# PROCEDURAL FLEXIBILITY IN EARLY UNIVERSITY MATHEMATICS

XXXX

YYYYYYYY

## ABSTRACT

*Procedural flexibility refers to the ability to choose efficient strategies for the solution of problems. We study procedural flexibility and accuracy when solving linear and quadratic equations as well as in systems of equations at the beginning of university mathematics. We also investigate the viability of the chosen domains for measuring flexibility in the setting of a tri-phase test. We conclude that university students show procedural flexibility and that accuracy and flexibility correlate positively with their course success. This contrasts the earlier results that only accuracy had a significant positive correlation with results of a delayed test among high school students. Finally, we show that we can measure flexibility for both quadratic equations and systems of equations.*

## INTRODUCTION

This paper focuses on mathematical flexibility, which is defined as the ability to solve mathematical tasks efficiently by taking into account task and context characteristics (Heinze et al., 2009; Rittle-Johnson & Star, 2007; Rittle-Johnson & Star, 2008). As Heinze, Star and Verschaffel (2009) state, students need to learn accurate and adaptive procedures that are reflected also in everyday, working and student life. Mathematical flexibility is an essential part of deeper understanding of strategies and their use in mathematics (Carpenter, Franke, Jacobs, Fennema & Empson, 1998). Flexibility is treated not as a trait but as a skill that can be improved by solving non-routine problems and by comparing multiple solutions to fairly routine tasks (Cigdem & Yeliz, 2015; Rittle-Johnson, Star & Durkin, 2012; Star & Rittle-Johnson, 2007; Maciejewski & Star, 2016).

Mathematical flexibility in arithmetic and elementary algebra has been studied for lower and upper secondary students (Blöte, Burg & Klein, 2001; Schneider et al., 2011; Xu et al., 2017) and in calculus for university students (Maciejewski & Star, 2016). Still, little is known about flexibility in other mathematical areas for university students and thus, this study concerns procedural flexibility at the beginning of university mathematics studies when solving algebraic tasks.

## PROCEDURAL FLEXIBILITY

There are different ways to conceptualize mathematical flexibility. Mathematical flexibility has been considered as one of five strands of mathematical knowledge alongside conceptual knowledge and procedural fluency (National Research Council, 2001). Baroody (2003) considers flexibility as "adaptive expertise" that is seen as the integration of conceptual and procedural knowledge, whereas Star (2007) has suggested that strategic flexibility constitutes deep procedural knowledge. There is a correlation between measures of conceptual and procedural knowledge and flexibility, although the direction of causation is not clear (McMullen et al., 2017, Rittle-Johnson, Star & Durkin, 2012, Schneider et al., 2011). In this study, we approach flexibility as deep procedural knowledge following the idea presented by Star (2007) and thus, we mean by flexibility such procedural, or strategic, flexibility.

The relationship between flexibility and accuracy has also been studied. Accuracy is a facet of procedural knowledge, namely the correct application of a viable solution strategy to a given task (Star, 2007). Knowledge of multiple strategies is related to greater accuracy and greater conceptual knowledge of arithmetic (Carpenter et al., 1998) and algebra (Rittle-Johnson & Star, 2007). In this study, we define accuracy as an ability to solve the task correctly, i.e. to present correct intermediate steps and to arrive at the correct final answer.

Flexibility can be studied with a tri-phase test based on tasks, which can be solved by using a well-known standard strategy as well as by using a computationally more efficient "situational strategy" (see Methods for more information on the test). In the test situation, participants are asked (1) to solve a task, (2) to provide additional solutions based on different solution strategies and (3) to mark which of their solutions they consider the best. A tri-phase test comprised of 12 linear equation tasks has been found to have good psychometric properties (Xu et al., 2017) and each task can be assessed for both flexibility and accuracy. This test was used to study flexibility of middle- and high-school students from China, Finland, Spain and Sweden (Xu et al., 2017; Liu, Wang, Star, Zhen, Jiang & Fu, 2018; Hästö, Palkki, Tuomela & Star, 2019; Star, Tuomela et al., 2022). The earlier studies have found relatively low levels of flexibility, quite high levels of accuracy and a correlation between the two (Xu et al., 2017; Lui et al., 2018; Star et al., 2022).

We study flexibility using a tri-phase test with several novelties in data gathering. First, we study first year university students and we include quadratic equations and systems of equations alongside linear equations. Second, we measure accuracy not only by the tri-phase test itself but also by a standard university exam. This allows us to address whether the correlation of flexibility and accuracy in previous studies is an artifact of the used method. Third, the tri-phase test was administered online (via Moodle) instead of face-to-face and it was shorter than in previous studies.

This study examines procedural flexibility in relation to accuracy among university students. The study has been guided by two research questions of which the first focuses on flexibility in elementary algebra. We investigate especially how viable quadratic equations and systems of equations are as domains for measuring flexibility in tri-phase tests.

- How flexible are university students in solving elementary algebra tasks?

The second question concerns the connection of flexibility and accuracy when the latter is measured in a separate exam and as part of the tri-phase test.

- How strong is the connection between procedural flexibility and accuracy?

## METHODS

### Participants

We invited students in a large introductory mathematics course at a Finnish university to participate in the study. Data was collected with a test administered as a voluntary Moodle examination in the beginning of the course in September 2021. The test also served as a check of prerequisite knowledge and thus, some students completed it without consenting to the use of their results. Consent to use course exam scores was asked separately. Results of 4 participants could not be used in the analysis due to technical difficulties in entering or extracting data from Moodle.

The course served both mathematics majors and students from other programs. There were 121 participants altogether of whom 98 had studied advanced mathematics in high school or had completed at least 10 credits of university mathematics. 43 participants majored in computer science, 29 in mathematics or statistics, 17 in economics and the rest of the participants represented a wide range of different subjects.

### Data gathering

The test consisted of two parts, a preliminary part and a tri-phase part for examining procedural flexibility. The former included four tasks on arithmetic and algebraic simplification and equation solving (see Table 1). Task 4 was a multiple-choice task whereas only the final answer was entered and graded in tasks 1-3.

Table 1. Preliminary part of the test

| | Task | Answer | % of correct answers |
|---|---|---|---|
| **1.** | Simplify $\frac{2}{5} - \frac{1}{2}$ | $-\frac{1}{10}$ | 92% |
| **2.** | Simplify $\frac{2/3}{1/5}$ | $\frac{10}{3}$ | 91% |
| **3.** | Simplify $(2x)^5(-4x)^{-1}$ | $-8x^4$ | 53% |
| **4.** | Which of the following criteria applies to the function $2(x-1) = 3x - (x+1)$?<br>a. The equation has exactly one solution<br>b. The equation has no solutions<br>c. Every real number satisfies the equation<br>d. None of the above | b. | 88% |

The tri-phase part followed the structure laid out by Xu et al. (2017). In Phase 1, participants were instructed to solve four task and write the solutions in the leftmost column of a 4 × 3 grid on a sheet of paper. In Phase 2, new instructions appeared and asked students to provide up to two additional solutions to each task in the remaining columns. Finally, in Phase 3, participants were asked to circle for each task the solution they felt was best. The answers were returned by taking a photograph of the paper and uploading it to Moodle. In contrast to previous tri-phase tests, we did not have a time limit for each phase.

Tasks 5 and 6 were linear equations from Star et al. (2022) whereas tasks 7 and 8 were designed for this study and concerned quadratic equations and systems of equations. The tasks are presented in the first column of Table 2.

**Data analysis**

In the tri-phase part of the test, every task was coded separately for correctness and for strategy. In addition, the solution marked as "best" was recorded. The coding was carried out jointly by two of the authors.

The accuracy score of the tasks 1-4 simply means that the answer was correct as no intermediate steps were provided. To be considered correct in tasks 5–8, both the final answer and all intermediate steps had to be mathematically valid. For these tasks an accuracy score of 1 was obtained if the solution to the task during phase 1 was correct. Thus, the accuracy score (0–4) of the tri-phase part was determined by the number of correct solutions in Phase 1. Correctness in phase 2 was not considered, as per the coding scheme of the tri-phase test established by Xu et al. (2017).

The solution in the tri-phase part was coded as standard strategy (S), situationally appropriate (situational, M), or other (O). The standard and situational strategies were defined in the four tasks as follows (see Table 2 for an overview).

Table 2. Tri-phase part of the test

| Task | Standard solution | Innovative solution | Other solution (example) |
|---|---|---|---|
| 5. $4(x+\frac{3}{5})=12$ | $4x+\frac{12}{5}=12$<br>$4x=12-\frac{12}{5}$<br>$4x=\frac{48}{5}$<br>$x=\frac{12}{5}$ | $x+\frac{3}{5}=3$<br>$x=3-\frac{3}{5}$<br>$x=\frac{12}{5}$ | $\frac{4(x+\frac{3}{5})}{12}=1$<br>$\frac{x+\frac{3}{5}}{3}=1$<br>$x+\frac{3}{5}=3$<br>$x=\frac{12}{5}$ |
| 6. $5(x+\frac{3}{7})+3(x+\frac{3}{7})=16$ | $5x+\frac{15}{7}+3x+\frac{9}{7}=16$<br>$8x+\frac{24}{7}=16$<br>$8x=16-\frac{24}{7}$<br>$8x=\frac{88}{7}$<br>$x=\frac{11}{7}$ | $8x+\frac{24}{7}=16$<br>$8x=16-\frac{24}{7}$<br>$8x=\frac{88}{7}$<br>$x=\frac{11}{7}$ | $5+3=\frac{16}{x+\frac{3}{7}}$<br>$\frac{8}{16}=\frac{1}{x+\frac{3}{7}}$<br>$\frac{1}{2}=\frac{1}{x+\frac{3}{7}}$<br>$x=2-\frac{3}{7}$<br>$x=\frac{11}{7}$ |
| 7. $2x^2+4x=0$ | $x=\frac{-4\pm\sqrt{4^2-4\cdot2\cdot0}}{2\cdot2}$<br>$x=\frac{-4\pm4}{4}$<br>$x_1=0,\ x_2=-2$ | $x(2x+4)=0$<br>$x_1=0,\ 2x+4=0$<br>$2x=-4,\ x_2=-2$ | $x^2+2x+1=1$<br>$(x+1)^2=1$<br>$(x+1)^2-1=0$<br>$x_1=0,\ x_2=-2$ |
| 8. $\begin{cases}3x+4y=2\\y=2x+7\end{cases}$ | $\begin{cases}3x+4y=2 & \|\|\cdot 2\\y=2x+7 & \|\|\cdot 3\end{cases}$<br>$\begin{cases}6x+8y=4\\-6x+3y=21\end{cases}$<br>$11y=25,\ y=\frac{25}{11}$<br>$\frac{25}{11}=2x+7$<br>$x=-\frac{26}{11}$ | $3x+4(2x+7)=2$<br>$3x+8x+28=2$<br>$11x=-26,\ x=-\frac{26}{11}$<br>$y=2\cdot(-\frac{26}{11})+7=\frac{25}{11}$ | Graphical solution |

In task 5, if the first step of the solution of the equation $4\left(x+\frac{3}{5}\right)=12$ was to distribute the 4 on the left side, then it was a standard strategy. If the first step was to divide the equation by 4, then it was a situational strategy.

In task 6, if the first step of the solution of $5\left(x+\frac{3}{7}\right)+3\left(x+\frac{3}{7}\right)=16$ was to distribute by multiplying with 5 and 3, then it was a standard strategy. If the first step was to combine the terms on the left side, resulting in $8\left(x+\frac{3}{7}\right)=16$, then it was a situational strategy.

In task 7, if the first step of the solution of $2x^2+4x=0$ was to use the quadratic formula, then it was a standard strategy. If the first step was to factorize as $x(2x+4)=0$ and to use zero-product property, then it was a situational strategy and marked as M. As a variant of this (coded M2), some students divided the equation with $x$ or $2x$ and concluded that either $x=0$ or $x+2=0$.

Task 8 required the most involved coding. In the situational strategy (M), the student substituted $2x+7$ in place of $y$ in the first equation. Another variant considered a situational strategy (marked M2) was to multiply the second equation by 4 and subtract it from the first one. The coding S was used for solutions in which the equations were multiplied with non-optimal real numbers and then added; S2 referred to solutions with matrices, and S3 to solutions where $x$ was solved from one equation and substituted or $y$ was solved in a nonoptimal way (e.g. from the first equation and substituted in the second one).

In order to use the same measures of flexibility and accuracy as Star et al. (2022), we first equated different variants of the standard strategy (S, S2 and S3) and of the situational strategy (M and M2). A task was deemed flexibly solved if three criteria were satisfied:

1. The task was solved using a standard strategy.
2. The task was solved using a situational strategy.
3. The strategy chosen as best was a situational one.

The number of flexibly solved tasks gave each participant a flexibility score between 0 and 4.

Over the course of our analysis, we observed that these criteria led to unreasonable results when only two situational strategies (M and M2) were presented. Therefore, we introduced an alternative operationalization for flexibility (see Table 3). For simplicity, this is called flexibility, and the operationalization from the previous paragraph is called "flexibility (old)". Note that they coincide for linear equations, since there is only one type of situational strategy in this context.

**RESULTS**

We obtained data for examining the procedural flexibility of the 102 students who completed the entire test. There were 19 students out of 121, who did not complete the tri-phase part. Comparing the score on the preliminary part (tasks 1–4) indicates that these 19 students had marginally lower average scores ($p=0.052$ according to a t-test). Thus, the students who chose to participate in the entire test seem to be somewhat self-selected. The 19 students are not included in the following analyses.

As can be seen in Table 3, all tasks were completed with rather high accuracy. The lowest accuracy scores were in Task 3 (56 %), Task 7 (82 %) and Task 8 (67 %) but the accuracy in the remaining tasks was 90–95 %. The most important reasons for the low accuracy scores were small calculation mistakes. In task 7 that was on second-degree equation, many students forgot to show the second solution. In particular, in approach M2 they divided the equation by $x$ and forgot to take into account the case $x=0$. In task 8 that was on system of equations, a typical mistake was to solve only one variable, for instance, solving $x$ correctly but not solving $y$.

The average number of solutions was just under 2 for each of the tri-phase tasks. For tasks 5 and 6, the standard strategy was very well-known (97–98 %) and the situational strategy was quite well-known, whereas the situation for tasks 7 and 8 was the opposite. Only 47 % presented the standard strategy while 94 % used the situational in Task 8.

Table 3. Overview of test results

| | Abbreviation | Task 1 | Task 2 | Task 3 | Task 4 |
|---|---|---|---|---|---|
| Accuracy | preAcc | 91 % | 92 % | 56 % | 90 % |
| | | Task 5 | Task 6 | Task 7 | Task 8 |
| Accuracy | Acc | 95 % | 93 % | 82 % | 65 % |
| #Solutions | | 1.98 | 1.84 | 1.91 | 1.83 |
| Standard | | 98 % | 97 % | 64 % | 47 % |
| Situational | | 80 % | 73 % | 81 % | 94 % |
| Flexibility (old) | | 71 % | 62 % | 39 % | 35 % |
| Flexibility | Flex | 71 % | 62 % | 43 % | 54 % |

Turning next to flexibility (old) as defined in the section Data analysis, we see in Table 3 higher scores in Tasks 5 and 6 (60–70 %) than in Tasks 7 and 8 (35–40%). We found that many participants presented the two variants of situationally appropriate solutions (coding M and M2) in Task 8 and, to a lesser extent, in Task 7. This is natural, since there is not so clear standard solution than in Task 8. We therefore considered a broader view on flexibility, with answers containing the two situational strategies with a choice of either as best also counted as flexible. These variables are shown on the last two rows of Table 3. As can be seen, the proportion of flexible participants rises from 1/3 to ½ in Task 8 with new flexibility. The remaining analyses use the new flexibility variable.

To investigate whether the four tasks measured the same flexibility we calculated the correlations between the flexibility scores in tasks 5–8. Since the scores are dichotomous (i.e. 0 or 1), tetrachoric correlation is an appropriate measure. As can be seen in Table 4, the pairwise correlations are very high, although the correlations with Task 7 are somewhat lower. The Chronbach alpha of the scale consisting of the four tasks' flexibility equals 0.782. This falls between Nunnally's (1978) recommended thresholds of 0.7 for early stage research and 0.8 for applied research.

Table 4. Tetrachoric correlation of flexibility in Tasks 5–8

| | T5 | T6 | T7 | T8 |
|---|---|---|---|---|
| T5 | 1.000 | 0.883 | 0.761 | 0.807 |
| T6 | 0.883 | 1.000 | 0.633 | 0.753 |
| T7 | 0.761 | 0.633 | 1.000 | 0.504 |
| T8 | 0.807 | 0.753 | 0.504 | 1.000 |

The students in the test represented a variety of majors. The largest groups were computer science (36), mathematics and statistics (26) and economics (15), with the remaining students scattered over a dozen fields each with fewer than 5 students in the sample. The flexibility score in the three large groups was 1.78, 2.77 and 2.47, respectively. Pairwise t-test (without correction) indicates that mathematics students were statistically significantly more flexible than computer science students $(p=0.008)$whereas the other comparisons were not significant. However, larger and more representative groups would be needed for more convincing results.

We also examined accuracy by measuring it through the course exam which took place some weeks after the test. Of the 102 students who completed the whole test, 57 allowed the use of their exam scores for the research whereas 45 did not. The former subgroup had higher Acc $(p=0.036)$ and Flex $(p=0.044)$ and no significant differences in preAcc $(p=0.11)$. Thus it seems that again the 57 students who chose to permit the use of the grade are not wholly representative of all students taking the test. Figure 1 indicates that most students for each flexibility score cluster at the high end of the exam score (15–18 points). The 8 students with zero points did not participate in the exam and are excluded from further analysis. Furthermore, the three isolated points (in red) had an undue influence on results such as group variance and so they were excluded from further analyses as outliers.

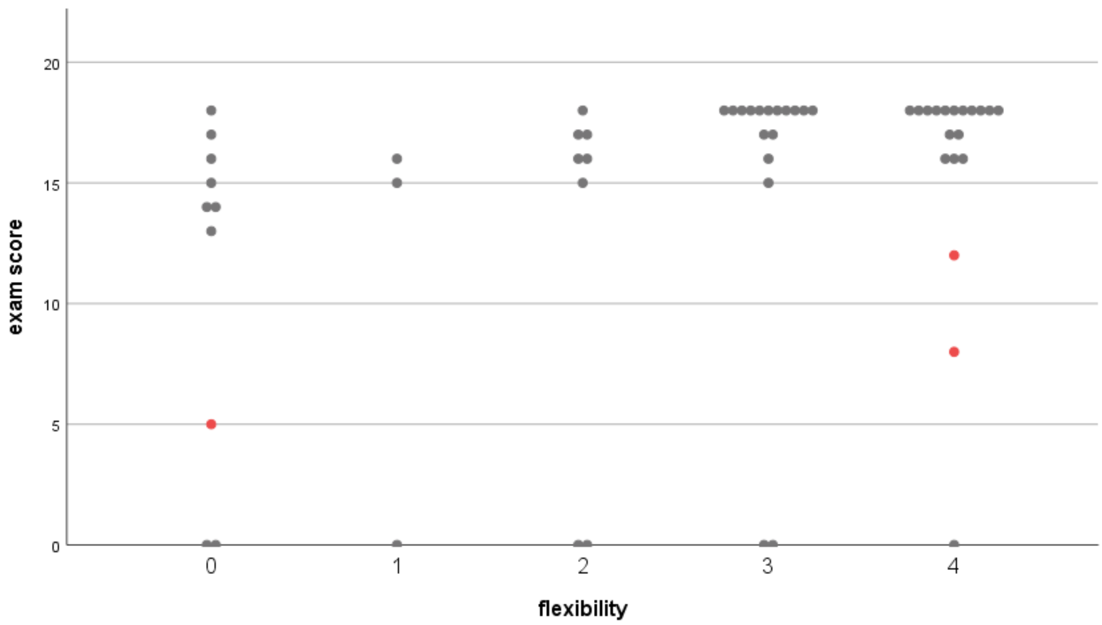


Figure 1. Exam score by flexibility

Table 5 shows that there is a slight trend towards higher exam scores with increasing flexibility scores; an ANOVA test confirms that the groups are statistically different from one another $(F=7.553, p=0.001)$. Corresponding analyses of the exam score's dependence on accuracy in the preliminary part (preAcc) or the tri-phase part (Acc) showed similar although not quite as clear trends. This is also indicated by the correlations shown in Table 6. The correlations between preAcc, Acc and Flex are mostly not significant in this group ($n=46$) but the correlations in the whole dataset $(n=102)$ lie in the .0285–0.390 range and are significant at the 0.01-level.

Table 5. Mean exam score by flexibility

| Flex | Score | N | Std. Dev. |
|---|---|---|---|
| 0 | 15.3 | 7 | 1.8 |
| 1 | 15.5 | 2 | 0.7 |
| 2 | 16.5 | 6 | 1.0 |
| 3 | 17.5 | 15 | 0.9 |
| 4 | 17.5 | 16 | 0.8 |
| Total | 17.0 | 46 | 1.3 |

Table 6. Pearson correlations between variables ). Asterisks indicate significance at the 0.05-level (*) and the 0.01-level (**).

| | exam | preAcc | Acc | Flex |
|---|---|---|---|---|
| exam | 1 | 0.386** | 0.416** | 0.616** |
| preAcc | 0.386** | 1 | 0.312* | 0.150 |
| Acc | 0.416** | 0.312* | 1 | 0.282 |
| Flex | 0.616** | 0.150 | 0.282 | 1 |

A linear regression model predicting exam score based on flexibility, pre-accuracy and accuracy was significant $(F=13.9, p<0.001)$, with an adjusted $R^2$-value of 0.463. Flexibility and accuracy in the preliminary part were statistically significant, accounting for 28 % and 6 % of the variance, respectively. The scatter plot in Figure 1 suggests that there was a ceiling effect with the most common score being the maximum of 18 when the flexibility score was 3 or 4. A stronger correlation might result if the exam had more power to distinguish high-achieving students.

## DISCUSSION

We have studied students' procedural flexibility in elementary algebra at the beginning of their university studies. In regard to the first research question, we found high levels of accuracy (65–95 %) and quite high levels of flexibility (43–71 %). As expected, these are higher than for Finnish 11th grade high-school students in the advanced mathematics track, who had 66 % accuracy and 36 % flexibility when averaged over all 12 tasks in the longer version of the tri-phase test (Star et al., 2022). It is good to notice that only two tasks of the original test, namely tasks 5 and 6, were used in our study and neither study is based on representative samples. Still, the similar but somewhat higher numbers can be considered a validating factor for the new, shorter test presented in this article.

Our tri-phase test included two new tasks designed for this study in addition to the two tasks from previous studies (e.g., Liu et al., 2019; Star et al., 2022). We derived measures of accuracy and flexibility following the model of previous research using the tri-phase test (Star et al., 2022). This means regarding flexibility that the participant should produce both a standard and a situationally appropriate solution, and mark the latter as "best". However, we found that for task 8 (system of equations), this requirement was not as clear due to the lack of an obvious standard algorithm to be used in the task. Students are typically taught to solve systems of equations by three different means (addition method, substitution method and graphical methods) without prioritizing any of them as a standard method. Furthermore, each method can be applied in various ways, e.g. by choosing which variable to solve first. Many participants produced two distinct solutions which we both considered situationally appropriate. To account for this, we considered a new, broader flexibility variable which also included the case of two situational solutions with either chosen as "best". The same applied also to task 7 (quadratic equation), albeit to a much lesser extent. This demonstrates the difficulty of applying the operationalization of procedural flexibility from outside the domain of linear equations, which have a very clear standard algorithm (cf. Xu et al., 2017; Hästö et al., 2019; Liu et al., 2019; Star et al., 2022).

Based on our study, we conclude that quadratic equations and systems of equations seem to serve viable domains for measuring flexibility with a tri-phase setup but more work is needed for producing a better, more reliable measure. The new flexibility measure performed well in many respects in this study. The correlations between the flexibility in different tasks were high and the reliability of the flexibility scale as measured with Chronbach's alpha was good for early stage research. There is still a need in future investigations to improve on this, for instance by refining the task instructions or by adding more tasks.

In the tri-phase test, one measures both flexibility and accuracy with the same tasks. These two measures correlate quite highly in most groups that have been studied (Star et al., 2022). With the new setup, we compared flexibility with accuracy not only based on the tri-phase test data but also drawn from the course exam taken several weeks after the tri-phase test. The correlation between accuracy and flexibility at 0.390 was lower than for Finnish students in middle and high-school (Star et al., 2022). This may be due to a ceiling effect in the accuracy variable of the tri-phase test. Interestingly, the correlation with the exam score was higher, especially for flexibility. Indeed, we found that flexibility in the tri-phase test was a much better predictor of exam score than accuracy, accounting for 29 % of the variance, while accuracy was not statistically significant. This contrasts with the results of Hästö et al. (2019) that accuracy in the tri-phase test was a better predictor than flexibility of exam results in the national matriculation examination one year after the test. This may reflect the higher correlation between accuracy and flexibility in high-school students and the ceiling effect, or it may mean that the course exam measures mathematical skills where flexibility is of greater value when compared to the national exam at the end of high-school. In response to the second research question, we found that accuracy and flexibility were correlated, but not as strongly as in previous studies. Moreover, we can say quite confidently that flexibility's connection with accuracy in the tri-phase test differed from its connection with accuracy in the exam. However, contrary to expectation the latter was stronger rather than weaker. This, again, may be due to a ceiling effect in the test's accuracy measure.

## LIMITATIONS AND FUTURE STUDIES

Our results are based on a convenient student sample from an introductory course in one university. This limits the generalizability to other populations. We had 121 participants, but many did not return a full set of data. After removing participants with incomplete data, those who did not consent to the use of their exam score, and a few outliers, only 46 participants remained for some analyses. Furthermore, each stage of self-selection may result in a group which is slightly more flexible and mathematically skilled. Thus the results may give too rosy a picture of the flexibility of university students. This self-selection may also be responsible in part for the ceiling effect that we observed both in the accuracy measurement in the test as well as in the exam. It is unclear how this affects the correlation between flexibility and accuracy, as it might either increase or decrease it.

Administering the tri-phase test through Moodle platform proved largely successful and it seems to be possible to continue developing the tri-phase test in this context. Improving the formulation of the questions as well as motivating the students to answer all questions could increase the completion rate. Furthermore, the new tasks on the quadratic equation and the system of equations can be further developed to fit better with the tri-phase framework and thus the reliability of the test. The resulting Moodle-based tri-phase test can then be scaled up to a larger study with a more representative sample.

## REFERENCES

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody, & A. Dowker (Eds.), *The Development of Arithmetic Concepts and Skills: Constructive Adaptive Expertise* (pp. 1–33). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410607218

Blöte, A., Burg, E., & Klein, A. (2001). Students' flexibility in solving two-digit addition and subtraction problems: Instruction effects. *Journal of Educational Psychology*, 93, 627–638. https://doi.org/10.1037/0022-0663.93.3.627

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29*(1), 3–20. https://doi.org/10.2307/749715

Cigdem, A., & Yeliz, Y. (2015). Common and Flexible Use of Mathematical Non Routine Problem Solving Strategies. *American Journal of Educational Research,* 3, 1519–1523. https://doi.org/10.12691/education-3-12-6

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

Hästö, P., Palkki, R., Tuomela, D., & Star, J. R. (2019). Relationship between mathematical flexibility and success in national examinations. *European Journal of Science and Mathematics Education, 7*(1), 1–13. https://doi.org/10.30935/scimath/9530

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., & Star, J.R. (2016). Developing flexible procedural knowledge in undergraduate calculus. *Research in Mathematics Education,* 18, 1–18. https://doi.org/10.1080/14794802.2016.1148626

National Research Council. (2001). *Adding It Up: Helping Children Learn Mathematics.* Washington, DC: The National Academies Press. https://doi.org/10.17226/9822.

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

Rittle-Johnson, B., & Star, J.R. (2007). Does Comparing Solution Methods Facilitate Conceptual and Procedural Knowledge? An Experimental Study on Learning to Solve Equations. *Journal of Educational Psychology,* 99(3), 561–574. https://doi.org/10.1037/0022-0663.99.3.561

Rittle-Johnson, B., Star J.R., & Durkin, K. (2012) Developing procedural flexibility: are novices prepared to learn from comparing procedures? *Journal of Educational Psychology,* 82(3), 436–55. https://doi.org/10.1111/j.2044-8279.2011.02037.x

Schneider, M., Rittle-Johnson, B., & Star J.R. (2011). Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology Journal*, 47(6), 1525–1538. https://doi.org/10.1037/a0024997

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

Rittle-Johnson B. and Star, J.R, (2008) Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, 18(6), 565–579. https://doi.org/10.1016/j.learninstruc.2007.09.018

Xu, L., Liu, RD., Star, J.R, Wang, J., Liu, Y. and 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