

## The Effect of Computer-Based Feedback in Game-Like Fluid Reasoning Tasks



Šárka Portešová , Michal Jabůrek , Adam Ťápal , Ondřej Straka 

The Institute for Research on Children, Youth and Family, Masaryk University, Brno, Czech Republic

We focused on the effect of various types of feedback in a game-based fluid reasoning test called *Triton and the Hungry Ocean* on elementary school students (ages 8-12; total  $N = 321$ ). The feedback types were four: no feedback (A), simple (correct/wrong feedback; B), elaborated (correct solution shown; C), and learner-controlled feedback (student chooses between feedback types; D). We did not observe an effect of any feedback type on performance (i.e., there were no between-group differences). However, within group D, students overall tended to choose elaborated feedback more often as task difficulty increased ( $r = .92$ ), and those in group D who generally tended to choose elaborated feedback also tended to perform better even after controlling for intellect.

*Key words:* feedback, learning, game-based performance tests, metacognition

### Introduction

Feedback is essential for learning. The concept of feedback is used for a variety of purposes in various disciplines, such as medicine, management, or sports. However, it gained significant research attention especially in educational contexts. Researchers in this field have been trying for almost a century to understand how feedback can be optimized to maximize value for students and lead

them towards growth-oriented development (Lipnevich et al., 2016).

Feedback is generally considered essential to improve the acquisition of knowledge and skills (e.g., Azevedo & Bernard, 1995; Moreno, 2004). For example, Hattie (1999), in his review of 196 studies of feedback in the classroom, described feedback as one of the most influential factors in learning. Hattie and Timperley (2007) define feedback as “information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding

Correspondence concerning this article should be addressed to Michal Jabůrek, The Institute for Research on Children, Youth and Family, Faculty of Social Studies, Masaryk University, Joštova 10, Brno 602 00, Czech Republic. E-mail: [jaburek@fss.muni.cz](mailto:jaburek@fss.muni.cz)

Description of game mechanics, translation of opening and closing stories of *Triton* game, table with estimated model parameters (items difficulty and the variance of the latent variable), dataset and R software script used in analyzes is openly available in figshare at <https://doi.org/10.6084/m9.figshare.13525886.v1>

Received December 2, 2020



aspects of one's performance or understanding" (p. 81). The most important function of feedback is to provide students with information about their learning or performance so that they can successfully regulate it (Butler & Winne, 1995; Hattie & Gan, 2011; Hattie & Timperley, 2007).

Various meta-analyses and systematic reviews have focused on the effects of feedback on learning. There are many variables that play a role in determining its potential to affect learning. However, the results of large-scale meta-analytical studies on the actual effectiveness of feedback lead to conclusions that could be described as contradictory or, at the very least, inconsistent. The most comprehensive review, based on 131 studies and meta-analysis of the effects of individual feedback on performance, was conducted by Kluger and DeNisi (1996). Their overall finding was that the average effect of evaluative feedback on performance was  $d = .41$  – in other words, groups receiving feedback on average outperformed their respective control groups by roughly four tenths of a standard deviation – the equivalent of moving from the 50th to the 66th percentile on a standardized test. However, more than 38 percent of the effect sizes from the analyzed studies were negative, that is, showed that control groups outperformed feedback groups. In the classroom context, Hattie and Timperley (2007) and Hattie and Zierer (2019) conducted meta-syntheses on the effects of feedback on student achievement. These indicated a large effect (between  $d = .70$  and  $d = .79$ ) of feedback on student achievement in general. However, the authors noted the considerable variance of feedback effects on achievement. When considering 435 studies, a recent meta-analysis (Wisniewski et al., 2020) revealed a medium positive effect size of feedback ( $d = .48$ ) on student learning. According to this extensive

study, the impact of feedback is significantly influenced by the information and content being conveyed. Overall, we can say that it is necessary to interpret different forms of feedback independently.

Most of the research into the effects of feedback is instructional. Its primary goal is monitoring students' learning in response to instructions and asking students to confirm, refine or clarify their misunderstandings. There has also been a considerable number of studies looking at the impact of performance (evaluative) feedback, whose primary goal is to evaluate students' performance on assessment tasks (Kluger & DeNisi, 1996). However, unfortunately, to the best of our knowledge, no study has been published in the last twenty years that has addressed in detail the effect of different types of feedback in logic tasks. This study intends to fill this gap. We sought to determine the effect of evaluative feedback in fluid reasoning tasks. Specifically, we focused on the effects of four types of computer feedback on performance in a video-game with fluid-reasoning tasks administered to elementary school students. We were also interested in whether and how, if given the chance, would the students choose a preferred type of feedback, and whether this choice is dependent on task difficulty and student's performance.

### **Feedback Characteristics**

Feedback can have different functions depending on the learning environment in which it is studied and the particular learning paradigm under which it is viewed. It occurs in different forms, has different purposes, acts upon different levels of learning (Mory, 2004). In the following sections, we introduce the most common criteria used to describe various kinds of feedback.

### *Feedback and Degree of Complexity*

With regards to the degree of feedback elaboration, one can distinguish between simple feedback, consisting of a short response (yes/no, right/wrong; Kulhavy & Stock, 1989; Mory, 2004), and elaborated (complex) feedback, presenting the correct answer or solution. The degree of elaboration strongly differs in various studies, elaborated feedback is usually more effective than simple feedback and is beneficial to both lower and higher order learning outcomes (Attali & Van der Kleij, 2017).

### *Learning Outcomes*

There are a number of hierarchical classifications with respect to learning outcomes. However, the distinction is usually made between lower-order and higher-order learning outcomes (Kulik & Kulik, 1988; Van der Kleij et al., 2012; Van der Kleij et al., 2015). Lower-order learning outcomes include recognition, understanding and memorization (Van der Kleij et al., 2012; Van der Kleij et al., 2015). Higher-order learning outcomes include intellectual skills such as analytic and procedural skills, and require the learner to apply knowledge and skills to new situations (Smith & Ragan, 2005). It can be said that more complex learning requires more complex, elaborated feedback that goes beyond mere correction (Attali & Van der Kleij, 2017; Smith & Ragan, 2005).

### *Feedback Target*

Feedback can be provided during instruction or serve as an evaluation tool. When provided during instruction, its primary goal is monitoring students' learning in response to instructions and asking students to confirm, refine

or clarify their understanding. Conversely, feedback provided to students regarding their performance in an assessment is based on the assumption that the feedback provides the examinee with useful information about their progress (Attali & van der Kleij, 2017; Kluger & DeNisi, 1996; Panero & Aldon, 2016; Wang et al., 2019).

The effect of feedback in both instructional and evaluative contexts on performance is quite variable. Under certain conditions, feedback improves performance, under others, no impact on performance can be determined, and under other conditions still, feedback can degrade performance (Hattie & Timperley, 2007; Hattie & Zierer, 2019; Kluger & DeNisi, 1996; Wisniewski et al., 2020).

### **Effects of Assessment Feedback on Logical Reasoning Test Achievement**

A specific type of evaluative feedback provided in the context of higher-order learning outcomes is simple feedback verifying the correctness of the response during logical reasoning assessment. Unfortunately, relatively little is known about feedback effects in psychometric assessment settings.

Because feedback in logical reasoning assessment offers an opportunity to learn during the test session, some studies view feedback in this context as a piece of important information that can result in performance improvement in a subset of tasks or on the entire test (Guthke & Beckmann, 2003; Guthke & Stein, 1996). These studies are conducted most often within the context of dynamic testing, in which intra-individual differences in performance scores are attributed to varying individual ability to process and learn from received feedback. This testing method has the added benefit of being able to recognize not only already developed abilities but also the learning potential of each participant

to develop in the future (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002; Veerbeek et al., 2017).

The ability to integrate further information from feedback while solving new problems is, in some cases, considered an integral part of fluid reasoning. For instance, the Concept Formation and Analysis-Synthesis test of the Woodcock-Johnson IV – Test of Cognitive Abilities battery (WJ-IV; McGrew et al., 2014) uses the ability to learn from immediate feedback as a part of its measurement. Interestingly, other tests of the same latent ability (fluid reasoning), such as those from the WISC 5 battery (Wechsler, 2014), do not account for feedback at all. Does that mean such tests are lacking in content validity? Or is the ability to learn from immediate feedback negligible from the perspective of the assessment situation? Naturally, the question of feedback standardization is an important one here – if one cannot ensure that feedback given will be relatively stable across administrations (i.e., available at all times and in the same form), the validity and fairness of tests utilizing feedback could be significantly impacted. This could be one of the reasons why feedback-giving is sought to be avoided in common assessment situations, however, with the ascent of computer-based assessment, it is becoming less and less relevant.

In methodological studies on the other hand, it is often hypothesized that feedback might improve outcomes by other means than by directly enhancing the measured ability, such as by removing anxiety or providing reassurance (Rocklin et al., 1995). Beckmann, Beckmann, and Elliott (2009) have found that feedback on item performance during cognitive ability testing on the whole did not affect performance; however, feedback interacted with self-confidence and goal orientation to produce positive effects in adolescents low in self-confidence and with high performance goal orientation.

Regardless of these findings, the integration of simple feedback mechanisms into standardized cognitive tests is often seen as being in direct contradiction to valid testing procedures (Guthke & Beckmann, 2000; however see above). It seems, however, that evaluating the effect of feedback in cognitive tests could actually benefit assessment of cognitive skills in many ways. To better reflect a person's current state and potential, it would be informative to know the extent to which feedback is followed by improvement in performance – i.e., to gauge the ability to learn (Guthke & Beckmann, 2000; Veerbeek et al., 2017). Furthermore, we consider it key to be aware of the extent to which it is indeed necessary to keep cognitive assessment feedback-free, and how big is the risk of test procedure bias when different types of feedback are given (Fuchs & Fuchs, 1986; Reynolds & Suzuki, 2012). Finally, we think the effect of feedback (or the lack thereof) is ultimately informative with respect to test validity.

### **Aims of the Present Study**

To fill the gap in feedback research, we focus on the effects of various types of feedback, as categorized by Shute (2008), on fluid reasoning task performance. We developed a video-game consisting of fluid-reasoning tasks with four different feedback mechanisms that are triggered after every response given by the participant, making this an immediate feedback in all cases. The feedback types are:

No feedback (condition A): Provides no information about the response.

Simple/verification feedback (condition B): Informs whether the response is correct or incorrect.

Elaborated/animated feedback (condition C): Shows the entire procedure needed to solve the task in an animated form.

Learner-controlled feedback (condition D): Allows for the deliberate choice of feedback

from any of the aforementioned alternatives (A, B, C), as many times as is requested.

Thus, our aim is to investigate the effect of the four types of feedback on learners' performance in the game. We formulate the following hypotheses and research questions.

### Effects of Feedback on Game Performance

Current research is mostly in agreement that simple/verification feedback is inefficient, especially in the context of higher-order learning outcomes (Van der Kleij et al., 2015). On the other hand, elaborated feedback is beneficial for both lower-order and higher-order learning outcomes (Attali & Van der Kleij, 2017). For example, in their meta-analysis, Van der Kleij et al. (2015) have found that elaborated feedback (EF; analogous to our condition C) produced larger effect sizes ( $d = .49$ ) than feedback regarding the correctness of the answer (KR; analogous to our condition B;  $d = .05$ ). EF was particularly more effective than KR for higher order learning outcomes.

Hence, it seems that in the evaluative context (namely in performance tasks), revealing to the learner the entire process of arriving at the correct solution is essential for facilitating improvement. When given elaborated feedback, the students can monitor their own thought process and confront it with the correct solution. In accordance with the meta-analysis findings, we expect the effect of condition B to be too small to be detectable ( $d < 0.1$ ), however for the elaborated feedback (condition C) we do expect to find a medium effect size ( $d = 0.4-0.5$ ). Furthermore, learner-controlled feedback does not seem to directly lead to improved performance (Alevén et al., 2006; Timmers & Veldkamp, 2011; see below for more details), hence we expect small effect of condition D ( $d < 0.1$ ).

*H1: Feedback conditions have an effect on game performance. Specifically, no effects of sim-*

*ple feedback condition (B) or learner-controlled feedback condition (D) will be found, whereas the elaborated feedback condition (C) will show an effect over no feedback (A), simple feedback (B), and learner-controlled feedback (D).*

### Learners' Control over the Feedback Message (condition D)

When assessing the efficiency of feedback it is common to allow learners to choose their preferred feedback type, however, as mentioned above, this itself does not directly lead to improved performance and learning (Alevén et al., 2006; Timmers & Veldkamp, 2011). Students must be willing to invest the necessary time and effort and possibly also possess certain metacognitive abilities to make the best possible use of the feedback they receive (Timmers & Veldkamp, 2011). This phenomenon is closely related to the students' interest in monitoring their own learning process and improving themselves (Bangert-Drowns et al., 1991). Hence, we hypothesize that students who more often choose elaborated feedback when given the chance (feedback C in condition D) use better metacognitive strategies and thus will attain better overall outcomes in the game compared to those who choose this type of feedback less often. Based on the findings of the meta-analysis cited above (Van der Kleij et al., 2015), we expect a medium effect ( $d = 0.4-0.5$ ).

*H2: In learner-controlled feedback (condition D), children who choose elaborated feedback more frequently will show better performance in the game than children who choose it less frequently.*

### Exploring the Strategies in Learner-controlled Feedback and How They Connect to Task Complexity and Game Performance

An interesting question arises when the learner is given control over the form of feedback

they receive. What form of feedback will be chosen under different circumstances? Will there be any observable consistent strategies? It has been shown that students choose elaborated feedback most frequently after an incorrect response to clarify errors and misunderstandings (Mory, 2004), and that the effect of feedback is moderated by task complexity (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996; Kulhavy & Stock, 1989; Mory, 2004). It is reasonable to assume, therefore, that feedback choice will differ when an incorrect (versus correct) answer was given, and the choice itself will be a function of the (perceived) task difficulty. As there is little previous research to draw any expectations from, we do not formulate any hypothesis here and instead will perform an exploratory analysis of our data.

### Methods

#### *CFT 20-R*

The CFT 20-R (Fajmonová et al., 2015) is a Czech adaptation of Cattell's non-verbal test, used to measure fluid intelligence in children aged 7.5 to 15. The test comprises four subtests and is divided into two parts for a total of 101 tasks. According to the test's manual, its internal consistency is  $\alpha = .88$ . Please note again that this reliability is reported from the test's standardization.

#### *Triton and the Hungry Ocean*

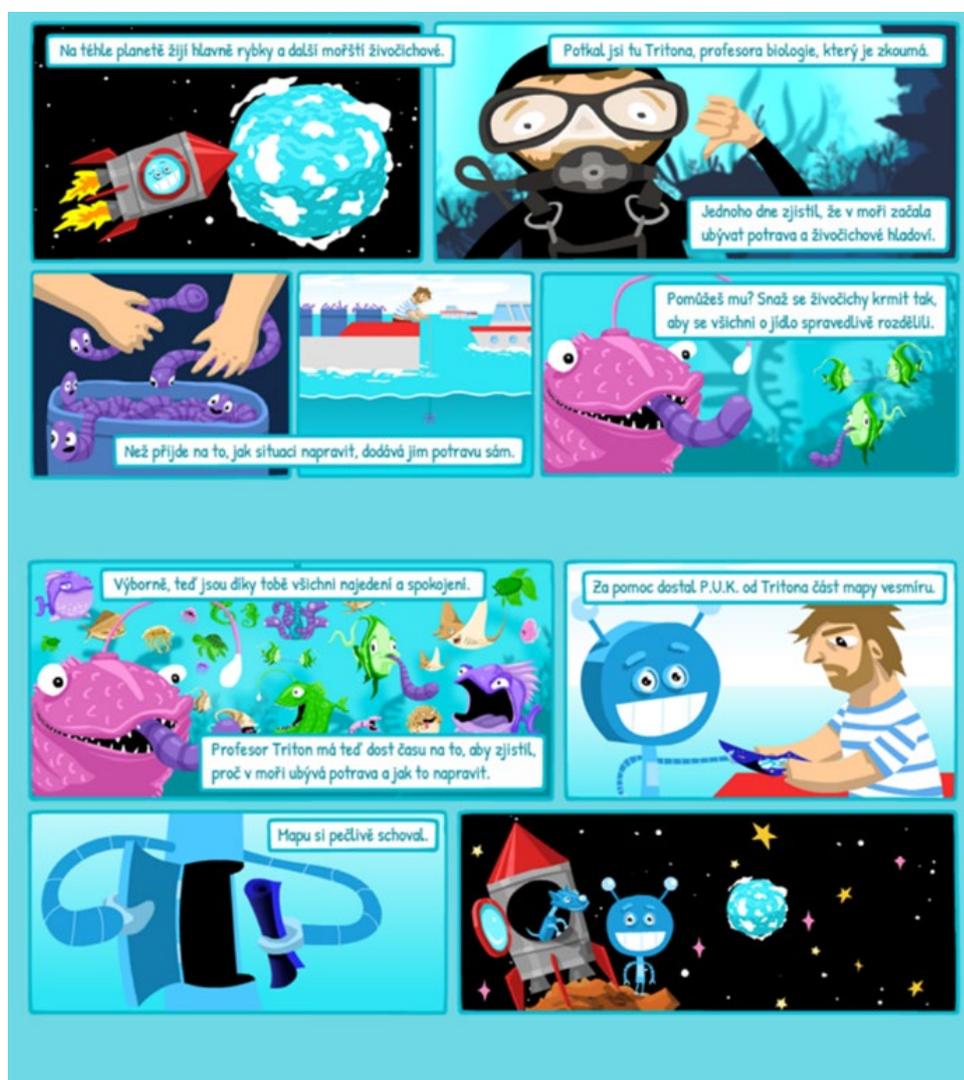
The logical game *Triton and the Hungry Ocean*<sup>1</sup> (referred to here as *Triton*) is based on the "balance beam task" of Inhelder and Piaget (1958), which was later adopted by other authors and is, for example, referred to as Figure Weights in the complex intelligence test batteries WAIS-IV (Wechsler, 2008) and WISC-V (Wechsler, 2014). The objective of

<sup>1</sup> A technical manual of *Triton* is available for download on the website [www.invenio.muni.cz](http://www.invenio.muni.cz).

the task is to choose a set of weights for one balance beam to counterbalance the weights on another one. *Triton* practically re-uses this principle, while bringing the task aesthetically to a submarine setting. The game starts and ends with a simple story (see Figure 1) and uses cartoon-like graphics and simple sounds. It also introduces additional game mechanics not typically present in similar tasks (see the supplemental material online) and does not have a time limit for individual tasks.

A sample task is shown in Figure 2, in which the individual features of the game are highlighted. The main part of the screen features a long hook with a worm (see Feature 1). On either side of this hook, there is a circle outlined by bubbles. On the left side, the circle contains a certain number of animals (Feature 2), while on the right side, the circle is empty; this we further refer to as a *slot* (Feature 3). The player is supposed to fill this empty slot by moving one of the five groups of animals from the bottom part of the screen (Feature 4) to balance out the strength of the animal group on the left side (Feature 2). The main rule is that animals of the same color, shape, and number have the same strength. In more complex tasks, the strength of individual animals is expressed via so-called conditions; shorter hooks with both sides already occupied and balanced (Feature 5). These conditions imply the relative strength of specific animal types.

Besides moving groups of animals from the bottom part of the screen to the slots and back, the player is allowed to reset the task (i.e., return all the features into the original state) by pushing the "reset" button (Feature 6). They can also proceed to the next task by pushing the "play" button (Feature 7). Furthermore, there are two more game buttons, which are, however, not necessary to solve the tasks – a sound on/off button (this affects only accompanying sounds, the instructions



Note. The translation is available online in the supplemental material.

Figure 1 The opening and closing stories of Triton.

remain audible; Feature 8) and a pause button to interrupt the game (Feature 9).

In order to solve the task, the player needs to deduce the relative strength of individual animals, applying primarily logical reasoning. In general, the abilities applied here fall into

the category of fluid reasoning, which is defined as logical reasoning intentionally and purposefully aimed at solving novel "on-the-spot" problems that cannot be solved using previously learned habits, schemas, or scripts (e.g., Schneider & McGrew, 2018).

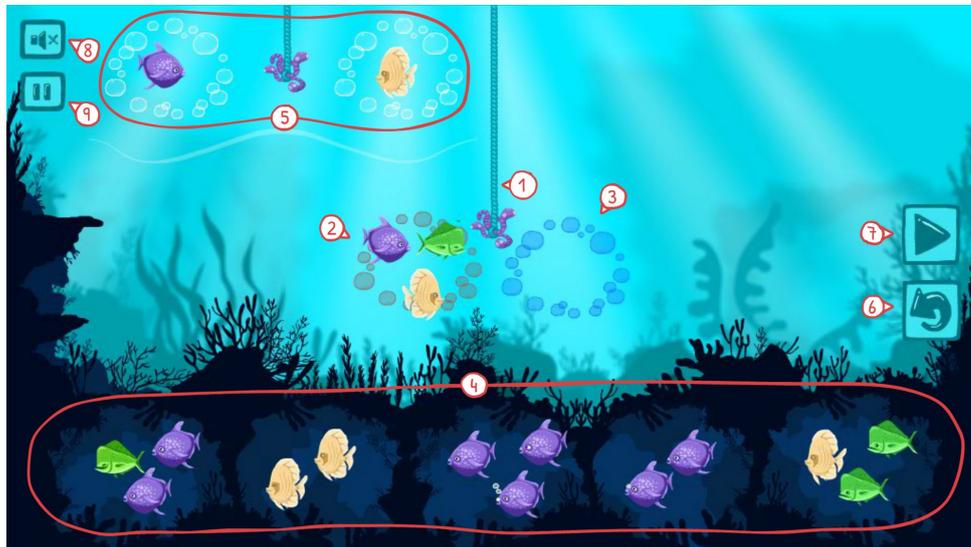


Figure 2 Sample task and individual game features of Triton.

The game consists of 32 tasks, the complexity of which gradually increases, as does the number of game mechanics involved. Description of these mechanics is available online in the supplemental material. Each new game mechanic is introduced by a narrated video demonstration, in which a task with the new mechanic is solved. All participants receive the same instructions on how to perform all tasks.

As mentioned previously, four feedback conditions were implemented into the game for the purposes of this study. Feedback was given to the players after completing each task upon pressing the “play” button. In Table 1, the form of the individual feedback conditions is specified.

The following data were logged for each participant: 1) Correctness/incorrectness of response to every task. 2) The sequence of choices of the individual feedback conditions (A, B, C) in condition D.

### Data Collection and Participants

The pilot version of *Triton* (with only simple/verification type of feedback) was first administered at two schools to a total of 127 students attending grades 3–5 in order to verify whether the instructions were comprehensible and confirm that the game’s technical aspects function correctly. As no problems occurred during data collection, the data were further used to construct the model of measurement.

Afterwards, five conveniently selected public schools participated in the main stage of the research. All students from participating grades (grade 3 – grade 6) whose legal guardians agreed with their participation ( $N = 321$ ) were first administered the CFT 20-R test by pencil/paper in a group setting. Afterwards, they were assigned to the four experimental groups so that the distribution of CFT 20-R

Table 1 *Characteristics of the individual types of feedback in Triton*

Feedback condition	Type of feedback	Detailed characteristics of feedback	Visuals
A	No feedback	No information on the correctness or incorrectness of the response, only a text showing "Thanks for your answer. Keep on playing."	
B	Simple/verification feedback	Correct responses will trigger a text showing "Correct answer!", while the message "Incorrect answer." follows incorrect responses.	
			

Table 1 continues

Table 1 (continued)

Feedback condition	Type of feedback	Detailed characteristics of feedback	Visuals
C	Elaborated/animated feedback	Regardless of whether the response was correct or incorrect, a text prompt showing “Show the correct solution.” is displayed. After clicking the text, the solution is presented and explained via a simple animation. After the animation is over, the solved task can be viewed until the “play” button is clicked, after which a new task starts. The button can also be clicked any time during the animation to fast forward to a new task.	 <p>Animated sample is available on the link: <a href="https://bit.ly/3g0j7q8">https://bit.ly/3g0j7q8</a></p>
D	Learner-controlled feedback	Regardless of whether the response was correct or incorrect, any of the previously mentioned types of feedback can be chosen (A, B, or C). If A is chosen, a new task starts. If B is chosen, the verification message described in B is shown. If C is chosen, the animation described in C is shown. The B and C feedbacks can be selected repeatedly – a new task is presented only after choosing A.	

would be balanced, secondary counterbalancing was done with respect to gender and school grade. Subsequently, each group was administered *Triton*, with a different feedback condition, otherwise identical to the pilot version. All students worked on their own and at their own pace on classroom desktop PCs. Students controlled the game with a mouse and were presented instructions into their headphones within the game environment. Note, the sample thus collected was convenient and its size was not driven by an *a priori* power

analysis. As such, the hypothesis tests performed here might not have sufficient power.

Table 2 breaks down the sample by school grade, gender, age, and individual feedback conditions.

## Results

### Hypothesis 1

First, a Rasch model was fit using the *mirt* package (Chalmers, 2012) in R using all avail-

Table 2 Summary of the sample for each version of the game/type of feedback

Feedback type	Grade 3	Grade 4	Grade 5	Grade 6	Total	Mean IQ (SD)*	Mean age (SD)	Mean Triton performance (SD)**
Feedback A	25 (48%)	20 (30 %)	17 (47 %)	23 (48 %)	85 (44 %)	109 (15.8)	10.6 (1.24)	0.72 (0.17)
Feedback B	17 (41 %)	17 (59 %)	16 (31 %)	26 (38 %)	76 (42 %)	110 (15.6)	10.6 (1.12)	0.71 (0.14)
Feedback C	21 (43 %)	19 (32 %)	19 (47 %)	22 (48 %)	81 (42 %)	108 (15.7)	10.5 (1.16)	0.71 (0.15)
Feedback D	19 (42 %)	21 (24 %)	15 (67 %)	24 (42 %)	79 (42 %)	110 (14.3)	10.5 (1.13)	0.74 (0.16)

*Note.* The percentages in parentheses refer to the proportion of girls in that subgroup of the sample. Within the pilot version, data on the gender of participating respondents were not collected.

\* CFT 20-R; \*\* Due to missing data, an average score was calculated for each student, and the mean and *SD* of the average score was taken.

able data (i.e., also utilizing the data from students who only participated in the pilot study). This model serves the integral purpose of modeling the game as a test where the correctness of a child's response to a test item (i.e., a task in the game) is a function of the child's (unknown) latent ability level and the task difficulty. Without a well-fitting measurement model of this kind, many subsequent analyses using the data could be called into question. The model fits reasonably well,  $M_2(496) = 750.3$ ,  $p < .01$ , RMSEA = 0.04 (95% CI: 0.03, 0.05), TLI = .94, SRMSR = 0.09, with an empirical reliability of .85. It is important to keep in mind that the fit of the model to data is approximate and not perfect, as evidenced by the test of the  $M_2$  statistic.

Subsequently, the model was refit as a generalized linear mixed model (with a logit link

function) using the *lme4* package (Bates et al., 2014) in R, now only using data from students who participated in the experimental conditions. Then, another such model was fit which was identical to the previous one except that the feedback type (A, B, C, or D) was added as a covariate to predict the probability that a task will be answered correctly, making it effectively a latent regression Rasch model. A likelihood-ratio test, comparing the latent regression model with the baseline measurement model, was conducted to see whether including the feedback type as predictor did, in fact, improve the model fit significantly. Based on the comparison (see Table 3), we conclude that feedback type has no effect on the probability of tasks being solved correctly, the difference between the two models' deviances being 1.4 with  $df = 3$ ,  $p = .69$ .

Table 3 Model comparison for Hypothesis 1

	<i>npar</i>	Deviance	$\Delta$ Deviance	$\Delta$ <i>npar</i>	<i>p</i>
Rasch model	33	7334.3	-		
Latent regression	36	7332.9	1.4	3	.69

Table 4 Model comparison for Hypothesis 2

	<i>npar</i>	Deviance	$\Delta$ Deviance	$\Delta$ <i>npar</i>	<i>p</i>
Rasch model	33	1746.0	-		
Latent regression	34	1739.8	6.23	1	< .05

## Hypothesis 2

The same baseline measurement model was used as in the previous hypothesis, but only with students who were given tasks with feedback type D. With this type of feedback, after solving any task participants were offered the choice of no feedback, correct/incorrect feedback, or elaborated feedback showing the correct solution. The students could select any feedback as many times as they wished until continuing with the game. The choice of feedback was recorded. For each time a child solved a task, they were assigned a score of 1 if they chose the elaborated feedback at least once, and a score of 0 if they did not. This dichotomization served the purpose to eliminate the effect of extreme observations for children who may have selected the elaborated feedback a large number of times for no apparent benefit. For each child, we computed the average of this score (in effect, the average frequency with which an elaborated feedback was chosen) across all tasks. The distribution of this average was strongly left-skewed and zero-inflated, with an average of 0.64 ( $SD = 0.33$ ).

Afterwards, a latent regression Rasch model was fit using the calculated average as a covariate. A likelihood-ratio test comparing

the latent regression model with the baseline measurement model showed that including the covariate improves model fit,  $\Delta$ Deviance = 6.23 with  $df = 1$ ,  $p < .05$  (see Table 4). The estimated latent regression coefficient was  $\beta = 1.24$ , while the estimated latent score variance without and with the latent regression was  $s^2 = 2.09$  and  $s^2 = 1.88$ , respectively. Thus, the frequency of choosing the elaborated feedback explains roughly 10% of the variance in latent score estimates for students in feedback condition D.

## Exploring the Strategies in Learner-controlled Feedback

Still using the sub-sample of students in feedback condition D, we then analyzed the choice of feedback per task. For each task, we calculated the proportion of students who chose either the correct/incorrect feedback, the elaborated feedback, or both. As shown in Figures 3 and 4, the frequency of a chosen correct/incorrect feedback was linearly independent of task difficulty (i.e., the proportion of incorrect responses) with a correlation of  $r = -.08$ , while the relationship was strongly linear for the elaborated feedback,  $r = .92$ .

Naturally, we expected that as tasks became harder, increasing the likelihood of an incorrect response, the use of elaborated

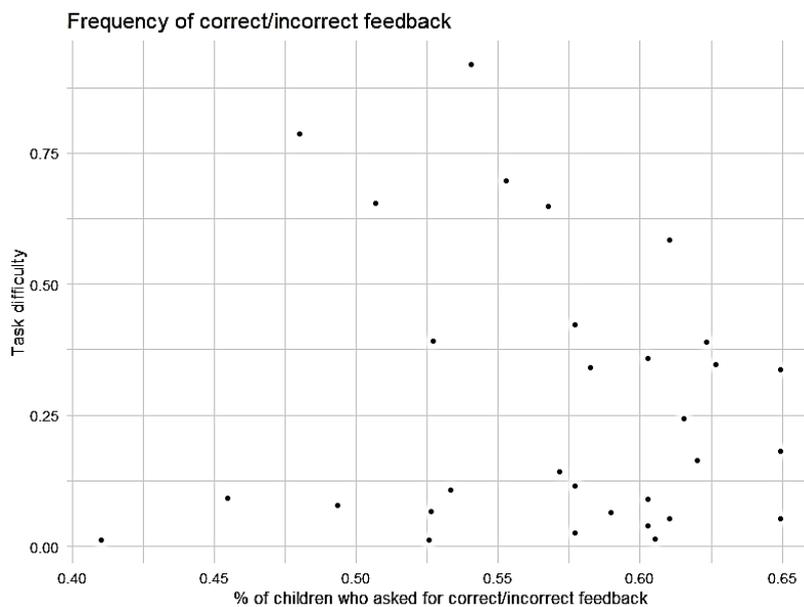


Figure 3 Frequency of correct/incorrect feedback based on task difficulty.

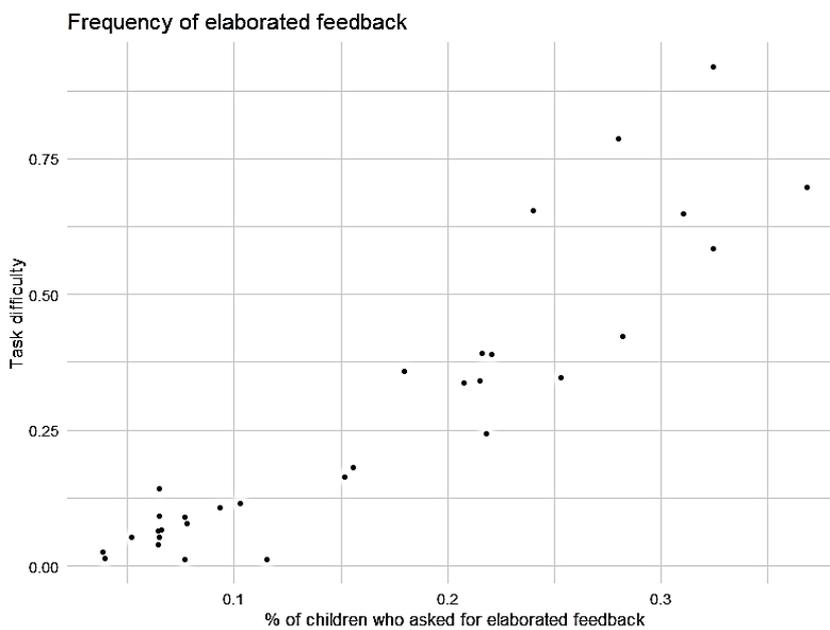


Figure 4 Frequency of elaborated feedback by task difficulty.

Table 5 *Relative frequency of feedback choice*

	Correctly answered tasks	Incorrectly answered tasks
No feedback chosen	36 %	36 %
Only correct/incorrect	59 %	17 %
Only elaborated	3 %	17 %
Correct/incorrect, then elaborated	1 %	28 %
Elaborated, then correct/incorrect	0 %	1 %

feedback would be chosen more often, especially after learning from correct/incorrect feedback that an incorrect response had been given. Indeed, the proportions of feedback use (Table 5) suggest this to be true.

### Discussion

Our experimental study deals with the effect of four types of computer feedback (no feedback, correct/incorrect feedback, animated elaborated feedback, and learner-controlled feedback) on performance in a game consisting of fluid reasoning tasks, while taking intelligence and age into account. Moreover, we also analyzed strategies of feedback use among students who were given a choice over their preferred feedback type.

We assumed that always receiving elaborated feedback will lead to improved overall performance in the game, as is the case in instructional higher-order learning (Attali & Van der Kleij, 2017). However, our study did not reveal any relationship between performance and elaborated feedback or any other feedback condition. We find this result surprising, as higher learning outcomes involve intellectual skills, such as analytical and procedural skills, and require the learner to apply (i.e., transfer) knowledge and skills to new situations. Yet, at the same time, this conclusion is in line with the few studies that exist on the effect of feedback on performance in the

evaluative context. For example, Beckmann and Beckmann (2005) and Delgado and Prieto (2003) reported zero or negative effect of feedback in performance tasks. The authors of these studies most often argue that elaborated feedback is perhaps only efficient in typically learning contexts, i.e., repeated practice of tasks in order to learn a general rule or a set of principles that can be transferred to subsequent tasks and thus allow improvement. A similar conclusion was reached in the area of mathematics by Attali and van der Kleij (2017).

However, in the evaluative context (in performance tasks), students do not and indeed should not learn any underlying principles that make the solving of subsequent tasks easier. In fact, well-designed performance tests are deliberately constructed so that each test item is different and thus the possibility of bridging the gap between old and new knowledge (transfer of knowledge about underlying principles between items) is negligible (Delgado & Prieto, 2003). Moreover, studies such as Beckmann and Beckmann (2005) found that simple feedback can, in the case of failure, increase respondents' fears, which can then negatively interfere with the following-task-related information processing.

Hence, it could be conjectured that in the evaluative context (in performance tasks), feedback predominantly affects meta-task processes of the participants themselves.

Although experimental group D (where students could choose between different types of feedback; none, simple, elaborated) did not exhibit on average better performance in the game, we were interested in the performance of participants who actively chose/used the elaborated feedback option, compared to students who used it less often or not at all. Based on the analysis of feedback choices, we discovered that these students have better performance in the entire game set of tasks without simultaneously having a higher average IQ. Hence, it is evident that their willingness to invest the effort in applying/using feedback is, as in other studies (Bangert-Drowns et al., 1991), closely linked to motivation and interest in effectively monitoring their process of learning. Students who have interest, motivation and invest more effort in task solving may already have a larger repertoire of metacognitive skills applicable to various tasks in different domains, and at the same time know how and when to seek appropriate feedback to promote error-detection and correction (Timmers & Veldkamp, 2011).

We introduced learner-controlled feedback specifically to be able to examine respondents' strategies, i.e., describe the sequence of choices among feedback types. We found that elaborated feedback, as opposed to simple, was chosen more frequently as the difficulty of the tasks increased. This effect is in accordance with a number of other studies (Kluger & DeNisi, 1996; Kulhavy & Stock, 1989; Mory, 2004) and highlights the fact that this type of feedback supports conceptual learning. Clearly, it is not related to evaluative tasks alone but is a more general learning phenomenon.

In line with other studies (Timmers & Veldkamp, 2011), our research has shown that the choice of feedback itself fulfills a largely corrective function, occurring more often

after an incorrect answer. Incorrect answers thus provide an opportunity to improve potentially poor understanding of the problem (Mory, 2004). Moreover, by analyzing the sequences of feedback type choices, we identified the following most frequently occurring order: First, choice B (verification correctly/incorrectly), then, if the solution is incorrect, choice C (elaborated feedback). We believe that this finding underlines the corrective aspect of elaborated feedback in particular (Mory, 2004).

### Limitations

The main limitation of this study is its (relatively small) sample size. Although we did expect relatively large effect sizes, this expectation did not come to fruition and it might be the case that small effects of feedback on task performance remained undetected by our study design. Replicating the study with a substantially larger sample is, according to us, key.

Secondly, one of our main assumptions was that the ability to learn and improve from feedback is primarily a function of fluid intelligence. Thus, we have balanced the experimental groups according to the students' intelligence as measured by the CFT 20-R. On the other hand, it is safe to assume that the ability to learn and improve from feedback is affected by a variety of other variables we did not control for, such as metacognitive abilities or motivation. Although one could argue for the possible counterbalancing effect of random assignment to one of the experimental conditions, this is never completely accounted for unless the potential intervening variables are directly controlled for.

Lastly, it is important to note the reported reliability of the CFT 20-R is taken from the test's standardization manual and does not reflect reliability calculated on our sam-

ple. Unfortunately, we have not collected item-level data for the CFT 20-R and, as such, were unable to calculate reliability.

### Conclusions, Implications, and Further Research

The literature on feedback suggests that there are complex relationships between feedback intervention, task characteristics, learning context, and characteristics of the learner, which affect the magnitude of feedback effects (Shute, 2008). Based on the results reported, we conclude that the provision of any (even elaborated) feedback in performance tasks (as opposed to intervention tasks) is generally not effective, since solving previous tasks does not provide information relevant to solving following tasks. These findings specifically concern the evaluative context of higher-order learning and as such differ from most high-order learning outcome intervention studies. These studies usually agree on the positive effect of elaborated feedback (Van der Kleij et al., 2015). To verify the generalizability of these effects, it would be desirable to conduct a similar survey with different logical tasks and populations (for example, with intellectually gifted students).

Detailed tracing of learners' behavior in individual feedback conditions allowed us to describe some mechanisms, perhaps more generally valid, by which elaborated feedback affects the learning process. These mechanisms appear not to be highly dependent on the evaluative context, and as such are more in line with research on feedback in educational studies. However, future research should concentrate also on specific traces of progress through the game (detailed tracing of in-game activity) while simultaneously measuring learners' metacognitive, motivational, and personality traits, which could provide additional valuable information on the regulatory,

metacognitive and motivational mechanisms in learning (Azevedo & Alevén, 2013). These traces might perhaps prove more valuable for understanding the students' learning process under individual feedback conditions than their overall performance in task solving.

### Acknowledgement

This work was supported by The Czech Science Foundation – GACR [GA17-14715S] and by the project MUNI/A/1548/2021.

### Authors' ORCID

Šárka Portešová  
<https://orcid.org/0000-0002-8107-5981>  
 Michal Jabůrek  
<https://orcid.org/0000-0001-5994-0441>  
 Adam Ťápal  
<https://orcid.org/0000-0003-2315-4590>  
 Ondřej Straka  
<https://orcid.org/0000-0001-5631-8825>

### References

- Alevén, V., McLaren, B. M., Roll, I., & Koedinger, K. R. (2006). Toward computer-based tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16*, 101–130.
- Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education, 110*, 154–169. <https://doi.org/10.1016/j.compedu.2017.03.012>
- Azevedo, R., & Alevén, V. (2013). *International Handbook of Metacognition and Learning Technologies*. Springer Publishing Company, Incorporated.
- Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111–127. <https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT>
- Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect

- of feedback in test-like events. *Review of Educational Research*, 61(2), 218–238. <https://doi.org/10.3102/00346543061002213>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beckmann, J., & Beckmann, N. (2005). Effects of feedback on performance and response latencies in untimed reasoning tests. *Psychology Science*, 47, 262–278.
- Beckmann, N., Beckmann, J. F., & Elliott, J. G. (2009). Self-confidence and performance goal orientation interactively predict performance in a reasoning test with accuracy feedback. *Learning and Individual Differences*, 19(2), 277–282. <https://dx.doi.org/10.1016/j.lindif.2008.09.008>
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245–281. <http://doi.org/10.3102/00346543065003245>
- Delgado, A. R., & Prieto, G. (2003). The effect of item feedback on multiple choice test responses. *British Journal of Psychology*, 94, 73–85. <https://doi.org/10.1348/000712603762842110>
- Fajmonová, V., Hönigová, S., Urbánek, T., & Širůček, J. (2015). *CFT 20-R – Cattellův test fluidní inteligence* [CFT 20-R – Cattell's Fluid Intelligence Test] [Measurement instrument]. Praha: Hogrefe-Testcentrum.
- Fuchs, D., & Fuchs, L. S. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *Review of Educational Research*, 56(2), 243–262. <https://doi.org/10.2307/1170377>
- Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124(1), 75–111. <https://doi.org/10.1037/0033-2909.124.1.75>
- Guthke, J., & Beckmann, J. F. (2000). The learning test concept and its application in practice. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic Assessment: Prevailing models and applications* (pp. 17–69). Oxford, UK: Elsevier.
- Guthke, J., & Beckmann, J. F. (2003). Dynamic assessment with diagnostic problems. In R. J. Sternberg, J. Lautrey, & T. I. Lubart (Eds.), *Models of intelligence: International perspectives* (pp. 227–242). Washington, DC: American Psychological Association.
- Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment*, 12(1), 1. <https://doi.org/10.1027/1015-5759.12.1.1>
- Hattie, J. (1999). *Influences on student learning*. Unpublished inaugural lecture presented at the University of Auckland, New Zealand.
- Hattie, J., & Gan, M. (2011). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York: Routledge.
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hattie, J., & Zierer, K. (2019). *Visible learning insights*. Routledge, NY.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.
- Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certainty. *Educational Psychology Review*, 1, 279–308. <https://doi.org/10.1007/BF01320096>
- Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, 58(1), 79–97. <https://doi.org/10.3102/00346543058001079>
- Lipnevich, A. A., Berg, D., & Smith, J. K. (2016). Toward a model of student response to feedback. In G. T. L. Brown & L. Harris (Eds.), *Human Factors and Social Conditions in Assessment* (pp. 169–185). Routledge.
- McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Woodcock Johnson IV: Technical manual*. Riverside Publishing.
- Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, 32(1-2), 99–113. <https://doi.org/10.1023/B:TRUC.0000021811.66966.1d>
- Mory, E. H. (2004). Feedback research revisited. *Handbook of Research on Educational Communications*, 2, 745–784. <https://doi.org/10.1007/s00127-009-0052-2>

- Panero, M., & Aldon, G. (2016). How teachers evolve their formative assessment practices when digital tools are involved in the classroom. *Digital Experiences in Mathematics Education*, 2(1), 70–86. <https://doi.org/10.1007/s40751-016-0012-x>
- Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment: An empirical review and recommendations. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of Psychology: Assessment Psychology* (2nd edition) (pp. 82–113). Wiley.
- Rocklin, T. R., O'Donnell, A. M., & Holst, P. M. (1995). Effects and underlying mechanisms of self-adapted testing. *Journal of Educational Psychology*, 87(1), 103–116. <http://dx.doi.org/10.1037/0022-0663.87.1.103>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (4th edition) (pp. 73–163). The Guilford Press.
- Smith, P. L., & Ragan, T. J. (2005). *Instructional design* (3rd ed.). Wiley.
- Sternberg, R., & Grigorenko, E. (2002). *Dynamic testing*. Cambridge, UK: Cambridge University Press.
- Timmers, C., & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education*, 56(3), 923–930. <https://doi.org/10.1016/j.compedu.2010.11.007>
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4), 475–511. <https://doi.org/10.3102/0034654314564881>
- Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, 58(1), 263–272. <https://doi.org/10.1016/j.compedu.2011.07.020>
- Veerbeek, J., Hessels, M. G. P., Vogelaar, S., & Resing, W. C. M. (2017). Pretest versus no pretest: An investigation into the problem-solving processes in a dynamic testing context. *Journal of Cognitive Education and Psychology*, 16(3), 260–280. <https://doi.org/10.1891/1945-8959.16.3.260>
- Wang, Z., Gong, S., Xu, S., & Hu, X. (2019). Elaborated feedback and learning: Examining cognitive and motivational influences. *Computers & Education*, 136(April), 130–140. <https://doi.org/10.1016/j.compedu.2019.04.003>
- Wechsler, D. (2008). *Wechsler Adult Intelligence Scale (4th ed.)* [Measurement instrument]. San Antonio, TX: Pearson.
- Wechsler, D. (2014). *Wechsler Intelligence Scale for Children (5th ed.)* [Measurement instrument]. San Antonio, TX: NCS Pearson
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>