European Journal of Science and Mathematics Education
Vol. 7, No. 1, 2019 , 1-13

Relationship between mathematical flexibility and success in national examinations

Peter Hästö[1,2,*], Riikka Palkki[2], Dimitri Tuomela[2], Jon R. Star[3]
[1]University of Turku, Department of Mathematics and Statistics
[2]University of Oulu, Department of Mathematical Sciences
[4]Harvard University, Faculty of Education
For correspondence: peter.hasto@oulu.fi

Abstract
Flexibility is an important element in learning mathematics. The purpose of this study was to investigate whether flexibility in linear equation solving predicts future academic achievement in mathematics and other subjects. Participants were 149 Finnish high-school students. Results show that flexibility was related to grades in both tracks of mathematics, chemistry, and mother tongue, as well as the total number of exams taken in the national matriculation examination. However, when controlling for accuracy in equation solving, only basic level mathematics and, to some degree, chemistry grade were related to flexibility. On the other hand, flexibility had an impact on students' choice to participate in the mathematics and physics exams. A theoretical analysis shows that student selection may mask part of the relationship between flexibility and grade.

Keywords: flexibility; equation solving; national matriculation examination; student choice

Introduction & theory

While there has been a great deal of research about mathematical flexibility (see, e.g., Baroody, 2003; Heinze, Star,& Verschaffel, 2009; Star & Rittle-Johnson, 2008; Verschaffel, Luwel, Torbeyns, & Van Dooren, 2009), less attention has been paid to the effects of mathematical flexibility on achievement in other school subjects and mathematics more generally. Of particular interest is the relationship between flexibility and students' future problem solving. In this study, we address this issue. Concretely, we compare the results of high school students' flexibility in a linear equation solving test taken in grade 11 with their results from the national matriculation examination 1–2 years later.

Flexibility in mathematical problem solving has been defined as the knowledge of multiple strategies and the ability to choose the most mathematically appropriate strategy for a given task (Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2008). Here, mathematically appropriate refers to the strategy or strategies that are optimally matched to the characteristics or features of a task (and thus produce a more efficient, elegant, or effective solution). Flexibility is related to, but distinct from, conceptual and procedural knowledge (Schneider, Rittle-Johnson, & Star, 2011). Numerous research reports and policy documents across several countries suggest that flexibility should be an instructional goal at all levels of mathematics instruction (e.g., National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; Opetushallitus, 2015).

When students exhibit flexibility in one mathematical domain, do they subsequently perform better in other mathematical domains? Recent research has not explored this issue. Flexibility has been generally found to influence procedural and conceptual knowledge; Rittle-Johnson, Star & Durkin (2012, p.

446) found that pre-test flexibility had a high correlation to procedural and conceptual knowledge in a post-test and a moderate correlation with the same variables in a delayed post-test. Relatedly, Schneider *et al.* (2011) found that procedural and conceptual knowledge of equations influenced future procedural flexibility, as well as procedural and conceptual knowledge, in the same area. A study by McMullen *et al.* (2017) explored a similar relationship, finding that adaptive number knowledge (a facet of flexibility) was related to future pre-algebra skills even while controlling for prior arithmetic fluency, grade level and conceptual knowledge.

Based on prior research on flexibility (e.g., Star & Rittle-Johnson, 2008; Xu, Liu, Star, Wang, Liu & Zhen, 2017), we distinguish between students' *knowledge* of strategies and the ability to implement or *use* strategies on given problems. Drawing upon this distinction, we use the following definitions related to flexibility.

A **flexibility-eligible task** is a task that can be solved in several different ways, including both by standard and non-standard strategies. On a flexibility-eligible task, a student who satisfies all of the following criteria is said to be **flexible**:

(a) Knows at least one standard strategy.
(b) Knows at least one mathematically appropriate non-standard strategy.
(c) Can determine the strategy that is more mathematically appropriate for the task.

In addition, a flexible student is **spontaneously flexibility** if he/she:

(d) Uses unprompted the more mathematically appropriate strategy when solving the task.

In this study, we explore the relation between students' flexibility and their future exam achievement. Our research question asks: Do Finnish students who demonstrate flexibility in linear equation solving subsequently exhibit greater proficiency in the national matriculation exam in mathematics and other subjects?

Method

Participants and data sets.

In the spring of 2016, we collected data on flexibility in linear equations solving from students across multiple grade levels and tracks (described in more depth below). The flexibility data is based on a convenience sample. Participating schools were diverse in size and geographic location (both rural and urban).

The present study draws from a subset of this larger dataset, using data of those 11th grade students ($n = 164$) who participated in at least one matriculation examination between the spring of 2016 and the spring of 2018 ($n = 149$).

The final exam in Finnish high schools is called the matriculation examination (described below). Data on all participants in target schools in the matriculation examinations as of the spring 2018 were obtained from the Matriculation Examination Board. It was combined with the flexibility data by comparing participant names and schools, which uniquely identified the individuals in the study. The total number of tests taken by the students in our sample is shown in Table 1.

Table 1. Number of exam results obtained from different exam times.

Time	Spring 2016	Fall 2016	Spring 2017	Fall 2017	Spring 2018
# of exams	2	179	505	76	36

Flexibility test

The 45-minute flexibility test measured students' ability to produce standard and innovative strategies in linear equation solving. The test contained 12 flexibility-eligible tasks and three phases. The equa-

tions are listed in Table 2. In the first phase (15 min), students were prompted to provide one solution for each equation. In the second phase (20 min), they were asked to write as many solutions as possible for each equation. In the third phase (5 min), students were asked to circle the solution they considered "the best" for each of the 12 problems. The test was previously used and validated in Xu, *et al*. (2017).

Table 2. The equations in the test.

1) $\mathbf{4}(\boldsymbol{x}-\mathbf{2})=\mathbf{24}$	2) $\mathbf{3}(\boldsymbol{x}+\mathbf{0.69})=\mathbf{15}$	3) $\mathbf{4}(\boldsymbol{x}+\frac{\mathbf{3}}{\mathbf{5}})=\mathbf{12}$
4) $\mathbf{4}(\boldsymbol{x}+\mathbf{6})+\mathbf{3}(\boldsymbol{x}+\mathbf{6})=\mathbf{21}$	5) $5(x+\frac{3}{7})+3(x+\frac{3}{7})=16$	6) $2(x-0.31)+3(x-0.31)=15$
7) $\mathbf{8}(\boldsymbol{x}-\mathbf{5})=\mathbf{3}(\boldsymbol{x}-\mathbf{5})+\mathbf{20}$	8) $8(x-\frac{2}{5})-11=6(x-\frac{2}{5})$	9) $5(x+0.6)+3x=5(x+0.6)+7$
10) $\frac{\mathbf{2x-6}}{\mathbf{2}}+\frac{\mathbf{6x-18}}{\mathbf{3}}=\mathbf{5}$	11) $\frac{x+3}{3}+\frac{3x-9}{9}=1$	12) $\frac{5x+5}{5}+\frac{6x+6}{6}=6$

For the problems on the flexibility test, the standard solution typically included four steps: distribute-parentheses; combine-like-terms; move x-terms to the left and non-x-terms to the right; divide-by-coefficient. This method is typically taught to Finnish students. The innovative solution was more efficient as measures by the number of steps and the technical difficulty of carrying out the steps.For instance, for the equation $4(x+6)+3(x+6)=21$, the innovative steps were:
$7(x+6)=21 \Leftrightarrow x+6=3 \Leftrightarrow x=-3$.
The solutions were coded for type (standard, innovative or other) and correctness.

In accordance with the definitions presented above, each student's performance in the flexibility test was coded by variables flexibility (*F*) and spontaneous flexibility (*SF*). Furthermore, each task was scored for **accuracy** (*acc*) based on the correctness of the solution provided in the first phase. Thus, each of these variables takes values in the range 0–12, 12 being the highest. See Xu, *et al*. (2017) for further details.

The Finnish national matriculation examination

In Finland, about half of every age group attends high school (grades 10–12). In high school, students choose between two tracks in mathematics: advanced and basic. Towards the end of high school, students take part in the so-called matriculation examination consisting of exams in different subjects of their choice.

The Finnish matriculation examination offers students many choices of which subjects to participate in. The only compulsory subject is mother tongue (Finnish). About a third of the students participate in the advanced mathematics exam, another third of the students participate in the basic mathematics exam, and a third do not take any mathematics exam. Grades are given using a weighted normal distribution (see Table 1) in order to make results from different exams in the same subject comparable. The exam can be taken twice a year (spring and fall), and a student may participate in exams on three consecutive instances (including retaking exams) for their diploma. Note that the matriculation examination is not a university entrance exam but a high-school leaving exam, the name being a historical legacy.

Table 3. Distribution of grades in advanced mathematics in our sample and in Finland.

Grade	Sample ($n = 85$)	Finland
laudatur (7) – *highest*	5 %	7 %
eximia cum laude approbatur (6)	31 %	22 %
magna cum laude approbatur (5)	36 %	27 %

cum laude approbatur (4)	14 %	22 %
lubenter approbatur (3)	9 %	16 %
approbatur (2) – *lowest*	4 %	4 %
improbatur (0) – *fail*	1 %	2 %

The mathematics exam consists of 13 tasks of which the student choses 10 (with several restrictions). The tasks mostly require not just an answer but also a complete argument for scoring points. For instance, in the spring 2017, Question 3 asked students to find the value of t which minimizes the length of the vector $\overline{c_t} = t\,\bar{a} + (1-t)\,\bar{b}$, where $\bar{a} = \bar{\imath} + 2\bar{\jmath} + 3\bar{k}$ and $\bar{b} = 2\bar{\imath} + 5\bar{k}$. Question 9 asked students to show that $(p+q)(q+r)(r+p)$ is even for all positive integers p, q and r. The exam also contained some more non-standard tasks (e.g. finding errors in presented solutions).

In this study, we treat the grades of the various matriculation exams as indicators of the **proficiency** of students in that subject.

The students in this sample performed slightly above the national average in the matriculation examination (see Table 1 for the data on advanced mathematics students).

Analysis

We divided students into four groups based on flexibility *F*: no flexibility (*F* score 0), low flexibility (score 1-3), medium flexibility (score 4-6) and high flexibility (score 7-12). These cut-offs were chosen partly to ensure reasonable size groups, and partly because the flexibility test was based on blocks of three similarly structured tasks. For spontaneous flexibility (*SF)* we used the groups corresponding to no SF (score 0), low SF (1-3), and medium SF (4-10), because of the low number of students with spontaneous flexibility.

We first analyzed the relationship between (spontaneous) flexibility groups and accuracy in the flexibility test in our sample. For this, we use cross-tabulations (Kerlinger & Lee, 2000, pp. 224–227) and Pearson correlation coefficients.

Second, we consider average grades in different subjects of students in the different flexibility groups. We use ANOVA analysis (Kerlinger & Lee, 2000, pp. 307–433) to determine whether the groups' averages differ statistically significantly from one another. We also use ANCOVA analysis (Maxwell & Delaney, 2004, pp. 399–468; see also Wright, 2006) to discount the effect of accuracy on grades.

In the third subsection, we tabulate how many students in each group participated in each of the exams. Aχ^2(chi-squared) analysis (Kerlinger & Lee, 2000, pp. 229–236) is used to determine whether participation in an exam differed statistically significantly between the flexibility groups. Participation is coded by a dummy variable, which takes value 1 if the student participated in a given exam and 0 otherwise. The dummy variables are used in a binary logistic regression analysis (Kerlinger & Lee, 2000, pp. 808–811) to determine whether participation in an exam differed statistically significantly between the flexibility groups when controlling for accuracy.

Finally, we study how big an effect selective participation could have on group averages based on a theoretical model presented in Appendix 1.

In all tables we have indicated by increasingly dark shades of gray results approaching statistical significance ($0.05 < p < 0.10$), marginally statistically significant ($0.01 < p < 0.05$) and statistically significant ($p < 0.01$).

Results

Flexibility and accuracy in the sample

Cross tabulations of accuracy (*acc*) and flexibility group (*gF*) are shown in Table 4. On the far right, we see that the number of students showing no flexibility was 54, while the low, med and high flexibility groups contained 37, 31 and 27 students, respectively. Furthermore, we observe a "triangle" type relation between accuracy and flexibility, viz. there are no students with high flexibility and low accuracy, whereas there are some students with low flexibility but high accuracy.

Note as well that the average accuracies in the flexibility groups were 4.54 (no flexibility), 6.65 (low), 8.13 (med) and 9.44 (high). The correlation between the *acc* and *F* variables was 0.607 ($p < 0.0005$).

Table 4. Cross tabulations of *acc* versus *gF*.

	acc													
gF	0	1	2	3	4	5	6	7	8	9	10	11	12	Tot
No	3	5	2	11	4	11	5	7	4		1	1		54
Low		1	1	4	4	3	4	5	6	1	5	2	1	37
Medium					1	5	3	6	4	3	1	2	6	31
High						1	1	1	4	7	5	4	4	27
Total	3	6	3	15	9	20	13	19	18	11	12	9	11	149

Cross tabulations of accuracy (*acc*) and spontaneous flexibility group (*gSF*) are shown in Table 5. On the far right, we see that the number of students showing no spontaneous flexibility is 110, while the low and med spontaneous flexibility groups contained 21 and 18 students, respectively. Furthermore, we observe an even stronger "triangle" type relation between accuracy and spontaneous flexibility, viz. there are no students with high spontaneous flexibility and low accuracy, whereas there are some students with low spontaneous flexibility but high accuracy.

The average accuracies in the spontaneous flexibility groups were 5.76 (no spontaneous flexibility), 8.62 (low) and 10.17 (med). The correlation between the *acc* and *SF* variables was 0.470 ($p < 0.0005$).

Table 5. Cross tabulations of *acc* versus *gSF*.

	acc													
gSF	0	1	2	3	4	5	6	7	8	9	10	11	12	Tot
No	3	6	3	15	8	19	13	14	13	3	5	3	5	110
Low					1	1		3	5	3	5	2	1	21
Medium								2		5	2	4	5	18
Total	3	6	3	15	9	20	13	19	18	11	12	9	11	149

The fact that only 26 % of students showed spontaneous flexibility in the test suggests that the spontaneous flexibility groups will not be very reliable when dealing with subsets of students taking a particular test. For instance, of students taking the basic mathematics test, 50 showed no spontaneous flexibility, only 1 showed low *SF*, and none showed medium *SF*. Consequently, results are presented below only for flexibility, not for spontaneous flexibility.

A comparison of Tables 4 and 5 shows that 56 students (38 %) exhibited flexibility but no spontaneous flexibility. Thus, a quite large proportion of participants are capable of flexibility but not inclined to use this skill spontaneously.

Relationship between flexibility and matriculation exam grades

Next, we consider the relationship between flexibility and matriculation examination results. The following table shows the average grade in different subjects for the different flexibility groups along with some statistical measures. The number of students in each group is shown in Table 7, below.

Table 6. Grade averages in advanced mathematics, basic mathematics, physics, chemistry, mother tongue (Finnish) and average of all other exams by flexibility group (no/low/med/high). The last row is the average total number of exams taken. The column Levene shows the *p*-value of Levene's test for equality of variances. The column ANOVA shows the *p*-value of an ANOVA test that all flexibility groups have statistically identical average scores. The column ANCOVA shows the *p*-value of the flexibility group effect on the grade average in an ANCOVA test with *acc* as covariate.

	Flexibility group						
Exam	No	Low	Med	High	Levene	ANOVA	ANCOVA
Adv math	4.69	4.62	4.96	5.56	0.009	0.041	0.538
Basic math	4.00	5.80	5.67		0.413	0.000	0.002
Physics	4.22	4.23	5.00	5.25	0.279	0.150	0.244
Chemistry	3.27	3.70	5.08	5.08	0.081	0.019	0.074
Finnish	4.12	4.86	4.57	4.84	0.579	0.028	0.243
otherAvg	4.35	4.74	4.78	4.97	0.365	0.120	0.948
#exams	5.0	5.5	5.5	5.6	0.796	0.011	0.229

From Table 6 we see a general positive trend, where higher flexibility is connected to higher average grades in all subjects. However, the ANOVA column in Table 4 indicates that the variation in Finnish and physics is not consistent enough for us to make statistically significant claims at the current sample size. In addition to grades, students showing at least some flexibility participated on average in 5.5 exams whereas those with no flexibility only participated in 5.0 exams.

Furthermore, the low *p*-values in Levene's test for advanced mathematics and chemistry indicate that the ANOVA *p*-values are not reliable in these cases. However, ANOVA is known to be a robust test, so the similar group sizes (for these two subjects, see Table 7) and the moderately small maximal difference in variances (largest variance is only 80 % higher than smallest variance) mean that the *p*-values can still be treated as fairly accurate (Harwell, Rubinstein, Hayes, & Olds, 1992).

Table 6 suggest that there may be a different effect of flexibility groups between the subjects. It appears that physics and chemistry have low scores for both no and low flexibility and a jump for medium flexibility, whereas advanced mathematics has better averages especially in the high-flexibility group and even low flexibility is sufficient to give a large boost in basic mathematics scores. However, the number of participants in the current study means that we cannot make reliable conclusions to this effect and we did consequently not follow-up the ANOVA test with *post-hoc* analyses.

We turn now to the ANCOVA column of Table 6. As observed above, the correlation between *F* and *acc* is quite high ($r = 0.607$). Therefore, we studied whether the average grades differed between flexibility groups also when we account for accuracy, i.e. use it as a covariate. As can be seen in the last column of Table 6, when accuracy is used as a covariate, only grades in Basic mathematics are statistically significantly affected by flexibility (with chemistry grades approaching statistical significance).

Relationship between flexibility and matriculation exam participation

It is important to notice that there is a lot of choice for students about which matriculation exams they participate in. This may affect the interpretation of the results on grade averages, as explained in the

next section. Before that, we study what effect flexibility group has on students choosing to participate in each of the five subjects under consideration, advanced and basic mathematics, physics, chemistry and mother tongue. Table 7 gives the total number of students participating in the exams and Table 8 shows the percentage of students in each flexibility group participating in the exams.

Note that participation in matriculation exams is related to some extent with the number of courses the student chooses to study in each subject, although students are not required to take the exams even if they have studied the courses. Nevertheless, students taking the physics exam will, on average, have taken substantially more physics courses than those not taking the exam, and similarly for other subjects. It is quite plausible that taking certain courses may influence one's flexibility (in equation solving). For simplicity, in this section, we only mention the reverse relation, how flexibility relates to taking exams, with further discussion on this issue in the discussion section.

Table 7. Number of students participating in exams in advanced mathematics, basic mathematics, physics, chemistry and mother tongue (Finnish) in each flexibility group (no/low/med/high). The final column shows the p-value of the χ^2test that the distribution of students in a given test is independent of the flexibility group.

	Flexibility group					
Exam	No	Low	Med	High	Total	χ^2
Adv math	13	26	28	27	94	0.000
Basic math	38	10	3	0	51	0.000
Physics	9	13	12	20	54	0.001
Chemistry	11	10	13	13	47	0.122
Finnish	49	36	30	25	140	0.987
Total	54	37	31	27	149	

Table 8. Percentage of students in the flexibility group participating in exams in advanced mathematics, basic mathematics, physics, chemistry and mother tongue (Finnish).

	Flexibility group				
Exam	No	Low	Med	High	All
Adv math	24	70	90	100	63
Basic math	70	27	10	0	34
Physics	17	35	39	74	36
Chemistry	20	27	42	48	32
Finnish	91	97	97	93	94

From Tables 7 and 8 we see that flexibility has a very large effect on student participation in mathematics and physics exams, namely, even a small amount of flexibility substantially increases the likelihood of choosing the advanced over the basic mathematics exam, whereas physics participation is very likely with high flexibility, unlikely with no flexibility, and moderate with some flexibility. There seems to be a modest trend in chemistry participation (higher flexibility related to higher participation), however, the differences are not statistically significant. The Finnish exam is compulsory for all students, so it is not surprising that there is no selection effect here.

We also investigate whether flexibility is related to participation in different exams over and above the influence of accuracy. To this end, a binary logistic regression model was used to predict exam participation with independent variables F and acc. The results are show in Table 9.

Table 9. Binary logistic regression analysis of the influence of flexibility and accuracy on the number of students participating in the exams. The middle columns show the significance level of the influence of the F and acc variables and the final column gives the Nagelkerke R^2 value of the two-parameter model on student participation.

Exam	p (F)	p (acc)	R^2
Adv math	0.0000	0.0020	0.600
Basic math	0.0001	0.0020	0.552
Physics	0.0690	0.0003	0.318
Chemistry	0.1750	0.0310	0.142
Finnish	0.5110	0.2060	0.032

From Table 9, we see that both variables were highly statistically significantly and independently related to participation in both mathematics exams, with high total explanation degree ($R^2 > 0.5$). Participation was quite well predicted also in physics ($R^2 = 0.3$) although flexibility's effect is only approaching statistical significance. The effect for chemistry was smaller, and only related to accuracy.

Estimating the effect of selection on flexibility-group averages

Above we suggested that the selection of which tests to participate in might influence the lack of grade average differences between different flexibility groups in mathematics, physics and chemistry after accounting for accuracy. The premise of the argument is that students tend to opt for the exams in subjects where they are (relatively) strong. To take an extreme case, say students have perfect knowledge of their strength and follow the previous maxim. In the no-flexibility group, 17 % of students participated in the Physics exam, which by the assumption would be the top-17 % of students in this group. In the high-flexibility group, analogously the top-74 % participated in the exam. Obviously, comparing the top-17 % of one group with the top-74 % of another group does not give accurate information about the differences between the whole groups.

In this section, we seek to estimate quantitatively the effect of this possible influence. The situation in mathematics is complicated by the fact that students have three options: take the advanced or basic mathematics test or take no mathematics test. Let us here focus on physics, which showed the clearest pattern of choice and relatively small difference in grade averages.

Table 10 shows some theoretical calculations on how much this kind of selection effect can have on the average (column Δz) and standard deviation (column sd) compared to an initial distribution with standard deviation equal to 1. (Details about the derivation of the table are given in Appendix 1). For instance, consider a medium influence level and selection rate of 20 %: the value 0.69 of Δz means that selective subgroup will have mean 0.69 standard deviations higher than the original group, and the value 0.92 of sd means that its standard deviation will be 92 % of the original group's standard deviation.

Table 10. How a selective subgroup differs from the whole population in terms of mean and standard deviation. In this table the whole population is assumed to be drawn from the normal distribution $N(0,1)$. The different rows describe three scenarios where the skill level of the individual has high/medium/low influence on whether he/she is included in the subgroup. Details on the construction of this table are given in Appendix 1.

	Selection rate							
	20 %		40 %		60 %		80 %	
Influence-level	Δz	sd	Δz	sd	Δz	sd	Δz	sd
high	1.32	0.56	0.91	0.62	0.61	0.70	0.33	0.79

medium	0.69	0.92	0.50	0.91	0.33	0.92	0.17	0.94
low	0.16	1.00	0.12	1.00	0.08	1.00	0.04	1.00

We note that in the "low influence" case there is a small effect on the average and a negligible effect on the standard deviation, whereas in the "high influence" case both effects are substantial, especially when selecting only a small proportion (20–40 %). From our observed data on performance in physics (Table 4), we can calculate what the flexibility-group averages and standard deviations would be if every student would have participated in the test in each group. According to Table 5, 17 %, 35 %, 39 % and 74 % of students in the no, low, med and high groups participated in the physics exam, respectively. Therefore, we select from Table 10 the selection rates 20, 40, 40 and 80 for these four groups. Tables 11 and 12 show the averages for the medium and low influence cases. For instance, let us consider the no-flexibility group with medium influence: we have estimated a mean of 3.24 with standard deviation 1.42; by the model, the selective subgroup (observed data) would have mean 0.69 standard deviations higher than the original group, i.e. $3.24 + 0.69 \cdot 1.42 = 4.22$, which is indeed the value given for this group in Table 6.

We can perform an ANOVA analysis to find out if the corrected group averages differ statistically significantly from one another. In the case of medium influence (Table 11) we get statistically highly significant differences, $p = 0.004$, $F = 5.0286$. In the low-influence case (Table 12), the group differences are approaching statistically significant levels, $p = 0.071$, $F = 2.4845$.

We have, of course, not collected any data on the choice behavior of students. While some theories of human behavior (e.g. the theory of rational choice, see DesJardins & Toutkoushian, 2005, and references therein) might give some estimates, these are likely to be much too vague for precise predictions. The analysis does, however, establish a plausible explanation for the counter-intuitive lack of effect of flexibility on performance in physics and mathematics.

Table 11. Averages and standard deviation estimates for the whole group adjusted by the values in Table 1, with medium influence and 20, 40 or 80 % selection rate as indicated.

gF	*n*	avg	sd	method
no	9	3.24	1.42	med-20
low	13	3.11	2.25	med-40
medium	12	4.34	1.33	med-40
high	20	5.04	1.23	med-80

Table 12. Physics group averages and standard deviations adjusted by the values in Table 1, with low influence and 20, 40 or 80 % selection rate as indicated.

gF	*n*	avg	sd	method
no	9	4.01	1.31	low-20
low	13	3.99	2.06	low-40
medium	12	4.86	1.21	low-40
high	20	5.20	1.17	low-80

Discussion

We measured students' flexibility and accuracy in the linear equation solving domain using 12 flexibility-eligible tasks. High school students showed a wide range of flexibility behaviors, with 64 % of students displaying at least some flexibility. In contrast, only 26 % showed spontaneous flexibility, i.e.

the kind of flexibility that would show also on a normal mathematics test. This indicates that the flexibility test tells us more about student flexibility than what could be gleaned from regular exams.

We found that flexibility in the equation solving test predicted success in national matriculation exams 1–2 years later in mathematics, chemistry and Finnish, as well as the number of different exams taken, whereas Physics or an other-subjects-average were not affected (Table 6, column ANOVA). However, when considering accuracy as co-variate, only basic mathematics grades were strongly affected by flexibility (Table 6, column ANCOVA).

We were surprised by this relative lack of flexibility effect. Flexibility is seen as a characteristic of mathematical experts (Heinze, Star & Verschaffel, 2009), and we expected high flexibilityto be related to good mathematics and physics grades. Furthermore, it has been established that procedural and conceptual mathematical knowledge predicts future procedural flexibility (Schneider, Rittle-Johnson and Star, 2011), and also the opposite direction seems natural.

There are several possible explanations for these discrepancies. First, we measured accuracy with flexibility-eligible tasks, which means that students using the innovative strategy had to perform fewer and less complex steps to arrive at the answer than students using the standard solution. Consequently, two students who are equally good at the performing the technical steps of a solution process would nevertheless get different scores on accuracy if one uses the innovative solution and the other uses the standard solution. This effect is limited by the fact that we assessed accuracy only during the first phase of the test, during which few students (39 out of 149) provided any innovative solutions. Nevertheless, our measure of accuracy also partially measures flexibility and an accuracy measure derived from non-flexibility-eligible tasks would likely have lower correlation with flexibility than ours. It follows by a comparison of the columns ANOVA and ANCOVA in Table 6 that taking *acc* as a covariate may lead to an overly conservative estimate of the role of flexibility as a predictor of success the matriculations exams.

Second, Heinze, Star & Verschaffel's (2009) argument for the relation between flexibility and mathematical expertise focuses on representational flexibility, so it is possible that the link does not pertain to strategy flexibility. Concerning the study of Schneider, Rittle-Johnson and Star (2011), we observe that it was conducted within a single area of mathematics, whereas we here consider the much broader generalization of flexibility in linear equations to mathematical proficiency in geometry, calculus, probability, etc.

On the other hand, we showed that different participation rates between flexibility groups might have a substantial effect on the observed exam performance, if we assume that students are likely to participate in those exams where they expect to perform best. In particular, in physics, this effect may explain why we did not find statistically significant grade averages between the flexibility groups. On the other hand, in chemistry there was less of a selection effect and a greater outright influence of flexibility on grades, which supports this hypothesized effect. However, an alternative and equally plausible explanation is that participating in physics and advanced mathematics classes leads to higher flexibility in equation solving contexts. The latter effect could be alleviated in future studies by collecting the flexibility data at an earlier date, at the beginning of high school when all students are still participating in the same courses.

This study raises several questions, which are left for future research. If accuracy were not measured with flexibility-eligible tasks, would its relation to flexibility still be as strong, or would flexibility emerge more clearly as an independent predictor?In Table 6, we observed some tentative differences between different subjects on what level of flexibility is required for higher performance, namely even low flexibility seemed better in basic mathematics, medium flexibility was required in chemistry and

physics, whereas advanced mathematics seemed to require high levels of flexibility. A broader study would be useful to lend credence to these observations.

In this study, we considered the relation between flexibility in equation solving to grades in different subjects. If flexibility had been measured not in equation solving but some other domain, would the results still be the same? One could also consider the relationship between flexibility in mathematics (equation solving) and flexibility in the other domains, i.e. the transfer of flexibility across domains.

Acknowledgements

This research was supported by the Finnish Cultural Foundation and the Jenny and Antti Wihuri Foundation.

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.), *Studies in mathematical thinking and learning. The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1–33). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

DesJardins, S.L., & Toutkoushian, R.K. (2005). *Are Students Really Rational? The Development of Rational Thought and its Application to Student Choice*. In: Smart J.C. (eds) Higher Education: Handbook of Theory and Research. Higher Education: Handbook of Theory and Research, vol 20. Dordrecht, Netherlands:Springer.

Harwell, M., Rubinstein, E., Hayes, W., & Olds, C. (1992). Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. *Journal of Educational Statistics, 17*(4), 315-339.

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.

Kerlinger, F.N., & Lee, H.B. (2000). *Foundations of behavioral research*. 4th Ed. Orlando, FL, Harcourt college publishers.

Maxwell, S.E., & Delaney, H.D. (2003). Designing experiments and analyzing data: A model comparison perspective. Routledge: Abingdon, United Kingdom.

McMullen, J., Brezovszky, B., Hannula-Sormunen, M.M., Veermans, K., Rodríguez-Aflecht, G., Pongsakdi, N., & Lehtinen, E. (2017). Adaptive number knowledge and its relation to arithmetic and pre-algebra knowledge. *Learning and Instruction*, *49*, 178–187.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common core state standards for mathematics. Washington, D.C.: Authors.

Opetushallitus (2015). Lukion opetussuunnitelman perusteet. Helsinki, Finland: Authors.

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.

Rittle-Johnson, B., Star, J.R., & Durkin, K. (2012). Developing procedural flexibility: Are novices prepared to learn from comparing procedures? *British Journal of Educational Psychology*, *82*(3), 436–455.

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*, *47*(6), 1525–1538.

Star, J.R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and instruction*, *18*(6), 565–579.

Verschaffel, L., Luwel, K., Torbeyns, J. & Van Dooren, W. (2007). Conceptualizing, investigating, and enhancing adaptive expertise in elementary mathematics education, *European Journal of Psychology of Education, 24*: 335–359.

Wright, D.B. (2006). Comparing groups in a before–after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology*, *76*(3), 663–675.

Xu, L., Liu, R., 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.

Appendix 1. The construction of Table 8.

All numbers in Table 10 are based on theoretical calculations in the context of a base population drawn from the normal distribution $N(0,1)$. The probability of being included in a subsample is modeled by a sigmoid curve given by $\frac{1}{1+e^{-(x-x0)/s}}$. The parameter s determines how steep the cut-off function is: a small s means that few individuals below the threshold value $x0$ are included in the sample, whereas a large value for s means that the curve is relatively flat and has less impact on the selection. The probability distribution function (pdf) for the sub-group was obtained as the product of the sigmoid function and the pdf of the normal distribution:

$$\frac{ce^{-x^2/2}}{1+e^{-(x-x0)/s}},$$

where the constant c is chosen such that the total mass of the distribution is 1.

For our analyses, we chose four sub-sample sizes r (20 %, 40 %, 60 % and 80 %) that we wanted to inspect. The parameter s was had values 0.2 (high influence), 1 (medium influence) and 5 (low influence). These two parameters determine $x0$ by the integral equation

$$\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{1+e^{-(x-x0)/s}}\,dx = r,$$

which was solved numerically with MATLAB. Once the parameters s and $x0$ were fixed, the resulting distribution was used to calculate the expected value and standard deviation. The resulting curves are shown in Figures 1–3.

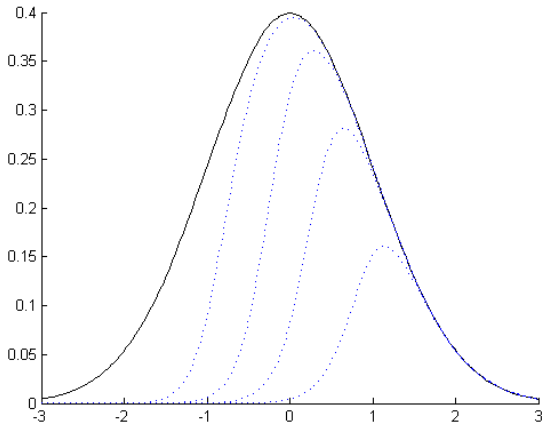

Figure 1. High influence of skill on choice ($s = 0.2$). The black curve is the normal distribution, whereas the dotted curves correspond to the 20–80 % sub-groups.

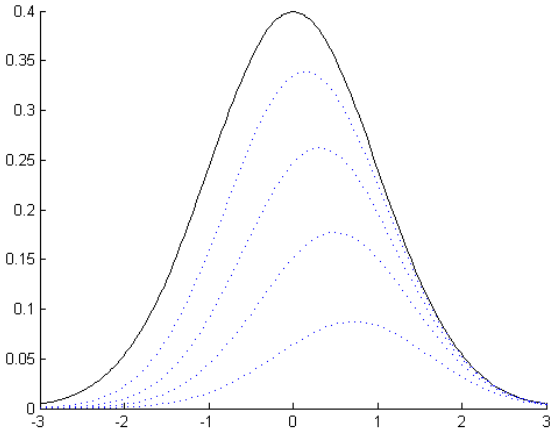

Figure 2. Medium influence of skill on choice on ($s = 1$). The black curve is the normal distribution, whereas the dotted curves correspond to the 20–80 % sub-groups.

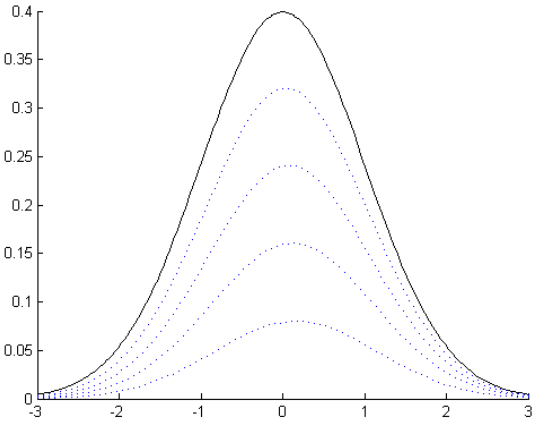

Figure 3. Low influence of skill on choice on ($s = 5$). The black curve is the normal distribution, whereas the dotted curves correspond to the 20–80 % sub-groups.