MEASURES OF STRATEGY FLEXIBILITY IN AN ABBREVIATED TRI-PHASE TEST ON LINEAR AND QUADRATIC EQUATION SOLVING

Anne-Maria Ernvall-Hytönen, Peter Hästö, Sara Parikka and Heidi Krzywacki[1]

University of Helsinki

ABSTRACT

This methodological paper is concerned with the issue of measuring strategy flexibility. We devised a 6-item tri-phase test from the domains of linear and quadratic equations. The test was completed by 306 students in an introductory university mathematics course. To validate the test we consider alternative definitions of key variables and conduct psychometric analyses of our strategy flexibility measures' internal consistency and factorial validity. We show that various choices for the definition of variables have no significant impact and that the test is internally consistent despite the inclusion of both linear and quadratic equations.

INTRODUCTION

Flexibility has been widely studied in mathematics education over the past decade, especially in China, Europe and North America (Hästö & Star, 2024; Hong et al., 2023; Verschaffel, 2024). Flexibility is also recognized as an explicit goal in mathematics instruction in many policy documents (e.g., NCTM, 2023), but, curiously, not in Finland (POPS, 2014; LOPS, 2019), the context of this research. Nevertheless, Finnish middle-school and high-school students have shown comparable levels of flexibility to other countries (McMullen et al., 2016; Star et al., 2022).

Many forms of mathematical flexibility have been considered (Heinze et al., 2009; Hickendorff et al., 2022) but our study is restricted to strategy, or procedural, flexibility. In his recent literature review, Verschaffel (2024) defined strategy flexibility by two criteria: (A) knowledge of multiple strategies and (B) tendency and ability to select the most appropriate strategy for the problem at hand. Furthermore, he points out that appropriateness has been understood in a variety of ways. The characteristics of the mathematical problem are always considered: using the same strategy regardless of the specifics of the problem is a sign of inflexibility. Some researchers, especially ones of elementary mathematics education, also consider characteristics of the person solving the problem and the sociocultural context in which the work is carried out (Hong

[1] The authors are listed in alphabetical order.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

et al., 2023; Verschaffel, 2024). The last two features are not utilized in this article in the definition of flexibility.

In this methodological paper, we consider the issue of measuring strategy flexibility following ideas of Xu et al. (2017) who devised a tri-phase test in the domain of linear equation solving. Their test operationalizes criteria (A) and (B) mentioned above and considers appropriateness in relation to problem characteristics only. While the main objective of the test is to measure strategy flexibility, it also yields a measure of general mathematical proficiency in the test domain that Xu et al. (2017) call "accuracy".

The tri-phase test was used by Xu et al. (2017) and Liu et al. (2018) to study strategy flexibility of Chinese middle-school students. The same test was then used by Star et al. (2022) and Jiang et al. (2023) in an international comparison of Finnish, Spanish and Swedish middle- and high-school students' strategy flexibility. these studies used a 12-item test of linear equation solving. The procedure underlying the tri-phase test was adapted by Maciejewski (2022) to measure flexibility in differentiation tasks of students in a university calculus class. We independently adapted the tri-phase approach and conducted a pilot study using a 6-item test including linear equations, a quadratic equation and a system of equations (Ernvall-Hytönen et al., 2022). This paper builds on the last-mentioned study. We have revised our test based on the pilot to include 4 linear equations and 2 quadratic equations. Here, we validate our test more rigorously in the university setting with a larger and more diverse group of students.

Validity means that "evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014). Establishing validity is related to five factors, namely evidence based on (1) test content, (2) response processes, (3) internal structure and (4) relations to other variables, as well as (5) consequences of test use (AERA et al., 2014, pp. 13-19). In this article we focus on factor (3) and touch upon factors (1) and (2). To this end, we considered alternative definitions of key variables and conducted psychometric analyses to establish internal consistency and factorial validity of our measures of strategy flexibility.

Of particular interest and in contrast to previous research, we compare flexibility demonstrated in the linear equation items with that in quadratic equation items. To validate our instrument we want to show that these items do not form overly independent scales. Our main hypotheses are as follows:

1. Different definitions of the accuracy-variable in the tri-phase test have negligible impact.
2. Different definitions of the flexibility-variable in the tri-phase test have negligible impact.
3. Items relate to the flexibility-variable in a similar way regardless of their equation type.
4. Shortening the test further will decrease its reliability.

METHODS

Participants

Students were given a tri-phase test at the beginning of an introductory mathematics course in the autumn of 2022 as part of the course work. Participants' consent to use their responses for research was asked and the same course credit was received for completing the test regardless of research consent. Altogether 306 students consented to participate in the study and submitted valid responses. This is the vast majority of the students who received the invitation to participate. Participants mainly majored in mathematics, natural sciences, computer science, statistics, and economics.

To check the statistical significance of certain observations we use a simple resampling technique of calculating a statistic on 10 randomly chosen subgroups. For this purpose, we created 10 random subgroups of the original group of students. Each student had an independent 50 % chance of being included in each group. There was no stratification. The size of the resulting subgroups varied between 139 and 161. We use this as a simple alternative to bootstraping, see Babu (1992) for a discussion on the subsample method and further references.

Data gathering

Our tri-phase test contains six mathematical tasks: four linear and two quadratic equations. The untimed test is completed in the Moodle environment of the course. Test instructions were provided in Moodle, but participants wrote their answers on paper and submitted their responses as photos in Moodle. Instructions for each phase were provided only after the previous phase had been completed, similar to earlier pen-and-pencil tri-phase tests. As preparation, students were asked to take two blank A4 papers and divide each into a 3*3 grid. Students were instructed to solve the tasks individually using only pen and paper and they had a bit over a week to solve the tasks.

In Phase 1 of the test, students solved all six tasks and wrote each solution in the leftmost column of their grid. In Phase 2, students were instructed to generate up to two additional solutions to each task and write them down in columns 2 and 3. In Phase 3, students were asked to circle the solution they considered "the best" for each task. These three phases are identical to those used by Xu et al. (2017) and others and constitute the "tri-phase approach" mentioned in the introduction.

Coding

The coding was carried out by two researchers (co-author Parikka and a research assistant) independently. Disagreements were discussed and when necessary resolved with the aid of co-author Ernvall-Hytönen. Each solution was coded in terms of accuracy as correct (1) or incorrect (0) and a strategy was identified based on the first step(s) of the solution. Each solution was categorized as situational, generic or other strategy in a manner explained next. The selection of the "best" solution for the task was coded if there were more than one solution and otherwise ignored.

In the tri-phase test in the domain of linear equations, strategy coding has been carried out with categories "standard" and "innovative/situational" (Xu et al., 2017; Star et al., 2022). However, other domains do not necessarily have a standard solution strategy. For this reason we prefer the term "generic" for a strategy which is widely applicable while we continue to use the term "situational" for a strategy which utilizes the situation specific features of the problem for a more efficient solution. We next describe in detail what this means in our 6 tasks; Table 1 summarizes the coding and the appendix contains an example solution for each type of strategy.

Table 1. Essential elements for strategy categorisation by task.

Task	Generic	Situational
1. $4(x+\frac{3}{5})=12$	Distribute	Divide by 4
2. $5(x+\frac{3}{7})+3(x+\frac{3}{7})=16$	Distribute	Combine like terms
3. $\frac{2x-6}{2}+\frac{6x-18}{3}=5$	G1. Expand with 3 and 2 G2. Multiply by 6	S1. Perform divisions S2. Take $2x-6$ as a common factor
4. $\frac{5x-5}{5}+\frac{6x-6}{6}=5$	G1. Expand with 5 and 6 G2. Multiply by 30	S1. Perform divisions S2. Take $x-1$ as a common factor
5. $2x^2+4x=0$	G1. Use quadratic formula G2. Complete the square	S1. Use zero-product property S2. Divide by$x$
6. $(x-2)(x+3)=0$	G1. Use quadratic formula G2. Complete the square	Use zero-product property

In Tasks 1 and 2 a strategy was considered generic if solving the equation started by distributing the parentheses. In Task 1, the situational strategy was to divide by 4 resulting in $x+\frac{3}{5}=3$ as the first intermediate step. In Task 2, the situational strategy was to combine like terms resulting in $8(x+\frac{3}{7})=16$.

In the generic strategy of Tasks 3 and 4, the fractions are either expanded to a common denominator (G1) or multiplied by the least common multiple (G2). The situational strategy involved either simplifying the fractions by performing the division (S1) or taking a common factor (S2) of either $2x-6$ (Task 3) or $x-1$ (Task 4). For instance, S1 in Task 3 starts with $x-3+2x-6=5$ and G2 in Task 4 starts with $6(5x-5)+5(6x-6)=150$.

In Tasks 5 and 6, the generic strategy was either to use the quadratic formula (G1) or to complete the square (G2). In Task 5, the situational strategy was either to take x as a common factor and use the zero-product property (S1) or to divide the equation by x (S2). In Task 6, the only situational strategy involved the zero-product property.

If a solution did not match any of the criteria mentioned above, it was categorized as "other". Only 1.0-3.4 % of solutions fell into this catch-all category, see Table 2. Thus, subsequent analyses concentrate on generic and situational strategies.

Table 2. Number of "other" solutions compared to the total number of solutions by task.

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
6/598	14/562	12/568	18/550	17/542	17/501

ANALYSIS

In this section we analyze the data to validate our version of the tri-phase test. We start with discussing accuracy and then examine different ways of measuring flexibility. The section is concluded by a study of the structure of the test and how different parts of it measure flexibility.

Definition of accuracy variables

We defined three accuracy variables. Traditional accuracy is 1 if the Phase 1 solution attempt is correct, and 0 if it is not, or if the task has not been attempted. This definition of accuracy has been used in prior research (Xu et al., 2017; Liu et al., 2018; Star et al., 2022; Jiang et al., 2023). We also defined two new accuracy variables. The partial accuracy score of a task is the number of correct solutions to the task divided by the total number of attempts to that task. The strict accuracy score of a task is 1 if all attempts are correct and 0 otherwise. For example, if a student gives two solutions, the Phase 1 one correct and other one incorrect, then traditional accuracy is 1, partial accuracy is 0.5 and strict accuracy is 0. The average accuracy in each task with these measures is shown in Table 3.

The AccTrad variable is obtained by adding a participant's traditional accuracy scores over all six tasks, similarly for the partial accuracy (AccPart) and strict accuracy (AccStrict). An alternative way to measure mathematical proficiency is by the total number of correct solutions (#cor) or even the total number of solutions (#sol). The correlations between these variables is presented in Table 4.

Table 3. Average accuracy scores in each task.

Average accuracy	T1	T2	T3	T4	T5	T6
Traditional	0.88	0.82	0.80	0.84	0.78	0.79
Partial	0.87	0.81	0.78	0.83	0.77	0.77
Strict	0.82	0.76	0.73	0.80	0.72	0.71

Table 4. Pearson correlations between proposed measures of mathematical proficiency.

Variable	1	2	3	4	5
1. #sol	-				
2. #cor	0.847	-			
3. AccTrad	0.524	0.811	-		
4. AccPart	0.329	0.787	0.957	-	
5. AccStrict	0.238	0.725	0.857	0.884	-

Notes. All correlations are significant at the 0.01 level.

In Table 4, we see two blocks, namely those formed by the first two variables (#sol and #cor), and by the last four (#cor, AccStrict, AccPart and AccTrad). The lower correlation between #sol and AccPart or AccStrict is no surprise, because the higher number of total solutions makes it harder to get good points in partial accuracy and strict accuracy.

To study the relations between these variables further, we extracted 1- and 2-factor solutions using principal component analysis; the eigenvalues for the first two factors were 4.050 and 0.787, with the remaining ones much smaller. The one-factor solution explains 81.0% of the variance with loadings (0.71, 0.95, 0.96, 0.92, 0.94). Overall, all variables measure mostly the same thing, despite the inherent negative relation between #sol and AccPart or AccStrict noted above. However, a 2-factor solution provides much better fit. After varimax with Kaiser rotation, we get the components with loadings (0.21, 0.64, 0.94, 0.95, 0.90) and (0.97, 0.75, 0.32, 0.22, 0.34), explaining 61.1% and 35.6% of the variance, respectively, and 96.7 % between them. We conclude that the total number of solutions measures something slightly different from the other variables and exclude it from further analysis.

To validate the observation that the remaining four variables behave the same way, we utilized the 10 subgroups described in the subsection "Participants". In each subgroup, we calculated the correlations between the variables. The range of these correlations is shown in Table 5. The results show that AccPart is extremely close to both AccTrad and AccStrict. Furthermore, AccTrad and AccStrict are very close to one another and all three "Acc"-variables are strongly related to the total number of correct solutions.

This confirms our first hypothesis that different definitions of the accuracy-variable in the tri-phase test have negligible impact insofar as the variables AccTrad, AccPart and AccStrict are concerned.

Table 5. Ranges of Pearson correlations in 10 random subgroups.

Variable	1	2	3	4
1. #cor	-			
2. AccTrad	0.780-0.841	-		
3. AccPart	0.812-0.859	0.950-0.970	-	
4. AccStrict	0.748-0.814	0.868-0.926	0.956-0.969	-

Definition of flexibility variables

While the general idea of the operationalization of flexibility in tri-phase tests is well established, details vary between the earlier papers: Xu et al. (2017) and Liu et al. (2018) use the practical and potential flexibility constructs whereas Star et al. (2022) and Jiang et al. (2023) use flexibility and spontaneous flexibility. We build upon the latter.

We define three flexibility variables. In each case the maximum score of 1 is given if a task has at least one generic strategy and one situational one and additionally the situational strategy is chosen as best. This operationalizes the definitions of flexibility in terms of criteria (A) and (B) given in the introduction. However, we wanted to investigate possibilities of identifying also partially flexible students. For this, the theory does not provide as much guidance. This ambiguity gives rise to the three different flexibility-variables.

These variables differ in the use of a partial score 0.5; the definitions are given in Table 6. FlexTrad is the flexibility variable used earlier in the literature (e.g. Star et al., 2022) and does not include the partial score. Star et al. (2022) utilized a second construct called Potential Flexibility. Their rationale behind this construct is the expectation that fulfilling some of the criteria of flexibility is an indication that the student has greater probability of developing flexibilility or showing it in other tasks. Our second variable, extended flexibility FlexExt, gives students a score of 0.5 if they fulfill the requirements of Potential Flexibility from Star et al. (2022). FlexMod is a variant of this which is slightly more restrictive in that one situational solution on its own is not sufficient for a partial score. Note that the score 0.5 can be obtained in several ways. They were lumped together by Star et al. (2022; cf. Tables 2, 3 & 5) and we adhere to this.

A slightly illogical detail in FlexExt is that it awards more points if a student has only one solution, which is situational, than if a student has done two solutions, both situational, and not chosen any solution as the best. This quirk follows from the earlier definition of Potential Flexibility (Star et al., 2022). Although this is uncommon in practice (e.g. in our data from 306 students only 16 tasks' solutions contained two situational strategies without a choice of best), it is nevertheless desirable to have an operationalization which is logically aligned with the contruct.

The student's TradFlex, FlexExt and FlexMod values are obtained as a sum over the six tasks. Thus, these variables have values between 0 and 6.

Table 6. Definitions of the flexibility variables.

Score	FlexTrad	FlexMod	FlexExt
		The solutions to a task include…	
1	a situational solution chosen as best and a generic solution.	a situational solution chosen as best and a generic solution.	a situational solution chosen as best and a generic solution.
0.5	-	1) a situational solution chosen as best and 1-2 additional solutions that are not generic - or - 2) both generic and situational solutions, but situational not chosen as best	1) a situational solution chosen as best and 1-2 additional solutions that are not generic - or - 2) both generic and situational solutions, but situational not chosen as best - or - 3) only one solution and it is situational
0	otherwise	otherwise	otherwise

Flexibility in different tasks

We first calculated the average number of solutions to each task, and the average of FlexTrad, FlexExt and FlexMod scores, which are shown in Table 7.

In Table 7, we notice that the FlexTrad values of Task 5 are somewhat low compared to the other tasks. We investigate this observation closer. The number of various types of solutions is shown in Table 8.

Table 7. The average number of solutions and flexibility scores in each task.

Task	#sol	FlexTrad	FlexMod	FlexExt
1	1.95	0.62	0.70	0.72
2	1.84	0.45	0.52	0.56
3	1.86	0.51	0.58	0.62
4	1.80	0.51	0.59	0.66
5	1.77	0.36	0.50	0.58
6	1.64	0.50	0.55	0.60

Table 8. Number and percentage of students who have given different types of solutions to each task.

Task	Has solution	Has generic	Has situational	Has both
1	303 (99.0%)	285 (93.1%)	253 (82.7%)	235 (76.8%)
2	302 (98.7%)	274 (89.5%)	204 (66.7%)	176 (57.5%)
3	300 (98.7%)	257 (84.0%)	229 (74.8%)	187 (61.1%)
4	297 (97.1%)	222 (72.5%)	252 (82.4%)	182 (59.5%)
5	292 (95.4%)	178 (58.2%)	249 (81.4%)	140 (45.8%)
6	296 (96.7%)	248 (81.0%)	218 (71.2%)	175 (57.2%)

The outlier-nature of Task 5 is even more apparent in Table 8, more specifically in the "Has generic"-column. The low value of 58.2 % means that a relatively small number of students produced a solution in Task 5 using the quadratic formula or by completing the square.

We calculated the correlations between the flexibility variables (see Table 9). These very high values indicate extremely strong relationships between the variables, as can be expected from the definition.

Table 9. Pearson correlations between the different "flexibility"-variables.

Variable	1	2	3
1. FlexTrad	-		
2. FlexExt	0.939	-	
3. FlexMod	0.949	0.967	-

It follows from the definitions that FlexTrad ≤ FlexMod ≤ FlexExt. To further compare the variables, we calculated pairwise differences between the variables. The values of FlexMod and FlexExt coincided for 212 students and differed by 0.5 for 46 students and by 1 for 25 students. For the remaining 23 students, the difference was bigger with a maximum of 3 for one student. Between FlexMod and FlexTrad, the difference was 0 for 161 students, 0.5 for 71 students and 1 for 34 students. For the remaining 40 students, the difference was bigger with a maximum of 3 for 5 students. The distribution of the difference between these two variables is shown in Figure 1. We further note that out of the 88 students with zero FlexTrad score, 45 nevertheless had positive FlexMod.

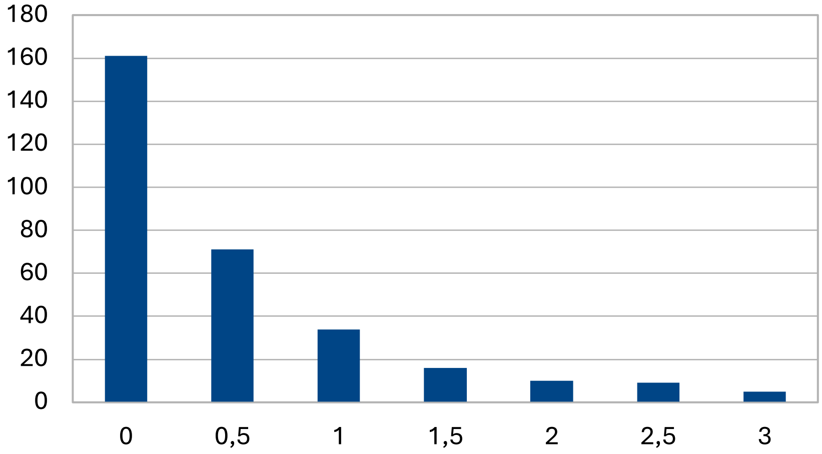

Figure 1. Distribution of difference between FlexTrad and FlexMod.

This confirms our second hypothesis that different definitions of the flexibility variable in the tri-phase test have negligible impact. Since all the flexibility variables measure the same quality, it suffices to look at one of them. For the rest of the section we will consider only the FlexMod-variable. This choice is justified in the discussion.

Structure of the test

The internal consistency of the test was measured with the Cronbach alpha coefficient. For the 6 FlexMod items it equals 0.849. One aim was to see how well the results from linear and quadratic equations are related. The subscale consisting of linear equations (items 1-4) has Cronbach alpha equal to 0.795 and the quadratic equations (items 5&6) subscale has Cronbach alpha equal to 0.805. Thus even these very short subscales have reasonable internal consistency. The Pearson correlation between the linear and quadratic subscales is 0.630 and it is significant at the 0.01 level.

The correlations between the FlexMod scores of different tasks is shown in Table 10. The highest correlations are between Tasks 1&2, Tasks 3&4 and Tasks 5&6.

Table 10. Pearson correlations between FlexMod in different tasks.

Variable	1	2	3	4	5	6
1. FlexMod1	-					
2. FlexMod2	0.527	-				
3. FlexMod3	0.489	0.383	-			
4. FlexMod4	0.460	0.377	0.731	-		
5. FlexMod5	0.412	0.428	0.413	0.458	-	
6. FlexMod6	0.440	0.500	0.478	0.492	0.675	-

To further study the structure, we used principal component analysis to extract 1- and 3-factors solutions. The 3-factor solution was mainly guided by theoretical considerations since the structure of tasks 1&2, 3&4 and 5&6 were pairwise similar. In terms of eigenvalues, only the first one exceeds 1, the second and third were around 0.75 and the remaining ones were considerably lower. Thus there some datadriven justification also for extracting three components.

In the 1-factor solution the component had loadings (0.73, 0.70, 0.78, 0.78, 0.75, 0.80) and explained 57.1 % of the variance. So by and large every item has a similar and rather strong correlation to this component.

For the 3-factor solution we used varimax rotation with Kaiser normalization. The factor loadings were (0.37, 0.10, 0.87, 0.86, 0.22, 0.27), (0.14, 0.33, 0.21, 0.28, 0.88, 0.82) and (0.79, 0.83, 0.24, 0.18, 0.19, 0.28) and together they explain 82,8 % of the variance of the 6 items. The factor loadings mostly indicate the same pairs as with correlations, 1&2, 3&4 and 5&6, although the first factor shows a correlation between Task 1 and the second pair and second factor shows a correlation between Task 2 and the last pair.

We thus created new subscales by adding the scores of the pairs. Table 11 shows the correlations between these variables. We see that all correlations are essentially the same, and the correlation between the linear subscales 1&2 and 3&4 is in fact the lowest. To investigate the significance of the differences between these correlations we used the 10 random groups described in subsection "Participants". The ranges of correlations for the subsets are shown in Table 11 to the right.

Table 11. Pearson correlations between subscales in the whole group and in 10 random subgroups.

	Whole group			Subgroups		
FlexMod	1	2	3	1	2	3
1. Tasks 1&2	-			-		
2. Tasks 3&4	0.523	-		0.521-0.636	-	
3. Tasks 5&6	0.559	0.542	-	0.475-0.654	0.492-0.644	-

Notes. All correlations are significant at the 0.01 level.

The significant overlap of the correlation ranges of the subsets confirms that there is no statistically significant difference between the correlations of the pairs. In summary, we have shown that the linear and quadratic equations subscales are highly correlated and quadratic equations differ from linear equations no more than different types of linear equations differ from one another. This confirms the third hypothesis.

The structure of the tasks as well as the correlations indicate that Tasks 1&2, 3&4 and 5&6 form pairs with more similar behavior within the pair. We wanted to see if it is possible to utilize this to shorten the test by using only the even or odd tasks of the test. The Cronbach alphas of the odd tasks and the even task subscales are 0.699 and 0.716, respectively. We note that these values are markedly lower than the alphas of the whole test and the linear and quadratic item subscales. The Pearson correlation of FlexMod between even and odd tasks is 0.807 and it is significant at the 0.01 level. Thus these subscales measure the same thing, but have lower internal consistency. This confirms Hypothesis 4.

DISCUSSION

We have explored the validity of a tri-phase test based on 6 items from the domains of linear and quadratic equations. We have particularly concentrated on different ways of measuring accuracy and flexibility, and the general structure of the test.

Hypothesis 1. Different definitions of the accuracy variable in the tri-phase test have negligible impact

We defined three variables for measuring accuracy: traditional accuracy, partial accuracy and strict accuracy. Furthermore, we considered the possibility of using the total number of solutions or the total number of correct solutions as alternatives. We concluded that all five variables measure roughly the same thing although the three accuracy variables were the most closely related by some margin (Tables 4&5).

One can argue that producing more correct solutions is always good, so that the number of correct solutions is the best measure of mathematical proficiency. However, attempting more solutions is also very likely to increase the total number of correct solutions, especially in a test without time-limit. Therefore the total number of solutions will contain an artificial correlation with flexibility which also requires the presence of multiple solutions.

Comparing the three accuracy-variables, their correlations with one another are very high. Since traditional accuracy only takes into consideration the first phase solution it does not register mathematical proficiency shown in Phase 2 of the test. Furthermore, this may also be confounded by flexibility, since those students who use the situational solution in the first phase have less complex calculations to perform. On the other hand, strict accuracy or partial accuracy have an inherent negative relation to the number of solutions given to a task.

Ideally, accuracy would be measured with items that cannot be solved in simpler ways using flexibility. However, this would necessitate additional test items, so it is not a practical suggestion in many situations. Since there are no clear advantages with the alternative definitions of accuracy, it seems that the best option is to use the AccTrad variable for comparability with earlier studies (e.g., Star et al., 2022; Jiang et al., 2023).

Hypothesis 2. Different definitions of the flexibility variable in the tri-phase test have negligible impact

In the context of linear equations, flexibility is operationalized in the tri-phase test by requiring a standard solution and an innovative solution (Xu et al., 2017; Star et al., 2022). In the context of linear equations in one variable there exists a widely taught standard algorithm which is directly related to the standard solution coding. In a more general context this is not usually the case and what should be considered "standard" is more debatable. We therefore dropped the concept of "standard" and instead required that among the 2-3 solutions presented at least one should be generic (i.e. apply to a large range of problems) and one should be situational (i.e. utilize the features of the particular problem), see Tables 1&6 for details.

This revised way of categorizing solutions seemed to work well in all tasks, with the possible exception of Task 5, see Tables 7&8. The number of generic strategies in Task 5 was unexpectedly low. One may argue that the situational solution is substantially superior for the equation $2x^2 + 4x = 0$ in Task 5. However, if one is asked to solve a task in several ways, it still seems reasonable to require also a generic solution. The results show that in Task 5 the biggest impediment to being deemed flexible was the lack of a generic solution. While unusual, this is not unprecedented: Star et al. (2022, Figure 1) found that Swedish middle- and high-school students were more adept at the situational rather than the generic solution in our Tasks 3&4 (Tasks 10&12 in the reference). That article suggested that the reason for this "reversal" of roles is that Swedish students had not learned the generic strategy. This is not the case for the students in this article since everyone has been taught the use of the quadratic formula. A more plausible explanation in our case is that they have been taught to solve such equations with the situational strategy and fail to notice the possibility of applying also the generic strategy. This is an alternative explanation also for the mentioned behavior of Swedish students in the earlier study.

In an earlier paper (Ernvall-Hytönen et al., 2022), we resolved the anomaly of lack of generic solutions in a quadratic equation by deeming students flexible also when they presented two situational solutions. This was mostly caused by observations from strategies chosen for the solution of a system of two linear equations, whereas the strategy choices in the quadratic equation in that test were more ambiguous. However, the more comprehensive analysis in this paper (e.g. Table 8) and the arguments above concerning Task 5 suggest that it was not necessary to make this revision for the quadratic equation. One possibility for future research is to see if asking for more than 3 solutions would yield a higher amount of generic solutions. However, from Table 7 we see that already more than one solution attempt on average was left "unused", so asking for more solutions might not be so effective.

The variable FlexTrad in this article is the traditional flexibility used in earlier articles. It is all or nothing: if all the conditions are met, then one gets a point, and otherwise nothing. We investigated the viability of a more nuanced measure of flexibility by combining the variables "flexibility" and "potential flexibility" from earlier stuies into one variable. Note that FlexTrad is a categorical variable but the same is not true for the variants (this was pointed out to us by Jake McMullen). However, the next phase in earlier studies (Xu et al., 2017; Liu et al., 2018; Star et al., 2022; Maciejewski, 2022; Jiang et al., 2023) is to add the scores over all tasks, after which the resulting variable is treated as an interval or ratio scale. Therefore, our departure from precedent is not as large as it seems at first.

We defined FlexExt as a combination of traditional flexibility and potential flexibility earlier used in research. This variable had some illogical features and so we also considered its variant FlexMod which is more streamlined. Furthermore, multiple solutions are usually considered a core feature of flexibility (Verschaffel, 2024), but this was not part of FlexExt. See Table 6 for the precise definitions. Based on the analysis performed, these variables are very highly aligned with correlations around 0.95 (Table 9). For the majority of students the scores on the variables are identical, and few have differences larger than 1. However, in the aggregate FlexMod and FlexExt still give noticeably larger averages (Table 7). In particular, the latter two variables are less susceptible to the exceptional nature of answers to Task 5. Furthermore, the number of students with a score of zero is markedly smaller. The new variables are thus able to pick up on smaller differences in performance. Based on these considerations, we recommend using FlexMod because it is more nuanced than FlexTrad and does not have the illogicality of FlexExt.

Hypothesis 3. Items relate to the flexibility variable in a similar way regardless of their equation type

Having established a valid way of defining the FlexMod variable, we proceeded to check the internal consistency of the scale and the structure of the test. The Cronbach alpha of the whole scale is 0.849 which exceeds the threshold 0.8 recommended by Nunnally (1978) for applied research. This is slightly smaller than the 0.92 value for the 12-item test found by Xu et al. (2017), as is to be expected with a shorter test with a more diverse set of items.

Quadratic equations were one novelty compared to earlier studies of flexibility. Thus we were especially interested in verifying that the subscales of linear equations (Tasks 1-4) and quadratic equations (Tasks 5&6) are compatible. The alphas of both subscales were around 0.80. Therefore they can be considered adequate subscales. The Pearson correlation between the subscales was 0.630, and significant at the 0.01 level. This combined with the reasonable Cronbach alpha for the whole test indicate that the items form a coherent measure of flexibility.

The linear equations formed two groups, as Tasks 1&2 had different structure and situational strategies compared to Tasks 3&4. The analysis showed that the differences between the pairs 1&2 and 3&4 was very similar to the difference of either of these pairs compared to 5&6 (see Table 11). This is further evidence that quadratic equations can be combined with linear equations of different types in a measure of flexibility. This conclusion was also supported by a principal component analysis of the 6 items. Thus the evidence supports Hypothesis 3.

Hypothesis 4. Shortening the test further will decrease its reliability

Since the items in each of the pairs 1&2, 3&4 and 5&6 were similar, we also considered the possibility of constructing an abbreviated measure of flexibility by using only one item from each pair. The subscales of odd and even items had Cronbach alphas around 0.7, which Nunally (1978) suggests is sufficient for early stage research. The correlation between the two scales exceeds 0.8. Therefore, it is possible to make a smaller flexibility test which measures roughly the same thing by taking just half of the questions at the expense of some reliability. An alternative with higher internal consistency (but lower correlation with one another) is to use only Tasks 1-4 or 5&6, as these subscales have alphas around 0.8 although (and probably because of) having only linear or quadratic equations, not both.

CONCLUSION

We have investigated the validity of a six-item test for strategy flexibility. We have shown that various choices for the definition of variables have no significant impact. The test is internally consistent despite the inclusion of both linear and quadratic equations. Finally, it is possible to abbreviate the test further to a 3-item version but this test has lower internal consistency.

ACKNOWLEDGEMENTS

We thank the referees for their helpful comments.

REFERENCES

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Babu, G.J. (1992). Subsample and half-sample methods. *Annals of the Institute of Statistical Mathematics, 44*, 703-720. https://doi.org/10.1007/BF00053399

Ernvall-Hytönen, A.-M., Hästö, P., Krzywacki, H., & Parikka, S. (2022). Procedural flexibility in early university mathematics. *FMSERA Journal 5*, 46-60.

Heinze, A., & Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM – Mathematics Education, 41*, 535-540. https://doi.apa.org/doi/10.1007/s11858-009-0214-4

Hästö, P., & Star, J.R. (2024). *Strategic flexibility in university mathematics.* Notices of the American Mathematical Society, *71* (7), 917-923.

Hickendorff, M., McMullen, J., & Verschaffel, L. (2022). Mathematical flexibility: Theoretical, methodological, and educational considerations. *Journal of Numerical Cognition, 8*, 326–334. https://doi.org/10.5964/jnc.10085

Hong, W., Star, J.R., Liu, R.-D., Jiang, R., & Fu, X. (2023). A Systematic review of mathematical flexibility: Concepts, measurements, and related research. *Educational Psychology Review, 35,* article 104. https://doi.org/10.1007/s10648-023-09825-2

Jiang, R., Star, J.R., Hästö, P., Li, L., Liu, R.-D., Tuomela, D., Joglar-Prieto, N., Palkki, R., Abánades, M., & Pejlare, J. (2023). Which one is the "best": a cross-national comparative study of students' strategy evaluation in equation solving. *International*

Journal of Science and Mathematics Education, 21(4), 1127-1151. https://doi.org/10.1007/s10763-022-10282-6

Liu, R.D., Wang, J., Star, J.R., Zhen, R., Jiang, R.H., & Fu, X.C. (2018). Turning potential flexibility into flexible performance: moderating effect of self-efficacy and use of flexible cognition. *Frontiers in Psychology, 9*, 646. https://doi.org/10.3389/fpsyg.2018.00646

Maciejewski, W. (2022). Between confidence and procedural flexibility in calculus. *International Journal of Mathematical Education in Science and Technology, 53*(7), 1733-1750. https://doi.org/10.1080/0020739X.2020.1840639

McMullen, J., Brezovszky, B., Rodríguez-Aflecht, G., Pongsakdi, N., Hannula-Sormunen, M., & Lehtinen, E. (2016). Adaptive number knowledge: exploring the foundations of adaptivity with whole-number arithmetic. *Learning and Individual Differences Volume 47,* 172-181. https://doi.org/10.1016/j.lindif.2016.02.007

NCTM. (2023). *Procedural fluency in mathematics: A position of the National Council of Teachers of Mathematics*. National Council of Teachers of Mathematics. Accessed March 31, 2024. https://www.nctm.org/Standards-and-Positions/Position-Statements/ Procedural-Fluency-in-Mathematics/

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

POPS (2014). *Perusopetuksen opetussuunnitelman perusteet 2014*. Helsinki: Opetushallitus.

LOPS (2019). *Lukion opetussuunnitelman perusteet 2019*. Helsinki: Opetushallitus.

Star, J.R., Tuomela, D., Joglar-Prieto, N., Hästö, P., Palkki, R., Abánades, M.Á., Pejlare, J., Jiang, R.H., Li, L., & Liu, R.D. (2022). Exploring students' procedural flexibility in three countries. *International Journal of STEM Education, 9*, 4. https://doi.org/10.1186/s40594-021-00322-y

Verschaffel, L. (2024). Strategy flexibility in mathematics. *ZDM Mathematics Education, 56*, 115–126. https://doi.org/10.1007/s11858-023-01491-6

Xu, L., Liu, RD., Star, J.R, Wang, J., Liu, Y., & Zhen, R. (2017). Measures of potential flexibility and practical flexibility in equation solving. *Frontiers in Psychology, 8*, 1368. https://doi.org/10.3389/fpsyg.2017.01368

APPENDIX 1. EXAMPLES OF STRATEGIES

Situational	Generic
1. $4(x+3/5)=12$	
$x+3/5=3$ $x=3-3/5$ $x=12/5$	$4x+12/5=12$ $4x=12-12/5$ $x=12/5$
2. $5(x+3/7)+3(x+3/7)=16$	
$8(x+3/7)=16$ $x+3/7=2$ $x=2-3/7$ $x=11/7$	$5(x+3/7)+3(x+3/7)=16$ $5x+15/7+3x+9/7=16$ $8x+24/7=16$ $8x=16-24/7$ $x=11/7$
3. $\frac{2x-6}{2}+\frac{6x-18}{3}=5$	
$x-3+2x-6=5$ $3x=14$ $x=14/3$	$\frac{2x-6^{(3}}{2}+\frac{6x-18^{(2}}{3}=5$ $\frac{6x-18}{6}+\frac{12x-36}{6}=5$ $6x-18+12x-36=30$ $18x=84$ $x=14/3$
$\frac{2x-6}{2}+2x-6=5$ $(\frac{1}{2}+1)(2x-6)=5$ $2x-6=5\cdot\frac{2}{3}$ $2x=10/3+6$ $x=14/3$	$3(2x-6)+2(6x-18)=30$ $6x-18+12x-36=30$ $18x=84$ $x=14/3$
4. $\frac{5x-5}{5}+\frac{6x-6}{6}=5$	
$x-1+x-1=5$ $2x=7$ $x=7/2$	$\frac{5x-5^{(6}}{5}+\frac{6x-6^{(5}}{6}=5$ $\frac{30x-30}{30}+\frac{30x-30}{30}=5$ $\frac{60x-60}{30}=5$ $60x-60=150$ $x=210/60=7/2$
$(\frac{5}{5}+\frac{6}{6})(x-1)=5$ $x-1=5/2$ $x=7/2$	$6(5x-5)+5(6x-6)=150$ $60x-60=150$ $x=210/60=7/2$

5. $2x^2 + 4x = 0$	
$x(2x+4) = 0$ $x = 0 \vee 2x + 4 = 0$ $2x = -4 \vee x = -2$ $x = 0, x = -2$	$2x^2 + 4x = 0$ $x = \frac{-4 \pm \sqrt{16 + 8 \times 0}}{4}$ $x = 0, x = -2$
If $x = 0$, then the equation holds. Otherwise we divide by x and obtain $2x + 4 = 0$ $x = -2$	$x^2 + 2x + 1 = 1$ $(x+1)^2 = 1$ $x = -1 \pm 1$ $x = 0, x = -2$
6. $(x-2)(x+3) = 0$	
$x - 2 = 0, x + 3 = 0$ $x = 2, x = -3$	$x^2 + x - 6 = 0$ $x = \frac{-1 \pm \sqrt{1+4\times 6}}{2}$ $x = 2, x = -3$
	$x^2 + x - 6 = 0$ $x^2 + x + 1/4 = 25/4$ $(x + 1/2)^2 = 25/4$ $x = -1/2 \pm 5/2$ $x = 2, x = -3$