1	P×n [P×-1]	Q26	66 (65.3%)	68 (67.3%)	148 (80%)
	n×P [-1 ×P]	Q22	69 (68.3%)	73 (72.3%)	151 (81.6%)
	N÷p	Q30	71 (70.3%)	79 (79.0%)	153 (82.7%)
	p×N	Q25	77 (76.2%)	80 (80.0%)	159 (85.9%)

Note. Sample size of each analysis: Accurate and Correct Sign Rasch Subsamples (n = 101); Accurate Full Sample (N = 185). Capital N or P indicates these integers have the greater absolute value. [a] Rasch level based on separation index in item summary statistics

Summary of Findings

The main finding from the Rasch Analysis for all three identified components (i.e., Addition, Subtraction, Multiplication/Division) were five-fold: (1) Person summary and fit statistics indicated good reliability and separation, (2) Item summary and fit statistics also indicated good reliability and separation, (3) The Item-Person Map corroborated the above findings, (4) Good model fit was evidenced, (5) there was strong internal consistency/reliability for all components. **Table 13** summarizes these main findings.

Table 13. Summary of primary Rasch analysis with related conceptual findings

Section	Analysis	Findings
Dimensionality	Rasch principal components analysis of residuals	Three components identified
Addition dimension* Subtraction dimension* MD[a] dimension*	Rasch dichotomous model	
	Person summary statistics	Good person separation and reliability No persons misfit to the model
	Item summary statistics	Good item separation and reliability No items misfit to the model[b]
	Item-person (Wright) map	Confirmed person/item separation and reliability
	Model fit	Good model fit (p > .05 for all)
	Reliability	Strong internal consistency (KR-20 > .70)

Note. * Order of analysis the same across dimensions. [a]Multiplication/Division. [b]One item identified as redundant after Rasch analysis in MD dimension

DISCUSSION AND IMPLICATIONS

We first address the micro level implications in terms of what we learned about particular item structures related to each research question. Then we explain the broader implications that the field now has a reliable and valid instrument for use with middle school students that consists of three forms, which are ethically available to any stakeholder in education.

Item Level Discussion and Implications

From a school mathematics perspective there are four primary operations and from a disciplinary mathematics perspective there are just two operations (addition and multiplication) by which the others are defined. Thus, we might have expected that the test would consist of four or two unidimensional constructs. Alternatively, because subtraction of a negative number results in addition operations (e.g., 4 – ⁻2 = 4 + 2) and multiplication and division follow regular rules that differ only by the number of negative factors, we might have expected some addition problem structures would load with some more mathematically related subtraction structures and that multiplication and division items might also separate based on the regularity of the rules. In fact, the 31-item Integer Test of Primary Operations cleanly separated into three constructs based on operations: addition, subtraction, and multiplication and division loading together as a single unidimensional construct. This answered RQ1: “What is the

dimensional structure of the ITPO?". Linacre (2018) contended that PCAR can identify groupings of items that measure a "strand" of a dimension rather than a specific dimension. That is, multiplication and division are likely strands of the same dimension (i.e., measuring the same dimension in different ways) rather than individual dimensions.

The Rasch person-fit results that overall the operation of integer addition was easiest, multiplication and division next easiest and subtraction most difficult was consistent with a pattern we noticed in the descriptive statistics of Ryan and Williams' (2007) study that used only three integer items as part of a mathematics measure. Typical integer instructional sequences follow this same pattern: addition, then subtraction and then multiplication and division. Yet, consider that to promote elementary student understanding of relationships between whole number operations, current standards combine addition and subtraction (NGO & CCSSO, 2010). Additionally, middle grade students should already have prerequisite whole number multiplication and division knowledge and -a is really (-1) (a). Thus, we question whether integer learning must follow the same sequence of operations that students experience for whole number instruction and suggest future research compare instructional sequences. The Integer Test of Primary Operations provides one measure that could be used as part of research designs that intend to make such comparisons and determine if the Item-Person Maps differ when different instructional sequences are used.

With regard to RQ1, notice that within each construct, same problem structures grouped together in terms of their difficulty. This was evident in the findings shown in Tables 6, 9 and 12 and due to placement on the item-person maps in **Figure 1**, **Figure 2**, and **Figure 3**. **Table 14** summarizes these key findings about the item structures to make these patterns easier to discern and more practically useful for stakeholders to apply this information to their future work. As stated in the purpose of the study, if RQ1 was satisfied, based on theoretical and practical reasons, then we would ask RQ2: "After instruction which integer problem structures did middle school students find most difficult or easiest?" Given their mathematical and instructional importance, the three subquestions of RQ2 are discussed next.

Table 14. Summary of integer operation problem difficulty

Rasch level	Addition		Multiplication & division		Subtraction	
	Integer structure	Q# (Digit structure)	Integer structure	Q# (Digit structure)	Integer structure	Q# (Digit structure)
4	--		--		p-N	Q13 (SD-SD) Q18 (DD-DD)
3	n+p	Q4 (DD+DD)	Division by -1	Q31& Q29	n-P	Q15 (SD-DD)
					N-p	Q9 (SD-DD)
2	n+N n+P N+p p+N p+n	Q7 & Q1 (SD +SD) Q2 (SD + SD) Q3 (DD + SD) Q6 & Q11 (SD+SD) Q17(DD+DD)	Products & quotients of two negative integers that are < -1	Q27, Q24, Q23, Q28, Q21	n-N	Q16 & Q20 (SD-SD)
1	p+n	Q8 (DD+DD)	Products of a positive integer and -1 N÷p p×N	Q26 & Q22 Q30 Q25	N-n p-P	Q10 & Q5 (SD-SD) Q12 & Q14 (DD-DD)

Note. This table summarizes the crucial findings from Tables 6, 9 and 12. Capital N or P indicates these integers have the greater absolute value. SD indicates single digit and DD indicates double digit regardless of magnitude comparisons. Levels indicate top to bottom most difficult Level 4 to easiest Level 1 within each operation dimension. Similarly, difficulty of items retains the vertical map order of the figures with the most difficult items listed first within each Q# column. Left to right the operation dimension is organized from least to most difficult (addition, multiplication/division, and subtraction).

Multiply or divide by -1 items

Given the ease that even young children have memorizing products of whole numbers that have the multiplicative identity as a factor, we would have expected that multiplying or dividing by $^{-}1$ would be the easiest items in that construct. Thus, RQ2a investigated whether this was true: "Is -1(X) or X divide -1 a more difficult construct than other multiplication/division items?" Note in **Table 14** that items involving multiplying by -1 loaded together along with other items that were the product or quotient of a positive and a negative number. These were the easier items, however multiplying by $^{-}1$ was not the easiest structure. Moreover, contrary to expectations, division by $^{-}1$ did not load with multiplication by $^{-}1$. This means that for students, factoring by $^{-}1$ was unrelated to multiplying by $^{-}1$. Division by -1 were the most difficult items of any products or quotients (see **Table 14**). This finding has qualitatively practical importance for success in later mathematics. Algebraic proofs or justifications often involve factoring out $^{-}1$, so this procedural skill and understanding the conceptual relationship that factoring $^{-}1$ is the inverse of multiplying by $^{-}1$ are crucial for these learners of integer arithmetic to master prior to an algebra course.

Additive inverses

"How difficult were additive inverse items?" (RQ2b) was asked due to its use in proofs in disciplinary mathematics as well as the reliance on the sum of additive inverses as the basis of determining chemical charges. Across all three constructs (addition, subtraction and multiplication/division), the two additive inverse items were the easiest items (see **Figure 1**, **Figure 2**, and **Figure**

3). That is, after having learned integer arithmetic with multiple models, including a chip model and a number line model, almost all students in this sample were able to answer these additive inverse item structures. When interpreting any assessment, however, it is crucial to consider the purpose (Bond & Fox, 2015). Therefore, if someone wishes to claim that additive inverse items are easiest to learn, this will need to be investigated in future studies. A suggestion for future research would be to replicate the study here with both pre and posttests with a similar population. If the items load in similar ways on the posttest as they did here and differently for the pretest, then this information combined with qualitative analyses could provide additional insights about effective instructional sequences.

Subtraction of negative integers

Whereas our analysis for addition and multiplication/division discerned three levels of difficulty within each dimension, four levels of difficulty were discriminated for subtraction operations (see **Table 14**). The answer to RQ2c (*Were items involving subtracting a negative number the most difficult items?*) was less straightforward. As predicted, the most difficult problem structures were subtraction items in which a negative number was subtracted to yield a positive solution (p–N; Q13 and Q18). Although in terms of mathematical structure, each specific subtraction structure grouped together in reliable ways (e.g., Level 3 n–P, N–p; Level 2 n–N, Level 1 N–n together before p–P; see **Table 14**), from a broader view the results do not allow us to claim that all problems in which subtracting a negative number yields a positive solution are most difficult for students. As **Table 9** and the Item-Person Map in **Figure 2** showed, beginning with a negative number and subtracting a positive (n–P or N–p) was less difficult than the structure of p–N, but more difficult than n–N. What is interesting and worthy of further consideration is that both of these items that seem contrary to theoretical and practical predictions were two items taken directly from existing measures and have about 40% accuracy in the Rasch subsample as well as the full sample. One of these items (Q9; $^{-}3-^{-}2$) was one of the anchor items appearing on every form taken from the Liebeck (1999) study. Although Liebeck's (1999) study that compared instruction with a chip model to a number line model is frequently cited, the design had multiple threats to experimental validity as well as a troublesome reliance on an unvalidated instrument that intended to assess student competency of addition and subtraction using only the digits 3 and 2. Given that quantities within three are subitizable (Kaufman et al., 1949), Liebeck's (1999) instrument limitation led us to predict that students would find these easy on the ITPO. After informally seeing the results of student accuracy across multiple implementations of the ITPO, we suspect that it may be the very size of the absolute values as subitizable that students in a sense take for granted and might passively answer this question quickly rather than actively using a learned strategy about how to subtract negative numbers. This supposition should be investigated with stratified interviews based on answers students provided to the open response items. For example, Liebeck's Q9 ($^{-}3-^{-}2$, N–n structure) loaded as similar difficulty to Q15, which was a single digit-minus a double-digit number (n–N structure) modeled after an item from Periasamy and Zaman (2009). When Periasamy and Zaman (2009) analyzed this item difficulty using only descriptive statistics, this was the most difficult of the single-digit minus double-digit items. Thus, when rigorously assessed with Rasch analysis in our study, it was not surprising that this item structure was the third most difficult subtraction item on the ITPO. Therefore, the answer to question RQ2c, as to whether problems that require subtracting a negative number are most difficult, is that it depends on the structure. These items occurred in higher as well as lower difficulty levels on the item-person map. The Level 4 p–N items had about 30% accuracy. Whereas the other subtraction of a negative items was Level 2, which included the mean and about 50% accuracy, indicating these items were typical of students in the Rasch analysis (see **Tables 9**).

The easiest subtraction structures (p–P; Q12 & Q14, see **Table 14**) were consistent with anecdotal experience that even without formal instruction learners find these problem structures more intuitive. Although after formal instruction, these were the easiest of the problem structures, the accuracy rates were below a criterion of competence. Each of these items had less than a passing criterion of accuracy for those in the Rasch subsample as well as when those with perfect scores on subtraction (i.e., full sample) were considered in the descriptive statistics. Moreover, even when we consider students with apparent conceptual understanding that these answers would be negative even if we disregard calculation errors, only 70% provided negative solutions. This was still barely competent against a typical classroom criterion of passing (see **Table 9** correct sign Rasch subsample). One reason we suspect students were less accurate than expected on these items is that students may overgeneralize commutativity of addition and misapply it to subtraction. This hypothesis could be tested by analyzing student answers and strengthened with interviews that ask students to explain their thinking. Such research would be especially important to conduct on its own or include as a statistical control. Although a prior study (Young & Booth, 2019) asserted that inaccurate answers on algebraic equivalence tests (e.g., 4x – 3 is equivalent to 3 – 4x, p.6) were due to students misunderstanding the negative sign, we suspect that an additional reason for the errors reported in that study were due to student misconceptions that commutativity is independent of operation.

Broader Implications and Future Research

For ethical reasons, we strategically chose this journal that allows any stakeholder access to the ITPO: Classroom teachers who want a valid assessment of their students' learning, researchers who need valid and reliable measures to assess instructional practices, as well as test developers who would create better tests of multiple constructs that include integer arithmetic if informed by this focused study of integer items. Teachers need valid and reliable no-cost assessments with low testing-time requirements. If teachers create their own unit tests or use textbook-provided assessments that primarily use items of the easiest structures, then they are likely to overestimate student proficiency with integer arithmetic that is crucial for later algebra. Practical needs of tests used in real classrooms must also always balance the time taken away from instruction for students to do the assessment with the information gained from the assessment. The ITPO provides this reasonable balance.

Integer test of primary operations is practical for teachers and scholars

In addition to the findings of the Rasch analysis (recall this requires removal of students who had 0% after instruction as well as those students who demonstrated complete mastery with 100%), the descriptive statistics of the full sample in **Table 6**, **Table 9**, and **Table 12** provide evidence that these assessments have practical value. The participating students had learned all four integer operations as well as other integer constructs but had not yet experienced their unit review prior to taking the ITPO. Regardless, at the time-point of the test, item accuracy of the full sample demonstrated that for addition and multiplication/division students met a criterion typical of classroom learning (i.e., better than 70 or 75%), because these ranged from 72% to 88% and 71% to 86% respectively. Descriptive item accuracy for subtraction was consistent with the Rasch analysis that this was the most difficult integer operation construct. The full sample did not meet criterion for any subtraction item, with each item below 68% accuracy (see **Table 9**) and just 7.5% of students demonstrating mastery of all subtraction items (see **Table 3**). Inspection of the Rasch subsample showed that integer subtraction item accuracy barely approached a passing criterion by ranging from 50% to 76%. Once negative sign accuracy was the criterion of competence to allow for simple errors during integer subtraction, students' subtraction competency was still troubling, although it included the acceptable criterion in the range from 62% to 80%. For teachers (or scholars) who offer partial credit for such calculation errors, this additional analysis we offered based on sign accuracy should provide confidence that the ITPO forms will be practically useful.

To further ensure these test forms are practically useful, we provide these forms in the **Appendix** as 30-item tests rather than 31 by omitting what was Q24, thus renumbering the remaining questions. An explanation for this follows. The multiplication and division construct consisted of 11 items, whereas addition and subtraction each have 10 items. There are two ways consideration of items could have been handled. First, as presented in results, 11 items diagnostically worked well with the other items, so based on the infit/outfit diagnostics as presented in the results there is no reason to dismiss or eliminate an item. Theoretically however, Q24 is the only item with a mathematical structure that loaded out of place in relation to the structure of other items. This may have been because each of these items had 9 as a factor such that students may have inaccurately calculated the magnitude due to errors in their whole number multiplication facts. Q24 was at the same location on the vertical ruler as Q27. This means they were at the same difficulty level and based on infit and outfit statistics of Q24, if we remove this item there would not be a large shift in the psychometric properties. Moreover, Q27 theoretically fits. Thus, given the practical interest that each construct might have exactly 10 items to offer a parsimonious test and so that qualitatively the problem structures load together in meaningful ways, we provide the test forms A, B, and C without item "Q24". If future research uses these versions of the forms and needs to report exact Rasch statistics of the forms used, it would be prudent to verify this as part of the study design.

Teachers or researchers can use the following approach to take best advantage of the affordances of these three forms that have been analyzed as sufficiently equivalent. A test-form sequence could be randomly assigned to each student. For example, prior to instruction one student might be assigned the form sequence BAC to use form B prior to instruction, form A after instruction, and form C either as a delayed test later in the same year or during a subsequent course within the same school to measure student growth and ensure students are retaining the necessary information for their algebra and science course work. When researchers use this three-form approach with random assignment of form sequence, this means that a researcher is implementing experimental controls to reduce threats to validity and more rigorous study designs that can assess learning growth as well as retention at three time points (Bhattacherjee, 2021, Chapter 10). Such a rigorous research design would promote stronger conclusions that matter over time than typical pre-post study designs (Nurnberger-Haag, 2018, 2020).

There are research questions for which assessments other than binary calculations would be more valuable. Yet, the field of mathematics education could make faster and more cohesive progress to understand the learning of integer arithmetic if at least some studies use the same measures to allow for more direct comparisons or syntheses. Consider that a common measure means investigations of varied instructional methods or same methods across varied populations can be compared in meaningful ways to draw more effective conclusions (Campbell & Fiske, 1959; Wiersma & Jurs, 2009). Furthermore, better communication and building of knowledge across disciplinary borders such as mathematics education, educational psychology, and psychology are crucial to foster improvements in what we know about mathematics learning (Nurnberger-Haag, 2018). Common measures such as the ITPO have the potential to contribute to this important cross-disciplinary building. The instrument we provided with this study can now be used to raise the competitiveness of grant proposals about integer learning by fulfilling the criteria that at least some measures used in the data analysis plan have been psychometrically analyzed. For example, mathematics education researchers in the United States frequently apply for National Science Foundation DRK-12 funding, which require documentation of the planned instruments or measures and how these will validly and reliably serve the intended purpose of the study (Research on Learning in Formal and Informal Settings, 2020).

Integer test of primary operations could be used to discriminate student performance

From a psychometric perspective and if the purpose of using the test is to discriminate students from each other, then a limitation of the individual dimensions was low person reliability (i.e., .51 for AddD, .64 for SubD, .66 for MDdim). Low person reliability may suggest the measure would benefit from more person ability range or the addition of more items to the dimension (Linacre, 2019), but is also common for scales with less items. However, the reason for this low person reliability is consistent with goals of regular classroom instruction in which the goal is not to obtain a good spread of scores, but to ensure every student succeeds above a criterion. This assessment was administered after three weeks of integer arithmetic instruction, but before their unit review and test. At that time 66 (61.68%) students included in the Rasch analysis met a criterion of at least eight of 10 addition items and 46 (45.54%) students accurately answered at least nine of the 11 multiplication and division items correctly. In contrast, 81 (54%) students on the SubD were low scoring (i.e., fewer than four of ten items correct). The uneven proportions of scores lessen concern that should be added to these dimensions. Although the MDdim was considered underfit to the model, Rasch

measurement is more concerned with the practicality of a set of items rather than those items statistical fit to the model (Linacre, 2019).

Due to limited test-taker time and capacity, textbook developers who create unit tests and especially standardized test developers select a few or even just one item to represent an entire aspect of knowledge. Thus, this study provides necessary insights for these developers to choose which one to three item structures would best serve the intended purpose of the assessment. If the purpose of the assessment is to use limited items to identify whether a large group of students has mastered integer operation knowledge, such as on mathematics placement tests, then choose the most difficult items. In contrast, the purpose of standardized measures as used for college admissions and so forth is to intentionally ensure that some people will be viewed as lower-performing and others ranked highest. In the words of the statistical methods themselves *person-discrimination*, that is to separate persons based on their test performance. For these purposes, test developers should use a range of item difficulty: select one item that the majority can get correct (low negative logit), one item that is around the mean difficulty level, and one item that is difficult for most students (highest positive logit).

A mathematics coach in a school district or a classroom teacher could use the results of our study to create quick formative assessments of just a few items. Given that the purpose of formative assessment is to gauge overall class and individual student learning in order to inform changes to instruction during the unit (Black & William, 2010), we suggest reserving our test forms for summative assessments and create formative assessments by using **Table 14** as a guide to select different numbers than provided on the ITPO but with a range of the same problem structures. For example, early in the unit, a teacher might choose one item structure within Level 1 for each of addition and subtraction, and one item from Level 3 of each construct. Later in the unit after all four primary operations have been taught, it might be important for a formative assessment to ensure that a range of difficulty is included within the necessarily short ungraded formative assessment, such as: one Level 2 addition structure such as p+N, one subtraction item from each Level 1 to 4 (see **Table 14**), a multiplication item of the form n x N and an item of the form X divided by -1. Other item structures could be selected. We simply offered these detailed examples to support busy practicing teachers or mathematics curriculum specialists and coaches within school districts who might be responsible for creating such assessments.

If the purpose of an assessment is to identify those with typical knowledge, then based on this analysis, item structures located around the mean on each Item-Person Map might best serve this purpose, such as single digit p+N, n–N, and n×n. We will use the integer items in Ryan and Williams' (2007) study of 15,000 students ages 4 to 15 and interpret these using our study findings to explain this more clearly. The addition item 4 + $^{-}$5 (Ryan & Williams, 2007, p.212) would serve the purpose of identifying typical addition knowledge, because this addition problem structure of p + N was located around the mean on the ITPO Item-Person Map (see **Figure 1**). Ryan and Williams' (2007) findings that about 66% of students correctly answered this suggests it was a good item for this purpose. As Ryan and Williams (2007) explained the accuracy rates of the subtraction and division items were quite troubling at 35% and 44%, respectively. Yet, these items probably overestimated actual integer knowledge when we interpret the item structures in terms of our study. Their subtraction item of -6–3 with a structure of N–p was slightly above the mean, which might discriminate slightly above average performing students. The division item, however, -24 divided by 6 has the easiest division structure (N ÷ p), which means the same students who answered this item correctly if given the most difficult structure of N divided by -1 (-24 ÷ -1), fewer than 44% would accurately answer this.

One practical limitation of Rasch analysis is that to meet the assumptions requires that those students with perfect scores (at ceiling) are eliminated along with those who scored 0 (at floor). Consequently, we discourage researchers from conducting a study of integer performance using Rasch to analyze student performance in ways that have any stake for the individual student or teacher. Similarly, this type of analysis would be problematic if used to determine the efficacy of an instructional method, because after instruction our goal would be that every student meets a criterion of 100% and no students 0%. If one instructional method had a disproportionate number of students at ceiling and another a disproportionate number with 0, the students kept in the analysis would therefore lead to faulty interpretations and conclusions about method efficacy for real classroom learning. Thus, we agree with Stacey and Steinle (2006) that Rasch analysis must be applied judiciously. As they stated, "Rasch theory has helped us see that different aspects of learning need to be tracked with fundamentally different tools" (Stacey & Steinle, 2006, p. 91). That is, scholars must recognize for what purposes this tool is useful and make appropriate interpretations to avoid any untoward outcomes both theoretically in research and practically in the field.

CONCLUSIONS

The primary practical and theoretical contribution of this study is the measure itself. Although rare in mathematics education, the development of reliable and valid instruments is such an important foundation of rigorous research designs that entire funding sources are dedicated to supporting the development of measures for mathematics knowledge (Callingham & Bond, 2006; Research on Learning in Formal and Informal Settings, 2020). Moreover, we ensured that this instrument is ethically accessible to any stakeholder and given that only the directions are written in words, translations would be simple and are unlikely to impact the validity of the assessment. Therefore, this test should be broadly useful to teachers and scholars across the world in any language that uses Arabic numerals.

Three Forms Strengthen Practical Value and Statistical Conclusion Validity

The Integer Test of Primary Operations is unique in that it consists of three forms. The contribution of three forms is psychometrically and practically important. Psychometrically, the three forms strengthen the statistical content validity of the instrument. Published tests for mathematics knowledge typically rely on a single form, which means the numbers used in specific items do not have the variability needed to make strong claims about why the particular items might be more or less difficult

(Wiersma & Jurs, 2009). Hence, due to the variability of the numbers used within each item across forms of the ITPO, the statistical conclusion validity evidence of the ITPO is stronger than most published measures of mathematics learning. This increases the interpretability that the similar loadings of items with the same structure are due to the underlying mathematical structure (e.g., p–N) rather than due to the particular numbers chosen for the item (e.g., 2 – ⁻5, or 3 – ⁻8). Practically, the studies of learning that scholars most care about in education is dependent on the quality of the measures used to assess and interpret that learning. The ITPO consists of the three forms necessary to conduct rigorous pre-post-delayed assessments of learning in future study designs.

Moreover, even if the ITPO is not used in its entirety, the findings of this Rasch analysis inform learning, teaching and assessment of integer knowledge in detailed ways. This study provided a critical and empirically based lens with which to critique existing items on classroom tests as well as higher-stakes instruments. Our study provides guidelines about the structures of items to include on other assessments that textbook developers, standardized test developers, teachers or school districts create. To ensure that students can complete such an assessment of mathematics within a reasonable time, only a few items or sometimes a single item are included for each topic or learning objective. For example, if a stakeholder reviews an assessment and finds problem structures located in Level 1 of **Table 14**, these are the easiest problem structures. This means that the test overestimates students' integer operation competence that is so crucial for their future algebraic and other STEM learning.

Future Research Needed with Particular Populations

Due to the nature of the ITPO questions as straightforward calculation problems without words or contexts, we anticipate the measure would be reliable with other learners and populations. Nevertheless, in order to claim this, this study should be replicated with these populations. A priority for next analyses should be high school students taking introductory STEM courses such as algebra, chemistry, and physics who should have already mastered the integer arithmetic that is crucial for success in these fields. Similarly, given the importance of integer knowledge to subsequent success, this should also be replicated with adult learners in varied mathematical contexts such as classes in which integer arithmetic is retaught (i.e., developmental mathematics) or introductory collegiate mathematics courses (e.g., Precalculus or Calculus I). Moreover, given the importance of teachers having accurate mathematics content knowledge and recent studies about the difficulties prospective teachers have with negative numbers (Rosyidah, et al., 2021), it would be important to replicate this study with mathematics content courses for prospective teachers of all levels (i.e., early childhood, elementary, special education, as well as middle school and high school STEM). Moreover, the ITPO could now be used in studies that assess the relationships between integer skills and later applications within mathematics such as algebraic expressions (e.g., Is -4x + 3 equivalent to 3 – 4x?) as in Young and Booth (2019). Furthermore, the ITPO could enhance practice and research to ask what relationship is there between integer knowledge and success in STEM disciplines more broadly?

Instruments to Assess Integer Knowledge Needed

This study provided an instrument to assess knowledge of the four primary operations with negative numbers, which are an important set of constructs of integer knowledge. Additional measures are needed to assess other constructs of integer knowledge such as ordering numbers, using a number line, conceiving of the negative sign as an operation, and real-life applications (Nurnberger-Haag, 2018; Vlassis, 2008). Many researchers study these integer constructs from varied theoretical perspectives and with diverse ages. If the field were to have several reliable measures of integer constructs such that researchers from diverse perspectives could use and compare results with varied populations, scholars could better build on each other's work to advance the field's understanding about ways to facilitate integer arithmetic learning across levels of schooling. Indeed, to be competitive for grant funding, agencies such as the National Science Foundation and Institute for Education Sciences are increasingly requiring evidence that proposals use valid and reliable measures (Institute of Education Sciences & National Science Foundation [NSF], 2013; Research on Learning in Formal and Informal Settings, 2020). Thus, we offer the three forms of the Integer Test of Primary Operations as an important step toward this ideal state of the field.

Author contributions: All authors have sufficiently contributed to the study, and agreed with the results and conclusions.

Funding: These data were collected with the support of a Michigan State University College of Education Dissertation Expense Fellowship awarded to the first author.

Declaration of interest: No conflict of interest is declared by authors.

REFERENCES

Bhattacherjee, A. (2012). *Social science research: Principals, methods, and practices*. Textbooks Collection, 3. https://digitalcommons.usf.edu/oa_textbooks/3

Bishop, J. P., Lamb, L. L., Philipp, R. A., Whitacre, I., Schappelle, B. P., & Lewis, M. L. (2014). Obstacles and affordances for integer reasoning: An analysis of children's thinking and the history of mathematics. *Journal for Research in Mathematics Education, 45*, 19-61. https://doi.org/10.5951/jresematheduc.45.1.0019

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81-90. https://doi.org/10.1177/003172171009200119

Bofferding, L., & Wessman-Enzinger, N. (2018). *Exploring the integer addition and subtraction landscape: Perspectives on integer thinking.* Springer. https://doi.org/10.1007/978-3-319-90692-8

Bofferding, L., Wessman-Enzinger, N., Gallardo, A., Peled, I., & Salinas, G. (July, 2014). Discussion group: Negative numbers: Bridging contexts and symbols. In C. Nicol, P. Liljedahl, S. Oesterle, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36* (Vol. 1, pp. 240). Vancouver, Canada.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Callingham, R., & Bond, T. (2006). Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal*, *18*(2), 1-10. https://doi.org/10.1007/BF03217432

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. https://doi.org/10.1037/h0046016

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory* (2nd ed.). Cengage Learning.

Fischbein, E. (1987). *Intuition in science and mathematics: An educational approach*. D Reidel Publishing Company.

Fowler, F. J. (2013). *Survey research methods* (5th Ed.). University of Massachusetts.

French, D. (2001). Two minuses make a plus. *Mathematics in School, 30*(4), 32-33.

Institute of Education Sciences & National Science Foundation. (2013). *Common guidelines for education research and development*. U.S. Department of Education, Institute of Education Sciences & National Science Foundation. https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498-525. https://doi.org/10.2307/1418556

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. https://doi.org/10.1007/BF02288391

Liebeck, P. (1990). Scores and forfeits-an intuitive model for integer arithmetic. *Educational Studies in Mathematics, 21*, 221-239. https://doi.org/10.1007/BF00305091

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Mesa Press.

Linacre, J. M. (2013). Reliability separation, and strata: Percentage of sample in each level. *Rasch Measurement Transactions, 26*(4), 1399.

Linacre, J. M. (2018). *Detecting multidimensionality in Rasch data using Winsteps table 23 [Video file].* https://www.youtube.com/watch?v=sna19QemE50

Linacre, J. M. (2019). *A user's guide to Winsteps Rasch-model computer programs.* https://www.winsteps.com/a/Winsteps-Manual.pdf

National Governors Association Center for Best Practices. Council of Chief State School Officers. (2010). Common core state standards for mathematics. Washington, DC: Author. http://www.corestandards.org/the-standards

Norton, A., & Nurnberger-Haag, J. (2018). Bridging frameworks for understanding numerical cognition. Editorial in special issue Psychology and Mathematics: Bridging Approaches to Research for Understanding the Learning and Teaching of Number. A. Norton and J. Nurnberger Haag (Eds.). *Journal of Numerical Cognition, 4*(1), 1-8. https://doi.org/10.5964/jnc.v4i1.160

Nurnberger-Haag, J. (2015, November). How students' integer arithmetic learning depends on whether they walk a path or collect chips. In T. G. Bartell, K. N. Bieda, R. T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 165-172). Michigan State University. http://www.pmena.org/proceedings/

Nurnberger-Haag, J. (2018). Take it away or walk the other way? Finding positive solutions for integer subtraction. In L. Bofferding & N. Wessman Enzinger (Eds.), *Exploring the integer addition and subtraction landscape: Perspectives on integer thinking* (pp. 109-141). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-90692-8_5

Nurnberger-Haag, J. (2020). Predictions of integer model affordances pan out for short-term, but not longer-term, learning of initial integer constructs. In M. Gresalfi & I. S. Horn (Eds.), *The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020* (Vol. 2, pp. 665-668). International Society of the Learning Sciences. https://repository.isls.org/bitstream/1/6722/1/665-668.pdf

Periasamy, E., & Zaman, H. B. (2009, Nov 11-13). *Augmented reality as a remedial paradigm for negative numbers: Content aspect* [Paper presentation]. The Visual Informatics: Bridging Research and Practice, First International Visual Informatics Conference. https://doi.org/10.1007/978-3-642-05036-7_35

Pettis, C., & Glancy, A. W. (2015). Understanding students' challenges with integer addition and subtraction through analysis of representations. In T. G. Bartell, K. N. Bieda, R. T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 221-224). East Lansing, MI, USA.

Research on Learning in Formal and Informal Settings. (2020). *Discovery Research PreK-12 (DRK-12)*. National Science Foundation, Directorate for Education and Human Resources. https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=500047

Rosyidah, A. N. K., Maulyda, M. A., Jiwandono, I. S., Oktaviyanti, I., & Gunawan, G. (2021). Misconceptions and errors in integer operations: A study in preservice elementary school teachers (PGSD). *Journal of Physics: Conference Series, 1779*(1), 012078. https://doi.org/10.1088/1742-6596/1779/1/012078

Ryan, J., & Williams, J. (2007). *Children's mathematics 4-15: Learning from errors and misconceptions.* Open University Press.

Shaw, F. (1991). Descriptive IRT vs. prescriptive Rasch. *Rasch Measurement Transactions, 5*(1), 131.

Stacey, K., & Steinle, V. (2006). A case of the inapplicability of the Rasch Model: Mapping conceptual learning. *Mathematics Education Research Journal*, *18*(2), 77-92. https://doi.org/10.1007/BF03217437

Stephan, M., & Akyuz, D. (2012). A proposed instructional theory for integer addition and subtraction. *Journal for Research in Mathematics Education, 43*(4), 428-464. https://doi.org/10.5951/jresematheduc.43.4.0428

Thompson, P. W., & Dreyfus, T. (1988). Integers as transformations. *Journal for Research in Mathematics Education, 19*(2), 115-133. https://doi.org/10.2307/749406

Tsang, J. M., Blair, K. P., Bofferding, L., & Schwartz, D. L. (2015). Learning to "see" less than nothing: Putting perceptual skills to work for learning numerical structure. *Cognition and Instruction, 33*, 154-197. https://doi.org/10.1080/07370008.2015.1038539

Vlassis, J. (2008). The role of mathematical symbols in the development of number conceptualization: The case of the minus sign. *Philosophical Psychology, 21*(4), 555-570. https://doi.org/10.1080/09515080802285552

Wiersma, W., & Jurs, S. G. (2009). *Research methods in education* (9th Ed.) Pearson.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370-371.

Young, L. K., & Booth, J. L. (2019). Don't eliminate the negative: Influences of negative number magnitude knowledge on agebra performance and learning. *Journal of Educational Psychology, 112*(2), 384-396. https://doi.org/10.1037/edu0000371

APPENDIX: PRINTABLE TEST FORMS

Printable Test Forms

The three forms (A, B, and C) of the 30-item Integer Test of Primary Operations (ITPO) are included here in a printable format to facilitate their practical use. Recall that Q24 was removed from the 31-item test for reasons explained in the discussion section. Thus, if cross-referencing these forms with data in the results section, Q1-Q23 will match; however, Q24-Q30 on these printable forms should be compared to items Q25-Q31 reported in the results section.

Integer Test of Primary Operations (Form A)

Directions: Please answer the addition and subtraction problems.

1. $^{-}2 + {}^{-}5 =$
2. $^{-}2 + 3 =$
3. $^{-}13 + 8 =$
4. $^{-}16 + 23 =$
5. $^{-}3 - {}^{-}2 =$
6. $3 + {}^{-}6 =$
7. $^{-}2 + {}^{-}3 =$
8. $^{-}12 + 12 =$
9. $^{-}3 - 2 =$
10. $^{-}5 - {}^{-}2 =$
11. $2 + {}^{-}3 =$
12. $16 - 23 =$
13. $2 - {}^{-}3 =$
14. $12 - 15 =$
15. $^{-}8 - 13 =$
16. $^{-}2 - {}^{-}5 =$
17. $15 + {}^{-}15 =$
18. $17 - {}^{-}25 =$
19. $^{-}5 + 7 =$
20. $^{-}3 - {}^{-}6 =$

Directions: Please answer the multiplication and division problems.

21. $^{-}5 \times {}^{-}7 =$

22. $^{-}1 \times 19 =$

23. $^{-}20 \div {}^{-}4 =$

24. $5 \times {}^{-}7 =$

25. $22 \times {}^{-}1 =$

26. $^{-}30 \div {}^{-}10 =$

27. $^{-}2 \times {}^{-}11 =$

28. $^{-}24 \div {}^{-}1 =$

29. $^{-}12 \div 6 =$

30. $28 \div {}^{-}1 =$

Integer Test of Primary Operations (Form B)

Directions: Please answer the addition and subtraction problems.

1. $^{-}3 + {}^{-}5 =$
2. $^{-}2 + 3 =$
3. $^{-}15 + 9 =$
4. $^{-}18 + 25 =$
5. $^{-}3 - {}^{-}2 =$
6. $2 + {}^{-}5 =$
7. $^{-}2 + {}^{-}3 =$
8. $^{-}17 + 17 =$
9. $^{-}3 - 2 =$
10. $^{-}4 - {}^{-}3 =$
11. $2 + {}^{-}3 =$
12. $17 - 25 =$
13. $2 - {}^{-}3 =$
14. $11 - 16 =$
15. $^{-}6 - 11 =$
16. $^{-}4 - {}^{-}5 =$
17. $13 + {}^{-}13 =$
18. $16 - {}^{-}23 =$
19. $^{-}4 + 8 =$
20. $^{-}2 - {}^{-}4 =$

Directions: Please answer the multiplication and division problems.

21. $^{-}5 \times {}^{-}9 =$

22. $^{-}1 \times 18 =$

23. $^{-}20 \div {}^{-}5 =$

24. $5 \times {}^{-}6 =$

25. $23 \times {}^{-}1 =$

26. $^{-}20 \div {}^{-}10 =$

27. $^{-}2 \times {}^{-}12 =$

28. $^{-}22 \div {}^{-}1 =$

29. $^{-}12 \div 4 =$

30. $24 \div {}^{-}1 =$

Integer Test of Primary Operations (Form C)

Directions: Please answer the addition and subtraction problems.

1. $^{-}2 + {}^{-}6 =$
2. $^{-}2 + 3 =$
3. $^{-}12 + 7 =$
4. $^{-}14 + 21 =$
5. $^{-}3 - {}^{-}2 =$
6. $4 + {}^{-}7 =$
7. $^{-}2 + {}^{-}3 =$
8. $^{-}19 + 19 =$
9. $^{-}3 - 2 =$
10. $^{-}6 - {}^{-}2 =$
11. $2 + {}^{-}3 =$
12. $15 - 22 =$
13. $2 - {}^{-}3 =$
14. $12 - 16 =$
15. $^{-}8 - 12 =$
16. $^{-}3 - {}^{-}5 =$
17. $14 + {}^{-}14 =$
18. $18 - {}^{-}23 =$
19. $^{-}4 + 7 =$
20. $^{-}2 - {}^{-}6 =$

Directions: Please answer the multiplication and division problems.

21. $^{-}5 \times {}^{-}6 =$

22. $^{-}1 \times 17 =$

23. $^{-}16 \div {}^{-}4 =$

24. $5 \times {}^{-}8 =$

25. $25 \times {}^{-}1 =$

26. $^{-}20 \div {}^{-}5 =$

27. $^{-}2 \times {}^{-}13 =$

28. $^{-}21 \div {}^{-}1 =$

29. $^{-}12 \div 3 =$

30. $26 \div {}^{-}1 =$