Title Page

1) Full title

The Role of Cultural Background in the Personalization Principle: Five Experiments with Czech Learners

2) Authors

Cyril Brom
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
brom@ksvi.mff.cuni.cz

Tereza Hannemann
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
Faculty of Arts, Charles University in Prague
U kříže 8, 15800 Prague 5, Czech Republic
tersel@seznam.cz

Tereza Stárková
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
tereza.starek@gmail.com

Edita Bromová
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
edita@email.com

Filip Děchtěrenko
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
filip.dechterenko@gmail.com

3) Corresponding Author

Cyril Brom

Faculty of Mathematics and Physics, Charles University in Prague, Room 312, Malostranske Namesti 25, Prague, 11800, Czech Republic.

E-mail: brom@ksvi.mff.cuni.cz

Tel: (420) 221 914 216; Fax: (420) 221 914 281

4) Acknowledgement

This study was primarily funded by Czech Grant Science Foundation (GA ČR) (Experiments 1, 2: Project nr. P407/12/P152; Experiments 3, 4, 5: Project nr. 15-14715S). T. Hannemann was also partly funded by a student research grant (GA UK nr. 1315) and F. Děchtěrenko was also partly funded by a student research grant (GA UK nr. 68413).

We thank research assistants who helped to conduct the experiments, namely: L. Kolek, V. Dobrovolný, M. Stejskalová, N. Bendová, and O. Smíšek. We thank Richard Mayer for discussing Experiments 1 and 2 with us and Kateřina Svobodová for discussing Experiments 3 and 4 with us. We also thank Slava Kalyuga for discussing certain aspects of the Cognitive Load Theory with us. We also thank Laboratory of Behavioral and Linguistic Studies in Prague, where part of the study was conducted.

The human data were collected with APA ethical principles in mind.

The Role of Cultural Background in the Personalization Principle: Five Experiments with Czech Learners

Abstract

Composing instructional texts in multimedia learning materials in a conversational style rather than a formal style can facilitate learning. We investigated whether a specific language/cultural background could present a boundary condition for this effect. In four experiments with a Czech sample ($N = 278$), we replicated a seminal experiment conducted on a US sample (with a short animation on the topic of lightning formation), which demonstrated a large effect size in favor of the instructional texts in the conversational style. In our four experiments, we varied between two types of audiences (a college and a high school audience) and two types of short animations (the original one and a complementary one). Instructional texts in a conversational style brought no overall advantage for the Czech audience ($\eta_p^2 = 0.00$; the high school audience: $d = 0.48$, 0.22; the college audience: $d = -0.45$, -0.04). Twenty-nine percent of participants who received the conversational instructional texts expressed explicit reservations regarding the style of the language. In the fifth supplementary experiment, Czech participants ($N = 138$) had to rate preferences on computer tutor's printed statements. Direct rather than polite statements were preferred. Limited benefits of conversational/polite instructional texts for Czech learners are probably related to the generally more formal approach to education in the Czech Republic compared to the US schooling system. We also failed to find a link between several affective variables and learning outcomes; with the exception of a relationship of generalized positive affect, levels of flow and perceived difficulty to some of the learning outcomes.

Keywords: personalization principle, multimedia learning, mental model, animation, positive affect

Article

1. Introduction

How should we design multimedia learning materials to best serve learning? One answer is to write instructional texts included in these materials using a language style that can be easily processed by the target group of learners and/or that is more appealing to them. One particular way how to do that is to write the instructional texts using a conversational rather than a formal style: this is the so-called *personalization principle* (Mayer, 2009). The major techniques for making instructional texts conversational include changing the text from a third person form of address to first/second person, adding statements directed at the learner, adding personal views or replacing direct statements with polite statements (Ginns, Martin, & Marsh, 2013; Mayer, 2009). Investigating the robustness of this principle beyond the context where it was established is within the scope of the present study.

The empirical base behind the personalization principle lies in studies that have demonstrated *the personalization effect* – superiority of learning from instructional texts in a conversational style (compared to a formal style). This effect was repeatedly demonstrated for English treatments up to 35 minutes long (meta-analyzed in Ginns, Martin, & Marsh, 2013; medium to very large effect sizes, as measured by transfer tests). The treatments included, for example, static multimedia presentations or instructional animations. Fewer studies have been performed with longer treatments (i.e., Clarebout & Elen, 2007; Doolittle, 2010; McLaren, Lim, Gagnon, Yaron, & Koedinger, 2006; McLaren, DeLeeuw, & Mayer, 2011a; McLaren, DeLeeuw, & Mayer, 2011b; Son & Goldstone, 2009; Tze Wei, Su-Mae, & Nuo Wi, 2014; Wang & Johnson, 2008; Brom et al., 2014a) and in non-English settings (i.e., Bol, van Weert, de Haes, Loos, & Smets, 2015; Clarebout & Elen, 2007; Dutke, Grefe, & Leopold, 2016; Kartal, 2007; Kartal, 2010; Kurt, 2011; Reichelt, Kämmer, Niegemann, & Zander, 2014; Rey & Steib, 2013; Schneider, Nebel, Pradel, & Rey, 2015; Schworm & Stiller, 2012; Stiller & Jedlicka, 2010; Tze Wei, Su-Mae, & Nuo Wi, 2014; Brom et al., 2014a). The findings from the studies performed with longer treatments and/or in non-English settings were mixed. Meanwhile, there is a demand for investigation of the personalization principle's boundary conditions (Mayer, 2009, p. 255); including in non-English settings and as concerns length of instruction time (Ginns, Martin, & Marsh, 2013, p. 467).

For the Czech language, we showed that in a 2-3 hour long educational simulation on the topic of brewing beer, college participants receiving formal instructional texts performed no worse than participants with instructional texts in the conversational style (Brom et al., 2014a). Was that result caused by longer exposure or by a different language/cultural context? To clarify this issue, and therefore to support one of the following two ideas – that the personalization principle's boundary condition is the length of exposure or that it is the participants' native language/cultural background – we performed (and present here) four experiments with interventions roughly five minutes in length in the Czech context. In Experiments 1 (with university students) and 2 (with high school students), we closely replicated, in the Czech context, the seminal experiment on the personalization effect conducted by Moreno and Mayer (2000; Exp. 2) with an animation on lighting formation with on-screen texts. To investigate the robustness of our findings, we then replicated the original experiment once again with a different animation (on how a biological wastewater treatment plant functions), which was comparable in terms of length and complexity to the lightning formation animation (Experiment 3: university students; Experiment 4: high school students). In all four experiments, participants received either a personalized or a non-personalized version of the respective animation. Taken together, the four experiments have the power to show how robust, or brittle, the personalization principle is in up-to-35-minute-long instructional materials in certain non-English contexts.

Because the mediating role of interest and motivational variables on learning outcomes in personalized vs. formal treatments has not yet been adequately studied (Ginns, Martin, & Marsh, 2013), we also administered questions on self-perceived learning, usefulness of the materials, interest, motivation, friendliness, perceived difficulty, and (in Experiments 3, 4) generalized affect and flow state. We also explicitly asked participants what they thought about the texts they read during the learning experience.

It is argued throughout this article that the Czech schooling system is more formal compared to the US one. This can be exemplified through its drawing on centrally-organized classrooms with teachers having relatively dominant role in student-teacher interactions and there being a higher focus on teaching subject-matter knowledge (primarily through transfer and reproduction of factual knowledge) and a lower emphasis on learners' personal development and the nurturing of student self-esteem and autonomy (Palečková, 1999; Stolinská, 2012; Straková & Simonová, 2013; see also Hirsch, 1997; Perry, 2005; Stigler & Perry, 1988; Průcha, 2015, ch. 11). Furthermore, in the Czech schooling system, students tend to be graded publicly, i.e., in front of the whole class, at the start of the lessons (U.S. Department of Education, 2003), and errors made are discussed publicly; whereas, evaluation tends to be more "invisible" in US schools and praise is a more typical type of public evaluation (Stigler & Perry, 1988).

Consequently, students who have grown up in such a formal schooling system probably have different expectations regarding how they would be addressed in educational contexts. Therefore, instructional texts written in a conversational style may have a different impact in Czech contexts. To support this argument, we also researched whether Czech learners' preferences for different types of educational statements differ from those of US learners. In Experiment 5, Czech participants had to rate their preferences for a computerized tutor's written statements. These statements were already evaluated by US learners in terms of politeness (Mayer, Johnson, Shaw, & Sandhu, 2006). Would Czech learners prefer educational statements considered polite by the US audience or would they prefer the opposite?

2. The Personalization Principle

The personalization principle is one of the design principles of multimedia learning (Mayer, 2009). Multimedia learning materials are defined, in this context, as materials combining pictures with words (narrated verbally or presented as text).

The personalization principle posits that people learn better when instructional texts included in multimedia materials are given in a conversational rather than a formal (or neutral) style (Mayer, 2009). There are more ways how a conversational style can be operationalized. Ginns and colleagues' meta-analysis (2013) coded two major forms: personalization and politeness. The personalization of instructional texts typically means changing a third person form of address to first/second person (e.g., "In this animation, one can learn how the human eye works." → "In this animation, you will learn how the human eye works." or "In this animation, I will tell you how your eye works."), or adding statements directed at the learner ("Now look at the retina.", "Congratulations, you now know what the retina is."). Politeness means using polite conversational statements (e.g., "Calculate the result." → "Let's calculate the result.") (see Mayer et al., 2006 for more on politeness). Sometimes, a personal view is also expressed (e.g., "This is how it works..." → "In my opinion, this is how it works...").

There is a strong empirical support for this principle in the context of up-to-35-minute-long English interventions, but evidence beyond this context is inconsistent and incomplete (Ginns, Martin, & Marsh, 2013). There are several complementary explanations for why this principle may work, but evidence supporting these explanations is scarce (as described in Section 4). What levels of personalization are best for facilitating learning is also still up for debate (Kartal, 2010; Mayer, Fennell, Farmer, & Campbell, 2004; Schworm & Stiller, 2012). Too much personalization may not be always beneficial (Mayer, 2009; p. 252).

2.1 Studies in non-English settings

Studies in non-English settings showed positive learning outcomes[1] in favor of instructional messages in a conversational style in Turkish (Kartal, 2010) and in German (Rey & Steib, 2013;

[1] When speaking about learning outcomes, we refer to transfer tests or their equivalent, because these tests are considered to be appropriate measures of deep conceptual learning (Mayer, 2009), as opposed to "surface" learning, which is often measured by retention tests in this context. Generally, retention measures

Schneider et al., 2015; Schworm & Stiller, 2012; Dutke, Grefe, & Leopold, 2016); but they also showed no difference in Turkish (Kartal, 2007), German (Reichelt et al., 2014), Czech (Brom et al., 2014a), Dutch (Bol et al., 2015), and in the context of the Malaysian educational system[2] (Tze Wei et al. 2014); and negative outcomes in Flemish (Clarebout & Elen, 2007, Exp. 1). One study reported mixed results where the overall null results masked a moderating effect of prior knowledge: a German study by Stiller and Jedlicka (2010). Unless stated otherwise in the following text, they all used a classical way of operationalization of conversational style, as described above.

In particular, for the German language, support for the personalization principle has been found four times, null results have been reported one time and mixed results also one time. Rey and Steib (2013) recruited 10 to 14-year-old Austrian students to learn about computer network topologies from a narrated 6-minute-long animation. The results revealed a personalization effect for both standard German and Austrian dialects. Dutke and colleagues (2016) found a personalization effect in a study with German high school students of roughly 16 years of age. They learned, for 15 minutes, about the anatomy and the functionality of the human eye from a textual material with one picture. Schworm and Stiller (2012) engaged German university students in learning about the circulatory system from a 12-minute-long presentation combining narration and static graphics. Two conversational styles were used: with weak and strong forms of personalization (they differed in the number of comments/sentences directed at the learner: 3 vs. 35). The strong form was better than the weak form, which was better than the non-conversational control; though significant difference was achieved only when the two groups with conversational treatments were tested together against the control group (the sample size was only around 20 per cell). Finally, in two experiments with 12 to 16-year-old German adolescents, who studied from 5-minute-long materials on photosynthesis, Schneider and colleagues (2015) found a personalization effect both for narration-only as well as text-only materials. These two experiments operationalized a conversational style as youth slang (i.e., language more familiar to the target audience).

However, in a study by Reichelt and colleagues (2014), German participants learnt over 25-30 minutes about principles of the gestalt laws from a computer-based learning environment combining texts and static pictures. The sample included college students and adults participating in continuing education. The authors found no personalization effect for any of the audiences (only the difference in retention test scores was significant). Stiller and Jedlicka (2010) engaged participants in learning about the human eye, for around 13 minutes, from an interactive computerized text-and-picture material. They found that, as concerns transfer, a conversational style improved learning for low prior

show smaller benefits from personalization compared to transfer measures (Ginns et al., 2013). If a study does not include transfer tests or an equivalent, we report on the knowledge measures available.

[2] The participants' native languages were primarily Malay and Mandarin and the expository texts were in English (since this is typical in the Malaysian educational system); (Liew Tze Wei, personal communication: email dating from 9 June 2015).

knowledge German high school learners (around 16-years-old) but hindered learning for high prior knowledge learners (but the sample size was less than 20 per cell). This is a manifestation of the so-called expertise reversal effect (Kalyuga, 2007), i.e., an interaction between learners' prior knowledge and treatment types.

Two studies were conducted with Turkish college audiences learning about stellar evolution and death from a brief computerized material (combining texts, pictures and animations) for around 20 minutes (Kartal, 2007; Kartal, 2010). Turkish distinguishes between two forms of second person singular pronouns, a formal one and an informal one, and these two forms were used as the basis for two types of personalization. No advantage has been reported regarding the formal personalization (Kartal, 2010) and beneficial effects of the informal personalization have been found in (Kartal, 2010) but not in (Kartal, 2007).

In our own study (Brom et al., 2014a), we have found no benefits of a conversational style for a Czech college audience learning how to brew beer from a 2-3 hour long educational simulation with on-screen instructions on the topic of brewing beer. To operationalize the conversational style, we used a classical form of personalization and framed the learning experience within a story about a family brewery. Bol and colleagues (2015) recruited Dutch adult participants (18 – 85 years old) to study (outside a lab) about a lung cancer treatment using a brief, web-based material. Either a text version or an audiovisual version with an actor introducing the content was used. No overall effect of the personalization has been found (but these authors used recall measures only); however, the audiovisual personalization group performed the best. Tze Wei and colleagues (2014) found no overall effect of the personalization when Malaysian college learners studied the basics of C++ for 40 minutes from a picture-and-text presentation accompanied by a narration. However, the materials were not in the native language of the majority of participants (see Footnote (2)), which could have influenced the results. Finally, in the study by Clarebout and Elen (2007; Exp. 1), 14 to 15-year-old Flemish students studied, over a period of 50 minutes, about ecology using a computer learning environment with an on-screen pedagogical agent. The group with non-personalized instructional texts outperformed the group with personalized texts.

2.2 Studies with Longer Treatments

Studies with treatments longer than 35 minutes showed no overall effect for the conversational style (Doolittle, 2010; McLaren et al., 2006; McLaren, DeLeeuw, & Mayer, 2011a; Tze Wei et al., 2014; Wang & Johnson, 2008; Brom et al., 2014a) or negative effect (Clarebout & Elen, 2007, Exp. 1; Son & Goldston, 2009, Exp. 2). One study with mixed results again showed a moderating effect of prior knowledge (McLaren, DeLeeuw, & Mayer, 2011b). Apart from the three studies already mentioned in the previous section (i.e., Brom et al., 2014a, Clarebout & Elen, 2007; Tze Wei et al., 2014), all of them were conducted with English-speaking samples.

Doolittle (2010) failed to demonstrate the effect in a 2.5 hour long, narrated multimedia tutorial on teaching the high-level skill of critical historical reasoning to college learners. Apart from traditional changes to grammatical constructions, the conversational style also included personal comments to foster engagement (e.g., "Imagine that you are reading a letter from your great grandfather."). Son and Goldston (2009; Exp. 2) even reported a negative effect of personalization on learning outcomes. This study engaged university students in learning about signal-detection theory from a computerized text-and-picture tutorial on signal-detection theory embedded in the context of a doctor trying to diagnose patients with leukemia. The personalization specifically biased learners to adopt a first person rather than third person perspective.

The effect was also not demonstrated in the study by McLaren and colleagues (2006), who used a roughly hour long, web-based, tutorial on stoichiometry with a college audience as a sample. However, the authors also reported that a substantial portion of the sample might have been non-native English speakers, who thus "missed the nuances of personalized English" (p. 326). Two different studies by McLaren's team (2011a; 2011b) employed a similar intervention, but politeness was used for operationalization of the conversational style. The former of these two studies recruited high school learners in a classroom setting and failed to find positive effects of the polite conversational style on learning outcomes. The latter study with a college audience in a lab setting showed a positive effect on learning outcomes for low prior knowledge learners, but a negative effect for high prior knowledge learners, i.e., the expertise-reversal effect. Wang and Johnson (2008) also tested the effects of a polite conversational style, but on foreign language learning (within a military set-up). Adult volunteers of 21 to 63 years of age interfaced with a 2 hour long intelligent tutoring system. No overall effects of the polite conversational style on learning outcomes were found.

2.3 Summary

Generally, studies in non-English settings showed mixed results. Studies with longer treatments tended to find null or negative results. Especially the short, non-English-context studies seemed to employ a classical form of operationalization of the conversational style (except for the youth slang used in Schneider et al., 2015). Therefore, findings from non-English contexts with the null or negative results cannot likely be explained by usage of "over-personalization", which is not always supposed to be beneficial (Mayer, 2009, p. 252). However, the null results of Bol and colleagues (2015) might have been caused by usage of retention tests only. The negative results of the study by Clarebout and Elen (2007) might have been due to longer exposure. The null results of Tze Wei and colleagues (2014) might have been due to longer exposure or the second language factor. The null results of our own study (Brom et al., 2014a) might, again, have been due to longer exposure or due to the usage of narrative as part of the personalization (i.e., "over-personalization"). Also, for the German language, there are more positive than null/negative findings.

Therefore, based on the research findings so far, it is impossible to conclude that a different language/cultural context can be one of the explanations for differences between findings from studies conducted in English and other language contexts. This warrants conducting additional studies looking at the language/cultural context as a possible boundary condition of the personalization effect; provided these studies are designed so that at least some of the alternative explanations can be excluded. Specifically, the present study is designed so that explanations based on long exposure (and also missing transfer tests and the second language factor) can be excluded. Findings from the present study, together with the findings from our previous study with a long treatment (Brom et al., 2014a), thus have the potential to support one of the following ideas: that the personalization principle's boundary condition is the length of exposure (i.e., if we now find the personalization effect with a short treatment) or that it is certain language/cultural backgrounds (i.e., if we now fail to find such an effect).

3. Cognitive-Affective Theory of Learning with Media

From a theoretical perspective, the design principles of multimedia learning are grounded in the Cognitive Theory of Multimedia Learning (Mayer, 2009), which is primarily cognitively focused. The principles are also based on its extension: the Cognitive-Affective Theory of Learning with Media (CATLM; Moreno, 2005), which adds affective and meta-cognitive factors. We use the CATLM as the explanatory framework for this work. These theories also run parallel to and draw on the Cognitive Load Theory (Sweller, Ayres, & Kalyuga, 2011).

Capitalizing on Baddeley's classical memory model (Baddeley, Eysenck, & Anderson, 2009) and Dual Coding Theory (Clark & Paivio, 1991), the CATLM postulates that multimedia information is processed by learners through two separate cognitive channels (verbal and visual), organized in their working memory into coherent mental models and integrated with prior knowledge "stored" in their long-term memory. The efficiency of this process depends on various aspects, two of which are important for our present purposes.

First, the learning task and materials impose cognitive load on the working memory of a learner, who then must allocate cognitive resources to deal with this load. According to the recent change in the concept of cognitive load (Kalyuga, 2011; see also Sweller, 1994; de Jong, 2010), the total cognitive load consists of two additive types of load: intrinsic and extraneous. Intrinsic cognitive load is imposed by the complexity of the learning task with respect to the learner's prior knowledge (what is complex for a novice may not be so complex for an expert). This type of load is "useful" in that it is essential for comprehending the learning material. Dealing with this load results in learning: with no intrinsic load, there is no learning. On the other hand, intrinsic cognitive load should not overwhelm available cognitive resources: that would impede learning. Extraneous cognitive load is caused by the redundant processing of suboptimally designed features of instructional materials (that nevertheless must be processed so that the core message can be understood). Accommodating this load means depleting cognitive resources that could otherwise be devoted to dealing with intrinsic load. Extraneous load is therefore a "bad" load and it should be minimized (Kalyuga, 2011; Mayer, 2009).[3]

The second aspect on which the efficiency of processing the learning message depends is the learner's willingness to devote cognitive resources to deal with the two types of cognitive load. In the context of the CATLM, the usage of the term "the level of active cognitive participation" (rather than devoted or allocated cognitive resources) is also used. Here, we will use these two terms interchangeably. For instance, the instructional materials may be boring: failing to make use of all required cognitive capacity for learning. Allocating fewer resources than needed to accommodate both types of load results in suboptimal learning (Mayer, 2009; Moreno, 2005; Moreno & Mayer, 2007). Engaging and motivating instructional materials may increase the level of the learner's active cognitive participation. However, at the same time, making the materials more engaging may increase extraneous cognitive load, because the learner must process the instructional features that make the materials engaging. Thus, there is a trade-off in creating motivational instructional messages (see, e.g., Um, Plass, Homer, & Hayward, 2012 for more on this point).

In the next section, we will link explanations underlying the personalization principle presented in the literature to the changes in cognitive load and to resources actually allocated for accommodating cognitive load.

4. Explanations Underlying the Personalization Effect

Why the personalization principle works remains elusive (cf. Reichelt et al., 2014; Rey & Steib, 2013). There are several complementary explanations, but limited evidence for their direct support. This is mainly because not enough studies have researched the impact of postulated explanatory variables on learning outcomes (see, e.g., Ginns, Martin, & Marsh, 2013). Moreover, it is hard to operationalize some of the key theoretical concepts and their measurement is problematic (as detailed below). Most explanations are thus made post hoc. We note that the present work seeks for some improvements, especially as concerns measurement of "positive affect" induced by the treatment. However, its primary objective is practical: to investigate the personalization principle's boundary condition, not to improve its theoretical underpinning.

First, Mayer (2009) argues, based on Reeves and Nass (1996) and Grice (1975), that conversational styles present certain "social cues" to the learner that prime the activation of a social

[3] The notion of cognitive load is inherently connected to Cognitive Load Theory (see Sweller, 1994; Sweller, Ayres, & Kalyuga, 2011; Plass, Moreno, & Brünken, 2010; Kalyuga, 2011) rather than to the Cognitive Theory of Multimedia Learning or CATLM. Nevertheless, the intrinsic and extraneous cognitive load are directly mapped onto the Cognitive Theory of Multimedia Learning by Mayer (2009, pp. 79-89) and Kalyuga (2011, pp. 5-8).

response in him/her. This might be, for example, “the commitment to try to make sense out of what the speaker is saying” (Mayer 2009; pp. 247-8). Thus, in terms of the CATLM, because of social cues in the instructional messages, the learner is engaged more in active cognitive processing and therefore learns better (see Figure 1). This idea was supported by Moreno and Mayer (2004) and Kartal (2010), but not by Wang et al. (2008). “Social presence” was operationalized primarily by the material’s “friendliness” and/or “helpfulness” (see Moreno & Mayer, 2004). This explanation is also typically employed when a polite conversational style is used (e.g., McLaren, DeLeeuw, and Mayer, 2011b).

-- Insert Fig. 1 around here --

The second major explanation is based upon a so-called self-reference effect (Rogers, Kuiper, & Kirker, 1977). This effect is linked to mnemonic aspects of the self-structure, cognitive-affective entity, which is believed to serve as a powerful organizing and elaborative device for processing and encoding self-related information (Symons & Johnson, 1997). Superior retention for self-related information (compared to other types of information) was demonstrated for incidental recall of a list of words (Symons & Johnson, 1997) and beyond (see Moreno & Mayer, 2000, p. 725), and it was also offered as a possible explanation for the personalization effect (e.g., Moreno & Mayer, 2000; Reichelt et al., 2014). In the terms of the CATLM, this explanation pertains to reducing cognitive load. We are less inclined toward this explanation, because in our opinion texts in a conversational style in personalization principle studies activate the self-structure to a varying degree. Also, as pointed out by Rey and Steib (2013), multiple studies support the personalization effect only for transfer performance (but not for retention; cf. Ginns, Martin, & Marsh, 2013) and it is not clear how this data pattern can be explained by the self-reference effect (while it can be explained by the social cues hypothesis; see Mayer, 2009). Also, in one experiment using a medical signal-detection scenario, participants were forced to adopt a self-perspective (i.e., a first person perspective) via a personalization of the instructional text, and this turned out to be detrimental to learning (Son & Goldstone, 2009, Exp. 2). The authors explained the findings in the following way: the lesson taught learners abstract knowledge (on signal detection theory), but the first person perspective probably facilitated the recall of medical-specific knowledge conflicting with abstract knowledge learned from the lesson. Consequently, participants sometimes answered test questions using their medical-specific knowledge rather than abstract knowledge.

Third, familiarity with the style of language used for instructional texts, i.e., whether learners encounter this style of language often, can also help to explain the personalization effect. Schneider et al. (2015) postulated familiarity as an antecedent to social cuing. This would mean, in the terms of the CATLM, that familiarity would help increase cognitive capacity actually devoted to dealing with cognitive load (i.e., via social cuing). Moreno and Mayer (2000) argued differently: if learners are more familiar with being addressed using a personal style, processing such materials may be easier for them. On the level of the theory, this would mean that familiar elements in the instructional materials may help to reduce cognitive load of learners. Both of these explanations predict that familiarity with the language style of instructional text would facilitate learning.

Finally, in agreement with the CATLM, interest, motivation, perceived difficulty, generalized positive affect or flow were sometimes explicitly posited as potential mediators between social presence and increased active cognitive processing (e.g., Schworm & Stiller, 2012; Brom et al., 2014a), or between using the self-structure as a reference point and capacity allocated for active cognitive processing (Mayer et al., 2004), or for both cases (Stiller & Jedlicka, 2010). For the sake of simplicity, we denote all of these variables as *affective* variables (even though some of the underlying processes are on the border between affect, motivation, and attention). Demonstration of positive effects of using conversational style on these potential "affective" mediators (or a fraction of these mediators) has been met with mixed success: it was demonstrated by Moreno and Mayer (2004) and Kartal (2010) but not found by Mayer et al. (2004), Kartal (2007) and Wang et al. (2008) (see also Ginns, Martin, & Marsh, 2013; p. 461). Consequently, it was suggested that some of the measures of affective variables, especially those assessing a construct with one or two items only, may be insensitive (e.g., Mayer et al., 2004; Wang et al., 2008). Using more complex instruments proved useful in (Brom et al., 2014a), where the positive effect of higher generalized positive affect and flow on learning outcomes was demonstrated; however, no influence of personalization on learning outcomes and on positive affect/flow was found in that study. This means that the existence of the second part of the hypothetical link "personalization → an increase in positive affect → learning" was supported for the 2-3 hour long educational simulation (Brom et al., 2014a), but the existence of the link's first part was not. On the other hand, Reichelt et al. (2014), who also used complex measures, showed that initial, in situ, and post hoc motivation were higher for personalized instructional texts in 25-30-minute-long gestalt laws learning material, but no effect of personalization on learning outcomes was demonstrated. This means that the study found support for the link's first part, but not for its second part. Unfortunately, it is especially rare to see reported correlations between cognitive and affective measures; making it difficult to find retrospectively support for the link's second part.[4]

Considering all points together (see Figure 1), and aligned with the CATLM framework, familiarity with the materials featuring instructional texts in a conversational style, and perhaps also

[4] There is also an ongoing "emotions-as-facilitator vs. emotions-as-distractor" debate; linked to the trade-off between allocating more cognitive resources to accommodate cognitive load due to positive impact of emotions (and related qualities) on learning and increasing extraneous cognitive load due to comprehending elements of learning materials that promote positive emotions (e.g., Um, Plass, Homer, & Hayward, 2012; Mayer, 2014). This issue is out of scope of the present work and we thus do not elaborate on it further for the sake of conciseness.

using self as a reference point, can decrease cognitive load (improving learning). Social cuing and again (perhaps) the self-reference aspect in the personalized instructional texts can help activate the learner, who will then be willing to allocate more cognitive resources for accommodating cognitive load (also improving learning). As concerns increases or decreases of allocated cognitive resources, certain affective variables can play a mediating role. The fading away of initially increased interest/motivation could also explain the lack of personalization effect for long interventions. However, so far, evidence for all of these ideas has been limited and mixed. At least partly to blame are some instruments with questionable sensitivity and/or validity used in the past (in the context of personalization principle studies) to investigate the constructs necessitated by the theoretical model on Figure 1, such as "social presence", "motivation" or components of "cognitive load" (see e.g., Mayer et al., 2004; Wang et al., 2008; de Jong, 2010).

Instructional texts in the conversational style can, in theory, also have negative effects. So far, this topic has not been discussed much in the literature (for example, see Mayer, 2009, p. 252). If learners are unfamiliar with the conversational style in learning contexts, this may cause distraction (in the cognitive domain) and/or aversion (in the affective domain). The former may result in increasing cognitive load. The latter may have unwelcome consequences in reducing allocated cognitive resources. These possible effects are also depicted on Figure 1.

As concerns personalization in non-US contexts, the different findings can be explained in multiple ways (after one excludes alternative explanations such as a long exposure). Learners in some cultural contexts may just not like the personalization (for instance, because they are not used to speech in the conversational style in textbooks/the school environment and therefore think its usage is childish), which can reduce the learners' active cognitive participation and thereby impede learning. Some learners may just not care about whether the instructional texts are formal or conversational, as long as they are easy-to-understand. An explanation pertaining to changes to cognitive load would be that learners with certain cultural backgrounds may be more familiar with personalized forms of address (compared to the formal ones) and therefore process them more readily. However, this can be the other way around for learners with a different background: lower familiarity with personalized forms of address can increase cognitive load, as already suggested. More than one process can also interact; for instance, negative influence of the increase in cognitive load can be offset by the potential "affective" benefits of social cuing or self-referencing, leading to overall null results.

Specifically, Czech learners are not used to a conversational and overly personal style of expository texts in the context of the institutionalized schooling system, which is traditionally more formal than the US schooling system.[5] Thus, on the one hand, students may welcome this "new" form

[5] A curricular school reform has been gradually initiated in the Czech Republic during the past decade. The influence of the reform on our samples was probably limited, as discussed in Section 8.2 in further detail.

of address; on the other hand, as suggested in the previous paragraph, they may consider its usage childish and/or it may be distractive for them in the educational context. There are thus reasons to expect that the results of this study may differ from the studies with US samples.

In this experiment, we primarily aim to investigate the *outcomes* of learning with personalized vs. formal instructional texts in the Czech context; not the underlying principles. Nevertheless, we adopt the positive-affect-as-mediator hypothesis. Therefore, as this study's secondary goal, we also investigate if differences in learning outcomes are related to differences in interest and related affective variables.

5. Research Questions and Design

This study's primary goal is to investigate if the personalization principle's boundary condition is a long treatment or a different language/cultural context. In the Czech context, we failed to replicate the personalization effect in a 2-3 hour long educational simulation with college participants (Brom et al., 2014a). Therefore, if we now *replicate* the effect in a shorter treatment, this would support the idea that the boundary condition is the length of exposure (especially given null and negative results of other experiments with treatments longer than 35 minutes; as discussed in Sec. 2.2.2). If, on the other hand, we *fail to replicate* the effect with the shorter treatment, this would support the idea that the boundary conditions are certain language/cultural backgrounds (while the issue of longer exposure would still be opened).

To give the study sufficient power, we recruited 278 participants (140 for the personalized and 138 for the formal condition) and split these participants among four experiments with the same design. To avoid alternative explanations due to differences in learning materials/research procedures, we first replicated as closely as possible the original study of Moreno & Mayer (2000; their Experiment 2) with an animation on lightning formation with on-screen texts (our Experiments 1, 2). We then replicated their Experiment 2 with another closely comparable animation we created – on how a biological wastewater treatment plant functions (our Experiments 3, 4). Because previous experiments were conducted mainly with high school and university audiences, half of our participants were university students (Experiments 1, 3) and the other half high school students (Experiments 2, 4) (see Table 1).

-- Insert Table 1 around here --

Because it was suggested that the effect of personalization may be reduced or even be negative for expert learners (see Stiller & Jedlicka, 2010; McLaren, DeLeeuw, & Mayer, 2011b; but see also McLaren, DeLeeuw, & Mayer, 2011a; Ginns, Martin, & Marsh, 2013, 2013), we recruited/included only low to moderate prior knowledge learners for Experiment 1 – 4.

The original lightning formation animation was system-paced and it lasted less than 3 minutes (Moreno & Mayer, 2000, p. 726). This was a very short amount of time, as our pilots indicated (the Czech and English texts are of a similar length). Because Ginns et al.'s meta-analysis (2013; p. 466) showed no substantial moderating effect of self-paced vs. system-paced interventions on the personalization effect, we opted for self-paced animations (the average time for completing the animation was about 5-6 minutes).

In Experiment 1 – 4, our primary dependent variable was a posttest score from transfer tests, but we also administered retention tests (as in the original experiment). We administered questionnaires on perceived prior knowledge as in the work of Moreno and Mayer (2000; Exp. 2). The secondary goal of this study is to investigate the possibility of a mediating effect of affective variables on learning. Therefore, we administered questions on self-perceived learning, usefulness of materials, interest, and motivation; based on the works of Moreno and Mayer (2000, Exp. 3) and Kartal (2010). We included one question on friendliness to assess social presence, according to Moreno and Mayer (2004). One additional question measured perceived difficulty. It would be useful to measure intrinsic and extraneous cognitive load, but reliable instruments distinguishing between these two constructs were lacking at the study's onset (e.g., Brünken, Seufert, Paas, & 2010; see also Leppink et al., 2014). Some actually link the perceived difficulty question to certain aspects of cognitive load (e.g., DeLeeuw & Mayer, 2008), but this has been disputed (e.g., de Jong, 2010). Even though the question's connection to cognitive load is problematic, we use it here, because it is often used in the context of personalization principle studies. It is also sensitive to the style of instructional texts (Ginns, Martin, & Marsh, 2013). Finally, we included two questions on perception of the level of formality of instructional texts with the aim to determine participants' opinion regarding the style of texts.

Because some of the "affective" questions above were criticized in the past (e.g., Mayer et al., 2004; Wang et al., 2008), we supplemented them, in Experiments 3 and 4, with widely used inventories assessing generalized positive and negative affect (Positive and Negative Affect Schedule, i.e., PANAS; Watson, Clark, & Tellegen, 1988), the level of flow (Flow Short Scale; Rheinberg, Vollmeyer, & Engeser, 2003), and initial motivation (simplified Questionnaire on Current Motivation, i.e., QCM; Rheinberg, Vollmeyer, & Burns, 2001). These inventories have so far been seldom used in studies on the personalization principle (see Reichelt et al., 2014; Brom et al., 2014a), but the inventory on generalized positive affect, i.e., the first dimension of PANAS, was used in the field of multimedia learning in the context of emotional design (e.g., Plass, Heidig, Hayward, Homer, & Um, 2014). Also, flow is supposed to be experienced in well-designed, game-based learning and similar educational experiences (e.g., Killi, 2006) and flow/engagement is reportedly experienced relatively often when learning from advanced learning technologies (meta-analysed by D'Mello, 2013).

If Czech learners are used to a formal pedagogical approach, as argued in this work, they should prefer direct statements in learning contexts (as opposed to more conversational or polite statements). To investigate whether this is indeed the case, we conducted a supplementary Experiment 5. Therein, we administered a questionnaire to 37 high school students and 101 college students, where they had to rate preferences for statements made by a computerized tutor in two learning situations.[6] These statements were previously rated in terms of politeness by US college participants (Mayer et al., 2006) and we hypothesized that Czech learners would most often prefer statements considered least polite by the US learners, and vice versa (i.e., a reverse pattern compared to what Mayer and colleagues have found).

6. Experiments 1 – 4

6.1 Method

6.1.1 Participants

Number of included participants, age, fraction of females, number of excluded participants, and reasons for exclusion for each experiment are given in Table 2. The university students had the following backgrounds: in psychology, psychology-special education, computer science, physics, economics, linguistics (only Exp. 1), engineering (only Exp. 3) or medicine (only Exp. 3). The largest subgroup of university participants was female students studying either a psychology or a psychology-special education major (Exp. 1: *n* = 23, ~ 40 %; Exp. 3: *n* = 49, ~ 66%). University students were recruited from two, above-average, Czech universities.

As concerns high school participants, they came either from a more theoretically-oriented high school (Exp. 2: *n* = 54; Exp. 4: *n* = 46) or from a more practically-oriented high school (Exp. 2: *n* = 18; Exp. 4: *n* = 28)[7]. All high school students, except for two in Exp. 4, planned to pursue university studies (diverse study programs). High school students were recruited from schools in the capital.

--- Insert Table 2 about here ---

[6] The questionnaire was administered as part of two larger experiments unrelated to the present study.

[7] The Czech schooling system, similarly to, e.g., the German schooling system, distinguishes between these two types of high schools (among others). The more theoretically-oriented is the "gymnázium" and the more practically-oriented is the "střední odborná škola" (technical high school).

All university students participated for course credit. In Exp. 2, high school students were invited to participate in a 2-hr. experimental "workshop". This workshop had two parts. The first part, which lasted around 50 minutes, consisted of an undemanding activity unrelated to the present experiment: students had to judge the quality of animations from a certain educational game (see Šisler et al., 2014 for the game). The second part was the present experiment (i.e., Exp. 2). In Exp. 4, we invited students to participate in a 3 hour long experimental "workshop". This workshop had also two parts. The first one was the present experiment (i.e., Exp. 4). The second part, which lasted around 80 minutes, was unrelated to the present experiment. In the second part, we investigated how students learn from a certain educational game (this was the same game as in Exp. 2). High school students were recruited via an online server advertising short-term jobs for students. The workshop was an extracurricular afternoon activity for them. Each high school student received a compensation for participating in the whole workshop (Exp. 2: 200 CZK, i.e. approximately 7 EUR; Exp. 4: 350 CZK, i.e. approximately 13 EUR).

6.1.2 Paper materials

6.1.2.1 Experiments 1 and 2 – the Lightning Formation Animation

The paper materials consisted of a pretest and a background questionnaire, a retention test with one question, a transfer test with four questions, and a feedback questionnaire. The background questionnaire yielded information about a participant's age, gender, and native language. For Exp. 1, it also included a question on study type. For Exp. 2, it included a question on school type and prospective fields of university studies. For both experiments, it also included one question on self-assessed knowledge of mathematics (a 6-point Likert scale; *1 – very good*; *6 – very weak*); one question on frequency of playing videogames (a 4-point scale; *1 – less than 1 hr. a week*; *4 – more than 10 hrs. a week*); and one question on the frequency of playing live action experiential/simulation games (*Larp*[8] *participation frequency* in the following text; a 5-point scale; *1 – never or I don't know what these games are*; *5 – at least once a month on average*). There are indications that these variables are predictive of learning outcomes when learning mental models from multimedia materials (correlations can be up to .5: Brom et al., 2014a; Brom & Děchtěrenko, 2015). Therefore, we have included these questions to control that the groups were balanced with respect to these variables. There were several additional questions in the background questionnaire, but they were not used for this study's purpose.

The questionnaire on perceived prior knowledge (called pretest for brevity in the following text) was based on Moreno and Mayer's questionnaire (2000; Exp. 1, 2) and it was intended to measure participants' perceived domain knowledge of meteorology (scale 0-20). We reasoned that the topic of lightning formation is more related to prior knowledge of electro-physics rather than

[8] LARP is an abbreviation for Live Action Role Play.

meteorology. Thus, we also included a second perceived prior knowledge questionnaire on electro-physics (scale 0-20). The questions are detailed in Appendix A.

The retention and transfer tests were also based on Moreno and Mayer (2000; Exp. 1, 2) (the wording of our questions was slightly different). The retention test included one sentence at the top of a paper sheet: “Please describe **in detail** how lightning works.” (emphasis in the original). Each of the four transfer test questions was administered on a separate sheet of paper. These questions were: “Imagine that you have limitless possibilities and resources. What would you do to reduce the amount of lightning during storms? Write down **all possibilities** that occur to you.”, “What does air temperature have to do with lightning? Write down **all possibilities** and explain.”, “Imagine you see a sky full of clouds, but you don’t see any lightning. Why? Write down **all** possible **reasons** that occur to you.”, “What causes lightning to occur? Name **all** possible **causes** that you can think of based on the animation that you saw today. Explain.” (emphasis in the original). There were blank lines for writing down the answer below each question.

The feedback questionnaire included the following questions with an 8-point Likert scale (*1 – very good/very much/definitely yes*; *8 – very weak/very little/definitely no*): two questions on self-assessed learning, one question on the animation’s usefulness, two questions on the learner’s level of interest, one question on the learner’s level of motivation, one question on the learner’s perception of difficulty in learning from the animation, and one question on the assessment of the animation’s friendliness. Scores on these questions were reversed for intuitive interpretation so that higher values meant “more” (including “more difficult” for the perception of difficulty). As said above, for simplicity’s sake, we denote all of these variables as affective variables. We also included one open-ended question: “What do you think about the texts you have read in the animation?” and one question on preference regarding the style of instructional texts (the *Language* variable; a 5-point scale: *1 – I would definitely prefer more formal language*; *5 – I would definitely prefer less formal language*). All questions are presented in detail in Appendix B. The questionnaire included five additional questions that were irrelevant for this study’s purpose.

6.1.2.2 Experiments 3 and 4 – the Wastewater Treatment Animation

The paper materials consisted of a pretest with a background questionnaire, an initial motivation questionnaire, prior PANAS, post hoc PANAS, post hoc Flow Short Scale, a retention test with one question, a transfer test with four questions, and a feedback questionnaire. The background questionnaire was the same as in Experiments 1 (or 2) with the following exception: it included two questions on prior tiredness: “How alert do you feel right now?“ (a 6-point scale; *1 – fresh as a spring chicken*; *6 – very tired*) and “How do you feel overall right now?“ (a 6-point Likert scale; *1 – very well*; *6 – very bad*). An average of these two questions is denoted as an *Energy* variable.

The perceived prior knowledge questionnaire (called pretest in the following text) was based on Moreno and Mayer's prior knowledge questionnaire (2000; Exp. 1, 2) and it was intended to measure participants' perceived domain knowledge of chemistry and biological wastewater treatment (scale 0-25). The questions are detailed in Appendix A.

The initial motivation questionnaire was a shortened version of a QCM inventory (Questionnaire on Current Motivation; Rheinberg, Vollmeyer, & Burns, 2001; see Appendix D for selected questions) and it was administered to assess participants' initial motivation to learn the given topic by means of the instructional animation. A shortened version was administered due to time constraints. The original instrument has four factors: challenge, interest, probability of success, and anxiety. We used items from two factors: interest (to which we added two original challenge questions) and anxiety (to which we added one original probability of success question) (8 items altogether).

The PANAS (Positive and Negative Affect Schedule; Watson, Clark & Tellegen, 1988) is an inventory that reliably measures generalized positive and negative affect (i.e., two scales). In our case, we measured affective state before and after the intervention, i.e., twice, to investigate affect induced by the intervention. The pre-post differences are the primary variables of present interest. This approach has already been used in the field of multimedia design (cf. Plass & et al., 2014). A PANAS may come with various initial instructions and our two PANAS assessments had the following instructions: "To what extent do you feel this way **right now?** [the list of 20 feelings]" (prior PANAS; the emphasis in the original); "To what extent did you feel this way **during your interface with the animation**? [the list of 20 feelings]" (post hoc PANAS). We point out that we measured affective *state* in both cases. In the following text, the schedule's positive dimension is denoted as *PANAS+* and its negative dimension as *PANAS-*, the pre-intervention PANAS as *PANAS1* and the post hoc PANAS as *PANAS2*.

Finally, we supplemented the post hoc PANAS with a Flow Short Scale inventory (Rheinberg, Vollmeyer & Engeser, 2003) measuring induced level of flow with 10 items with a 7-point Likert scale. The flow state is usually defined as a pleasant absorption from an activity one is undergoing (Csikszentmihalyi, 1975) and its components include, among others, focused and intense attention. Flow and flow-like states are known to be experienced when learning from advanced learning technologies (see D'Mello, 2013). Positive affect and flow tend to correlate when participants are engaged in interesting tasks (e.g., Rogatko, 2009) and they can predict learning outcomes in computer simulation tasks (Brom et al., 2014a). In Pekrun's terms (Pekrun & Linnenbrink-Garcia, 2012), these two are related to positive activating emotions (cf. Hussain, AlZoubi, Calvo, & D'Mello, 2011). In addition, prior motivation is related to flow under some conditions (Vollmeyer & Rheinberg, 2006).

The retention and transfer tests were based on Moreno and Mayer (2000; Exp. 1, 2), but with questions on how a biological wastewater treatment plant functions. The tests were created with an expert on the topic, according to guidelines by Mayer (2009, pp. 39 – 45), and calibrated on a sample different from the experimental sample. Both tests were administered in the same way as in the present Experiments 1 and 2. The retention test included the following question: "Based on the animation you just saw, describe **in detail** how biological wastewater treatment works". (emphasis in the original). The transfer test included the following four questions: "What would happen if a fungus first appeared in the treatment plant and then bacteria? Write down **all consequences** that come to mind based on the animation you saw today.", "What all does the presence of nutrients in wastewater have to do with the biological wastewater treatment process, i.e., with the functions and functioning of the treatment plant, with bacteria and with fungus? Write down **all possibilities** that occur to you and which relate to any phase of water treatment presented in today's animation. Write an explanation for each **possibility**.", "Based on the animation that you saw today, write down **all factors** that might cause the water treatment devices to not work properly. For each factor, write down **the reasons why** the treatment device would not work properly." and "How could we operate a properly functioning biological wastewater treatment plant using only fungi? **Explain** in detail." (emphasis in the original).

The feedback questionnaire included the same questions as in Experiments 1 and 2. The questionnaire also included ten additional questions that were irrelevant for this study's purpose.

6.1.3 Interventions

The interventions included a black-and-white animation with schematic graphics explaining the process of lightning formation (Exp. 1, 2) or the process of biological wastewater treatment (Exp. 3, 4). Each animation had two versions: a personalized one and a formal one. The complementary version had exactly the same graphics; the only differences were in the language style of the instructional texts.

As concerns the lightning formation animation, its original version (Moreno & Mayer, 2000; Exp. 1, 2) was no longer available; we therefore remade it. The animation used in the present experiment was as close a replica as possible of the original animation.[9] As concerns the biological wastewater treatment animation, we created it anew with an expert on the topic. To ensure similarity to the lightning formation animation, we started with the following requirements: a) it had to have a length similar to the first animation; b) it had to be about a process of the same complexity (and explain it in the same degree of detail); c) it had to avoid numerical calculations; and d) it had to involve a process about which the majority of participants would not have high prior knowledge.

[9] Discussed with Richard Mayer (email dating from 15 March 2013).

For the lightning formation animation, the instructional texts were translated to Czech from the original animation based on Moreno and Mayer (2000, Appendix A). The F (formal) version was written in the third person. The P (personalized) version was formulated from the author's perspective (i.e., in first person) and addressed the learner in the second person. The P version also featured six comments directed at the learner (see Appendix C for the instructions). The same method of personalization as in Experiments 1 and 2 was used also in Experiments 3 and 4. The P version of the biological wastewater treatment animation featured five comments directed at the learner.

The Czech language features two forms of singular, second person pronouns. One is more informal ("*You* [informal] *can* do it." is translated as "*Ty* to *můžeš* udělat.", where "ty" is, syntactically, the singular form of "you") and the other more formal ("*You* [formal] *can* do it." is translated as "*Vy* to *můžete* udělat.", where "vy" is, syntactically, the plural form of "you"). The former is often used when interlocutors are familiar with each other, such as in conversations between friends or family members, or in conversations with children up to around 10-14 years of age (puberty is usually the upper limit for this informal address). The latter is typically used when the interlocutors are not familiar with each other or when there is a formal relationship between them, for instance in teacher – student conversations. In the present study, we used the singular (more informal) version for the personalized animation; as in our previous work (Brom et al., 2014a). Some other languages have a similar feature; for instance, Reichelt et al. (2014) used the plural version (in German) and Kartal (2010) tested both versions (in Turkish) and showed higher benefits for the informal form. We point out that different languages may employ the two forms in slightly different ways.

In both animations, the instructional text, 1-3 sentences on each screen, was placed right below the animation (Figure 2, 3). In the bottom right corner, there was the "next" button. If pressed once, the static image would begin to animate. If pressed a second time, the next set of instructions appeared. Learners had no way of going back.

The whole lightning formation animation consisted of 16 screens. The whole biological wastewater treatment animation consisted of 19 screens. The P version of the lightning formation animation had 317 words; the F version 260 words. For comparison, the original P version had 355 words; the original F version 269 words (including articles). The P version of the wastewater treatment animation had 348 words; the F version 300 words. Table 3 shows average times participants needed to complete the animation.

--- Insert Table 3 about here ---

6.1.4 Procedure

Participants were tested in groups of 1-9 per session. The sampling to conditions was random (gender balanced); participants were assigned to different treatments within each session and worked independently. Each participant was seated at one computer with at least a 17"-wide screen, so that they cannot see on screens of other participants. Procedures somewhat differed for college and high school participants because high school participants attended an experimental workshop composed of two experiments; whereas, college participants attended a single experiment.

6.1.4.1 Experiments 1 and 3 - College Participants

In the introduction, college participants were informed that they would learn from a self-paced animation about the process of lightning formation (or biological waste water treatment) and complete five knowledge questions afterwards. They were then introduced to the animation's interface and how to control the animation (i.e., using the "next" button); without being showed any instructions from the animation. Participants then filled in, at their own pace, the pretests and background questionnaires (and the questionnaire on initial motivation and the prior PANAS in Exp. 3). Afterwards, participants received the animation according to their condition; all participants started at once. In Exp. 1, after the last participant finished, all participants received the feedback questionnaire at once and filled it in at their own pace. In Exp. 3, immediately after a participant finished, he/she received the Flow Short Scale followed by the post hoc PANAS. After all participants completed the post hoc PANAS, they filled in the feedback questionnaire (all at their own pace).

Afterwards, the retention test was administered, followed by the transfer test. The timing was strict for these two tests (Exp. 1, retention: 6 minutes; transfer: 3 minutes for each question; Exp. 3, retention: 7 minutes; transfer: 2:30, 3, 4, 2:30 minutes for each question, respectively). Each of the five knowledge test questions were collected before the next one was distributed (as in Moreno & Mayer, 2000). Afterwards, in Exp. 1, a questionnaire on what learners imagined while learning from the animation was administered – this questionnaire was not used for the present purpose. Likewise, in Exp. 3, a brief personality inventory, irrelevant for present purpose, was administered. Finally, participants were quickly debriefed and thanked.

6.1.4.2 Experiment 2 - High school participants, the Lightning Formation Animation

At the beginning of the workshop, participants were informed that the day's activity would have two parts: in the first one, they would judge the quality of animations in a certain educational game; in the second one, they would learn from a self-paced animation about the process of lightning formation and complete five knowledge questions afterwards. After this introduction, participants filled in pretests and background questionnaires at their own pace. Afterwards, they took part in the activity involving judging animations. Then, after a break, participants received the lightning formation animation according to their condition; all participants started at once. From there on out, the procedure was the same as in Exp. 1.

6.1.4.3 Experiment 4 - High school participants, the Wastewater Treatment Animation

At the workshop's beginning, participants were informed that the day's activity would have two parts: in the first one, they would learn from a self-paced animation about how a biological wastewater treatment plant functions and complete five knowledge questions. Afterwards, in the second one, they would interact with a certain educational game and learn from it. Following this introduction, the procedure was the same as in Exp. 3 (until the end of the first part of the experiment).

6.1.5 Scoring

Two independent evaluators graded retention and transfer test questions. The evaluators have sufficient background in STEM disciplines in order to grade the questions fairly. Inter-rater agreement was measured by two means. First, Pearson's correlation coefficient, *r*, was higher than .93 in all experiments and for both tests. Second, weighted Cohen's κ was used to inspect interrater agreement with weights set to zero on the diagonal and to the squared distance off the diagonal (weights indicate the seriousness of raters' disagreement) (Cohen, 1986). Cohen's κ were in the range .94 - .98 for total scores and .85 - .99 for individual questions (in all four experiments).

The scoring procedures were based on (Mayer & Moreno, 1998). Participants could receive one point for each key idea-unit in the retention test, or half a point for a partially correct idea-unit. These idea-units in the retention tests reflected key steps in the process of lightning formation/wastewater treatment, as described by the animation. There were eight key idea-units for the former animation and 19 for the latter.[10]

Likewise, participants received one point for every creative solution to a transfer test question; or half a point for a partially correct solution (an open-ended scale). Solutions based only on prior knowledge were not rewarded.[11]

[10] The difference in the number of idea-units is not because the latter process is more complex; rather the scoring was more fine-grained compared to the scoring used by Moreno and Mayer (2000). Roughly one idea-unit was related to every animation screen in Experiments 3 and 4. Therefore, had we used the same granularity for the first animation's test, we would have ended up with around 16 idea-units for the lightning formation test.

[11] Absolute values of our transfer tests scores cannot be directly compared to absolute values of transfer test scores from the original experiment (i.e., Moreno & Mayer, 2000; Exp. 2), because the scales for the transfer test could have been calibrated a bit differently. Neither Moreno and Mayer (2000), nor Mayer and Moreno (1998) provided an exact list of "creative solutions" for transfer tests (they provided only examples). (They provided the eight idea-units for the retention test though and we used these eight idea-units.)

6.1.6 Analysis

The data was analyzed in the statistical program R (2015). We averaged the values of two interest variables and two perceived learning variables. Scores from the two raters of the retention and transfer tests were averaged as well. Scores from the Flow Short Scale were analyzed through T-norms provided within standardized Flow Short scale (Rheinberg, 2004). The tests' internal consistency was measured using Cronbach's α (Experiment 3/4: PANAS1+: α = .84/.81; PANAS1-: α = .76/.83; PANAS2+: α = .86/.86; PANAS2-: α =.83/.88; Flow: α = .81/.84; QCM:Interest: α = .74/.71; QCM:Anxiety: α = .72/.69).

To follow the analysis in the key previous studies, differences between the F and P groups were tested using a two-sample t-test in individual experiments. The differences between groups were expressed as Cohen's *d* and classified into small (*d* ~ 0.2), medium (*d* ~ 0.5), and large (*d* ~ 0.8) effect sizes as suggested by Cohen (1988). The correlation between variables was expressed using the Pearson correlation coefficient and classified into small (*r* ~ .1), medium (*r* ~ .3), and large (*r* ~ .5) effect sizes (Cohen, 1988). We also report a 95% confidence interval for Cohen's *d*. For analysis of the general effect of personalization from the results of Exp. 1 – 4 together, we used analysis of variance with additional factors based on the respective tests (usually factors corresponding to the type of animation [lightning animation vs. wastewater treatment] and education level [college vs. high school]). We used η_p^2 for expressing effect size with classification into small (η_p^2 = 0.01), medium (η_p^2 = 0.06), and large (η_p^2 = 0.14) effect sizes (Cohen, 1988). Confidence intervals for η_p^2 were computed by bootstrapping the sample (*N* = 1,000) and computing the adjusted bootstrap percentile metric (Efron, 1987). In a case where all bootstrapped effect sizes were zero, both upper and lower limits of the confidence interval were also zero.

Correction for increased chance of Type I Error (i.e., "correction for multiple comparisons") has not always been used in the context of personalization principle studies, especially not in the seminal studies. We follow the original type of analysis and report the plain t-tests' results as the main findings. However, to inform the reader, we also run an alternative analysis in which we controlled false discovery rate by Benjamini and Hochberg procedure (Benjamini & Hochberg, 1995) and report the findings along with the t-tests' results. As concerns testing whether the groups were balanced with respect to a certain variable, we report only results without this correction (which is stricter for this purpose).

6.2 Results and Discussion

6.2.1 Assignment of Participants to Conditions

With one exception, there were no significant differences between the F and P group for the following variables in any of the experiments: pretest (or both pretests in Experiments 1 and 2), self-assessed knowledge of mathematics, frequency of videogame play, frequency of Larp participation, and the average time during which participants completed the animation. For Experiments 3 and 4, we also found no significant between-group difference regarding both factors of initial motivation, energy, and both prior PANAS+ and PANAS-. Thus, we can assume that the groups were balanced with respect to these variables, with the one exception. This exception was frequency of Larp participation in Exp. 4, for which we found a significant between-group difference (*t*(72) = -2.61, *p* = .011, *d* = -0.61, 95% CI [-1.08, -0.13]; M_P = 1.86 ± 1.11 [±SD]; M_F = 2.54 ± 1.12). However, this variable did not correlate (in Exp. 4) with retention (*r* = -.04) or transfer test scores (*r* = .14). We therefore did not take it as a covariate.

6.2.2 Knowledge Tests – Main Results

The primary question of this study is whether there is a positive effect of the conversational style on test scores; particularly on transfer test scores. As summarized in Tables 4 and 5, in Experiment 1, the F group was marginally better, when uncorrected for multiple comparisons, compared to the P group (medium effect size). In Experiment 2, it was the other way around: the P group outperformed the F group (medium effect size), and this difference remained marginally significant even after correcting for multiple comparisons. In Experiments 3 and 4, no significant difference was found regarding transfer tests. Also, no significant between-group difference was found for retention tests in any of the experiments.

--- Insert Table 4 around here ---

--- Insert Table 5 around here ---

To obtain overall effects, we used a three-way ANOVA with the language style of instructional texts (personalized vs. formal), animation type (lightning formation vs. wastewater treatment), and school level (high school vs. college) as factors. We found no main effects for personalization (transfer: *F*(1, 270) = 0.29, *p* = .594, η_p^2 = 0.00, 95% CI [0.00, 0.01]; retention: *F*(1, 270) = 0.12, *p* = .728, η_p^2 = 0.00, 95% CI [0.00, 0.01]). For the purpose of comparison with previous studies, we converted the effect size estimates to Cohen's *d* using formulas described in Cohen (1988). Corresponding values for Cohen's *d* were 0.04 for retention and 0.07 for transfer. This is the most important finding from this study.

As concerns the main effects for school level, college students scored better in both transfer tests (*F*(1, 270) = 31.92, *p* < .001, η_p^2 = 0.11, 95% CI [0.04, 0.18]) and retention tests (*F*(1, 270) = 13.19, *p* < .001, η_p^2 = 0.05, 95% CI [0.01, 0.10]), but in the case of transfer, this relationship was slightly moderated by the effect of personalization (*F*(1, 270) = 5.31, *p* = .022, η_p^2 = 0.02, 95% CI [0.00, 0.06]) such that the difference between the two types of participants was larger in the F group than in the P group (see Figure 4). Generally, this means that for high school participants (across the whole sample) personalized instructional texts were indeed slightly better for deep, conceptual learning compared to formal instructional texts; it was the other way around for college participants. This is primarily due to Experiments 1 and 2, with some contribution from Experiment 4.[12]

-- Insert Figure 4 around here –

Generally, no difference for retention is not surprising (given past results), but what does not agree well with past results is the limited advantage of the personalized instructional texts as concerns transfer. For comparison, the meta-analysis of Ginns and colleagues (2013) reports for the key variable, the transfer test scores, weighted mean effect size 0.54 and 95% CI [0.25, 0.83]. We found the personalization effect as in Moreno and Mayer (2000, Exp. 2) in one out of four experiments only and with a high school students rather than the college audience (and the effect size was around one-third of that in the original experiment). Also, present results are somewhat fragile in that the pattern revealed in Experiments 1 and 2 (i.e., a weak personalization effect for a high school audience but a weak reverse effect for a college audience) was not replicated in the second couple of experiments with the second animation.

6.2.3 Knowledge Tests – Moderating Effects of Prior Knowledge and Participants' Backgrounds

As showed in the previous section, the data indicated the possibility that the overall null result masks a weak moderation effect for study level (i.e., high school vs. college). We therefore tested the possibility of other moderating effects. The findings of this supplementary analysis should be interpreted with caution, because they are exploratory.

First, we used 6 one-way ANCOVAs to test for possible moderating effects of perceived prior knowledge with a test score as a dependent variable, the language style of instructional texts (personalized vs. formal) as a factor, and a pretest score as a covariate (including an interaction).

[12] For the sake of brevity, we do not report all effects of the models here but in Appendix F. The unreported effects are irrelevant to the main points of the study. For instance, there is always a main effect of animation type on retention (because of different scales for the two retention tests).

Because there were two pretests in Experiments 1 and 2, we run two tests for each of these experiments. No interaction between the pretest score and the style of instructional texts was found for any of the experiments and pretest types (all *p*s > .10). Additionally, we also used 6 two-way ANCOVAs with a test score as a dependent variable, with the language style of instructional texts (personalized vs. formal) and school level (high school vs. college) as factors, and a pretest score as a covariate (including all interactions). Again, no interaction between the pretest score and the style of instructional texts was found for any of the experiments and pretest types (all *p*s > .10). In the six models, only one three-way interaction between the style of instructional texts, school level, and the pretest score was significant with a small effect size (in the case of retention and wastewater treatment animation: $F(1, 135) = 3.96$, $p = .049$, $\eta_p^2 = 0.03$, 95% CI [0.00, 0.11]), which may be an artifact (see Table F3 for details).

Second, the original experiments (Moreno & Mayer (2000); Exp. 1, 2) were conducted with participants from “psychology subject pools” (p. 726, 727) and a couple of other experiments also used psychology students as participants. Our college samples are more heterogeneous; with the second largest subsample being students from technical schools (technical: $n = 30$; psychology or psychology-special education: $n = 76$). This warrants looking for possible moderating effects of participant background (i.e., technical vs. psychology): perhaps the null results were caused because instructional texts in conversational styles are beneficial for psychology audiences but detrimental for technical school students? For both retention and transfer tests and for both participant subgroups, the F group always outperformed the P group (Figure 5). However, when we run 4 two-way ANOVAs (i.e., [language style of texts × animation type], for retention or transfer test scores as dependent variables, and only on participants with psychology or technical background), the difference was significant only in the case of ANOVA for transfer and participants with technical backgrounds (i.e., main effects for language style: $F(1, 26) = 7.51$, $p = .011$, $\eta_p^2 = 0.22$, 95% CI [0.01, 0.48]). Both major subgroups of participants thus contributed to the formal-better-than-conversational pattern of results for college audiences; although participants with technical background contributed more.

-- Insert Figure 5 around here --

6.2.4 Affective Variables – Main Effects

The secondary question of this study is whether there is a positive effect of the conversational style on affective variables. As showed in Tables 6 and 7, generally, we found no such effect. The only notable difference approaching significance (without correction for multiple comparisons) pertained to the Interest variable in Exp. 2 (marginally higher for the P group). We therefore did not perform mediation analysis since there is no main relationship between variables.

--- Insert Table 6 around here ---

--- Insert Table 7 around here ---

6.2.5 Affective Variables – Induced Positive and Negative Affect

Differences between scores from the first and the second presentation of the PANAS questionnaire were significant for both positive (Exp. 3: $t(72) = 7.06, p < .001, d = 0.83$; 95% CI [0.60, 1.05]; Exp. 4: $t(69) = 4.83, p < .001, d = 0.58$, 95% CI [0.34,0.81]) and negative scale (Exp. 3: $t(73) = -6.90, p < .001, d = -0.80$, 95% CI [-1.11, -0.48]; Exp. 4: $t(72) = -6.98, p < .001, d = -0.82$, 95% CI [-1.12, -0.50]) with large effect sizes. Both differences remained significant with $p < .001$ after correction for multiple comparisons. This means that the learning experience induced positive affect (and reduced negative affect).

6.2.6 Affective Variables – Exploratory Analysis

As shown in exploratory correlational analysis (Appendix E), correlations between affective variables and test scores were generally negligible to small (probably caused by noise alone); with three notable exceptions. The first one was perceived difficulty. Participants who judged the animation to be easier scored better in both tests in Experiments 1, 3, and 4 (moderate effect sizes). To investigate the relationship between perceived difficulty and test scores across the entire sample, we ran 2 two -way ANCOVAs (with retention or transfer test scores as dependent variables, with school level and animation type as factors, and perceived difficulty as a covariate; including all interactions). The main effects for perceived difficulty were in medium range and they were significant for both types of tests (transfer: $F(1, 270) = 16.26, p < .001, \eta_p^2 = 0.06$, 95% CI [0.01, 0.12]; retention: $F(1, 270) = 15.17, p < .001, \eta_p^2 = 0.05$, 95% CI [0.02, 0.12]).

The second exception was the level of flow, available to us in Experiments 3 and 4. Participants with higher flow levels had higher test scores (moderate effect sizes, except for the negligible retention × flow correlation in Exp. 4). To investigate the relationship between the level of flow and test scores across the entire sample, we ran 2 one-way ANCOVAs (with retention or transfer tests scores as dependent variables, with school level as a factor, and with flow as a covariate; including all interactions). For transfer, the main effect for flow was in medium to large range (F(1,

141) = 15.51, p < .001, η_p^2 = 0.10, 95% CI [0.02, 0.20]). For retention, the main effect for flow was small, yet marginally significant (F(1, 141) = 3.11, p = .080, η_p^2 = 0.02, 95% CI [0.00, 0.09]).[13]

The final exception was the relationship between generalized positive affect and retention tests scores in Experiments 3 and 4 (small to medium effect sizes). The one-way ANCOVA (with retention test scores as dependent variable, school level as a factor, and positive affect as a covariate; including all interactions) revealed a moderate main effect for positive affect (F(1, 139) = 8.10, p = .005, η_p^2 = 0.06, 95% CI [0.00, 0.14]). For comparison, the main effect of positive affect on transfer was negligible (F(1, 139) = 0.28, p = .598, η_p^2 = 0.00, 95% CI [0.00, 0.02]).

6.2.7 Language Variable – Quantitative Findings

We found a difference between groups in the Language variable with higher scores in the F group than in the P group (Tables 6 and 7). The difference was significant without correction for multiple comparisons, except for Exp. 3, and it was consistent across the four experiments (small to medium effect sizes).

Based on this finding, we explored whether participants in the P condition preferred more formal texts and participants in the F condition preferred less formal ones (across the whole sample). We point out that the medium point for this variable is 3 (i.e., “I was fine with the version of the texts used in the animation.”). The answer to the former question is positive (P group: M_{Language} = 2.82 ± 0.60; t(130) = -3.34, p = .001, d = -0.29, 95% CI [-0.46, -0.12]) but the data does not support the second idea (F group: M_{Language} = 3.06 ± 0.51; t(130) = 1.38, p = .171, d = 0.12, 95% CI [-0.05, 0.29]).

These results mean that some of the P group participants would prefer slightly more formal texts. Meanwhile, the F group participants were more or less satisfied with the texts they were exposed to. This is a useful manipulation check, but this finding also shows that some Czech learners (those who scored 2 or 1) have some reservations regarding the usage of the conversational style in the animation. This fits with findings from our previous study (Brom et al., 2014a), where we also found several such participants. We point out that this finding is actually counterintuitive for Exp. 2: on average, some of the P group participants would have preferred more formal texts, which were

[13] For comparison, flow levels were in the same range in this study as they were for Czech college learners undergoing the 2-3 hour long brewery educational simulation (Brom et al., 2014a), but nearly a standard deviation higher compared to Czech high school students playing a complex, team-based, educational game over 5 hours (Brom et al., 2014b) and one and a half standard deviation higher compared to Czech high school students who received a typical, five hour long, discussion-based, educational workshop without any gaming element (Brom et al., 2014b). The post hoc positive affect (i.e., PANAS2+) was in the same range as in the case of the brewery simulation and the five hour long, educational game. However, it was a standard deviation higher compared to Czech high school students who took part in the discussion-based workshop.

worse for learning for this audience. This indicates that what participants prefer may not always be better for them in the terms of learning gains.[14]

One can ponder as to whether learning was hindered for participants having reservations regarding the conversational style. The data indicates that the truth about this is probably somewhat more complex. We investigated if transfer test scores and perceived difficulty among participants (no matter the treatment condition) preferring “more formal instructional texts” (Language < 3) differed from those who preferred “less formal instructional texts” (Language > 3). Both variables were significantly different between these two subgroups, when corrected for school level and animation type (i.e., main effects for the language style preference [‘ more formal’vs. ‘ less formal’] style in a three-way ANOVA with a transfer test score or perceived difficulty as dependent variables, and the language style preference, school level and animation type as factors; including all interactions: transfer: $F(1, 62) = 4.05$, $p = .048$, $\eta_p^2 = 0.06$, 95% CI [0.00, 0.21]; difficulty: $F(1, 62) = 14.05$, $p < .001$, $\eta_p^2 = 0.18$, 95% CI [0.02, 0.39]). The participants who preferred more formal instructional texts scored, on average, higher on the transfer tasks and perceived the learning to be easier ($M_{\text{transfer}} = 7.00 \pm 2.58$; $M_{\text{difficulty}} = 1.53 \pm 0.98$) compared to those who favored less formal instructions ($M_{\text{transfer}} = 5.78 \pm 2.09$; $M_{\text{difficulty}} = 2.74 \pm 1.75$). This means that reservations regarding personalized instructional texts tended to be expressed primarily by students who performed better on the task and who perceived the task to be easier. Even if learning was hindered for these better learners, they would still outperform those who would prefer more polite instructional texts. Therefore, the reservations expressed by the “Personalized-I-want-more-formal” participants were probably to same extent a matter of taste rather than deriving from substantial problems with learning.

6.2.8 Preference for Language Style of Instructional Texts – Quantitative Findings

We analyzed, across the whole sample, the open-ended question “What do you think about the texts you read in the animation?”. Would the answers explain the preference of some P group participants for more formal instructional texts? After a preliminary screening, we grouped all the answers into the following categories:

a) the answer was irrelevant to the language style of instructional texts;

b) the participant was distracted by the language style of instructional texts;

c) the participant felt the instructional texts were for a younger audience;

d) the participant thought the instructions were too simple and should have been more detailed;

[14] A more intuitive result, one that would support the positive-affect-as-mediator hypothesis, would produce a mean around 3 for the P group (“the present texts are ok”) and a mean higher than 3 for the F group (“please, give me less formal texts”).

e) the participant thought the instructions were too familiar (or not enough formal);

f) the participant thought the instructional texts were too blunt;

g) there were superfluous comments in the instructional texts, according to the participant;

h) the participant considered the topic/instruction to be too complex;

i) the participant considered the instructional texts favorable and, at the same time, explained why he/she thought so;

The majority of answers belonged to group (a). Most of these answers had a positive tone and they mainly claimed that the texts were "short", "succinct", "accurate", and/or "comprehensible". Except for this category and category (i), the statements tended to be negative. One-hundred three answers altogether fell into the categories (b) – (i); these answers were from 91 individual participants (some longer answers were assigned to two groups). Generally, the categories contained similar number of comments from high school and college participants. Certain categories contained statements primarily from the P group participants; one category (f) statements from the F group participants (see Table 8). Only one larger category (d) contained statements from participants from both groups. Upon close inspection, 20 statements from this category (d) concerned themselves primarily with the description's lack of complexity; not with the language style of instructional texts. Of the remaining 83 statements, 45 comments from 40 individual P group participants (~29% of the whole P group; *n* = 140) had a negative tone, but only 14 statements, each from an individual F group participant (~10% of the whole F group; *n* = 138), had a negative tone (and 24 were positive – category (i)). This fits well with the quantitative results above concerning the instructional style preference (i.e., the Language variable) and corroborates the idea that roughly one-third of P group participants had reservations regarding the conversational style whereas the formal style was accepted by nearly all learners.

-- Insert Table 8 around here --

Of the 59 (i.e., 45 + 14) negative statements, only 14 were from the participants working with the wastewater treatment plan animation. For comparison, of the 24 positive (i) category statements, 10 were from the wastewater treatment plan animation. This indicates there might have been differences in the conversational styles between the two treatments, but close comparison (Appendix A) reveals that only one congratulatory comment was in fact missing from the wastewater treatment plan instructions. Could this difference be so important that it caused a dissociation between

Experiments 1 & 2 vs. 3 & 4, as concerns the lack of moderating effect of the school level for the latter pair?

7. Experiment 5

It has been shown (Mayer et al., 2006) that when US college participants rated printed statements (spoken by a computer tutor) in terms of politeness, the participants rated direct statements (e.g., “Click the Enter button.”) as substantially less polite compared to requests (e.g., “I would like you to click the Enter button”), statements formulated as tutor goals (e.g., “I would now click the Enter button”) or statements presented as Socratic hints (e.g., “Do you want to click the Enter button?”). This finding indicates that US college learners would *not* prefer a direct way of address when interfacing with a computer-based tutor.

The purpose of Experiment 5 was to investigate Czech learners’ preferences for statements by a computer tutor; based on the work of Mayer and colleagues. If Czech learners’ preferences are reversed compared to the US sample; that is, if Czech learners prefer a direct way of address when learning, this would support our argument that Czech learners are not used to a polite/conversational way of address in the context of the Czech formal schooling system. This would help to explain the pattern of results from Experiments 1 – 4.

7.1 Method

7.1.1 Participants

There were 37 Czech high school and 101 Czech college participants (different from participants in Experiments 1 - 4 but with the same backgrounds; high school: $M_{\text{age}} = 17.53 \pm 0.65$; college: 22.97 ± 2.51). Both categories of participants took part in an experiment unrelated to the present study. In these experiments, they worked with an educational intervention (high school students with an animation and college participants with a simulation), for which they were recruited as in Experiment 3 (college) or Experiment 4 (high school). These interventions are not relevant for present purposes.

7.1.2 Procedure, materials

One month after both of these experiments, participants underwent a second round of knowledge testing. In this second testing session, we administered several tests and questionnaires, including one questionnaire relevant for present purposes.

In this questionnaire, participants were instructed as follows: “Imagine you are using a learning application with a computer tutor that can help or advise you. In this application, your task is to learn by solving quadratic equations.” Participants were then asked “Would you like the tutor to speak to you in the following way when you should [Situation a) press the Enter button; Situation b) solve the equation based on a worked-out example]?”. Then they rated four types of printed statements related to Situation (a) and an analogical four types of statements related to Situation (b) (a

5-point Likert scale; *1 – definitely no*; *5 – definitely yes*). These statement types were picked from Mayer et al.'s study (2006). The statement types are:

a) Do this... (a direct statement)
b) I would like you to do this... (a request)
c) Do you want to do this...? (a Socratic hint)
d) I would now do this (if I were you)... (a tutor goal)

We point out that all of these statement types can be considered as polite in the Czech language in certain contexts; i.e., in general, they do not sound "foreign" or irritating. These statements are also not colloquial and do not use slang. The particular statements are shown in Table 9. These statements were chosen so that we could contrast the least polite statement (which was the direct one in Mayer et al.'s study (2006)) with the more polite statements (the remaining three).

The Czech language features two forms of singular, second person pronouns: one is more formal and the other more informal (see Section 6.1.3). Therefore, one version of each of the four statements, for each situation, was expressed in the more formal and the other in the more informal tone. Overall, each of the four statement types came in two forms for each situation (i.e., 4 x 2 x 2).

7.1.3 Analysis

For both Situation (a) and Situation (b) differences between the preferences toward the direct statement vs. the other three more polite statements were analyzed as planned contrasts using a within-subject two-way ANOVA with the type of form (formal vs. informal) and the question preference (direct statement vs. rest) as factors . For college students, we also computed three-way ANOVAs with two within-subject factors, the type of form (formal vs. informal) and the question preference (direct statement vs. rest), and one between-subject factor, school type (technical vs. psychology). Both two-way and three-way ANOVAs included all interactions. The classification for effect size measure (i.e., η^2) was as in Experiments 1 – 4. To follow the analysis in the original study, we also correlated scores of corresponding statements from Situation (a) vs. Situation (b) (i.e., eight correlations) and scores of corresponding formal and informal form of statements (i.e., eight correlations).

7.2 Results and Discussion

The raw data, together with the findings from the original study (Mayer et al., 2006), are given in Table 9 (in the original study, the participants rated two forms of politeness for each statement: we present both values). The current main finding is that Czech students preferred direct statements (Situation (a): $F(1, 135) = 328.08$, $p < .001$, $\eta^2 = 0.23$, 95% CI [0.17, 0.28]; Situation(b): $F(1, 137) = 399.8$, $p < .001$, $\eta^2 = 0.26$, 95% CI [0.20, 0.32]). The effect sizes are large. Additionally, students preferred formal statements over the informal ones (Situation (a): $F(1, 136) = 65.18$, $p < .001$, $\eta^2 = 0.06$, 95% CI [0.03, 0.09]; Situation (b): $F(1, 137) = 41.55$, $p < .001$, $\eta^2 = 0.04$, 95% CI

[0.02, 0.06]). The effect sizes are moderate. For both situations, interaction between preference for direct vs. polite statements and preference for formal vs. informal statements was non-significant (*p* > .10). The interaction between school level and preference for direct vs. polite statements also was not significant (*p* > .10).

Our findings generally stand in contradiction to the findings from the original study. In addition, Czech learners prefer, on average, the formal rather than the informal version of the statements.

Mayer and colleagues also found that the most frequent computer users (*n*=7) had a tendency to judge direct statements as more polite compared to the least frequent computer users (*n*=7), whereas the non-direct statements tended to be judged by the least frequent users as more polite. We did not collect information on computer experience, but our participants had diverse backgrounds, so the diverse computer experience could be expected. In addition, we had information about the types of study for college participants. No differences in preferences were found between Czech students studying computer sciences/physics and other students regarding their preferences for the statements (Situation (a): $F(1, 98) = 0.03$, $p = .859$, $\eta^2 = 0.00$, 95% CI [0.00, 0.00]; Situation (b): $F(1, 99) = 0.11$, $p = .745$, $\eta^2 = 0.00$, 95% CI [0.00, 0.01]). For the purposes of internal validation, we conducted a correlational analysis similar to the one in the original study. Participants' preferences for corresponding statements from the two situations highly correlated (*r*s = .59 - .84). In both situations, preferences for the formal and informal forms of each particular statement also correlated (*r*s = .57 - .75); with the exception of the direct statements (in both situations: $r < |.10|$). This probably means that every participant dislikes both forms of each non-direct statement to a similar degree. However, every participant may have differing preferences for the two forms of direct statements.

--- Insert Tab. 9 around here ---

Overall, the data indicates that Czech participants, both college and high school ones (and with diverse backgrounds), prefer on average those statements that were considered the least polite by the US college sample. Czech participants are probably comparable only to the most frequent computer users in the US sample (however, this is quite a small subsample). This strongly supports the argument that preferences for statements made by computer tutors, with differing levels of politeness, may differ across cultures.

8. General Discussion

This study sought to contribute to clarifying boundary conditions for the personalization principle. Specifically, considering our previous study with a Czech college sample (Brom et al., 2014a), in which we failed to replicate the personalization effect in a 2-3 hour long educational simulation, we wondered if we would find the effect (on Czech participants) for animations that were minutes in length; i.e. where large effects were shown with the US audience (Moreno & Mayer, 2000, Exp. 2). The study was successful in answering this question. We found no main effect for personalized instructional texts in short animations (Experiments 1 – 4). Congruent to this finding, we found that Czech learners prefer a computerized tutor to address them with direct and formal, rather than polite and informal, statements in educational settings (Experiment 5). This indicates that the participants' cultural/language background (Czech, in this case) presents a boundary condition, where the personalization principle is fragile.

An exploratory analysis also indicated the possibility of a small moderating effect for school level, such that personalization was slightly beneficial for high school students but it slightly hindered learning for college audiences (as measured by transfer tests). This moderating effect was unstable: it was primarily due to Experiments 1 and 2 (and not 3 and 4). However, we cannot ignore it entirely, because both our largest college subsamples (technical and psychology) contributed to it. In Section 8.1, we will detail our explanation of the main results, and in Section 8.2, we will return to the differences between our high school and college samples.

8.1 Explaining the Main Results

Originally, in terms of the CATLM framework, we adopted the positive-affect-as-mediator-of-learning hypothesis. We speculated that between-group differences in affective variables would either predict between-group differences in learning outcomes or, if no differences in learning outcomes were found (which turned out to be the case), the affective variables would be related to learning outcomes. Such an outcome (if it had happened) could have been interpreted as follows: higher "positive affect" had increased the level of active cognitive processing and thereby improved learning. However, we found a very limited relationship between "positive affect" and learning (as measured by transfer tests) in this study; with the notable exception of perceived difficulty and flow levels. Therefore, our data does not support the straightforward positive-affect-as-mediator hypothesis (we will return to this point in Section 8.3). Nevertheless, the links between flow levels/perceived difficulty and learning outcomes deserve some attention and we will now consider them from the CATLM perspective.

The flow state is typically defined as a pleasant absorption from an activity one is undertaking. This state includes, among other factors, focused and intense attention. The flow intensity can thus be considered as a certain proxy for the level of active cognitive processing: the higher the flow level, the higher the level of active cognitive processing and thereby better learning. If we adopt this view, it would mean that our two conditions did not differ much, on average, in the level of active cognitive processing (because they did not differ much in the flow intensity). From this conjecture, it would also follow that average total cognitive load did not differ much between the two conditions. (There were negligible between-group differences in the level of active cognitive processing (indexed by flow) and negligible differences in learning outcomes. Therefore, cognitive load should also not differ between the conditions, on average.) In terms of Figure 1, this means that total cognitive load would be similar in both the personalized and the formal conditions. The same would apply for allocated cognitive resources. This interpretation obviously depends on accepting the flow level as an index of the allocated cognitive capacity/the level of active cognitive processing and accepting the existence of a linear relationship between cognitive load and learning outcomes. Neither of these presumptions is granted. Therefore, this interpretation should be treated cautiously.

Still, one additional point supports this interpretation. As apparent from Czech learners' preferences (both high school and college level ones) in Experiment 5 for direct and formal statements made by a computerized tutor and the fact that roughly 29% of participants in Experiments 1 – 4 assigned to the personalized condition (both high school and college level ones) expressed reservations regarding the personalized instructional texts, Czech learners were not excited much by the conversational and informal style of the personalized instructional texts. Their comments pertained to cognitive (e.g., "Phrases like 'Let me tell you what happens...' bothered/distracted me...") as well as affective explanations (e.g., "...the texts were written for little kids.") (Table 8). At the same time, while we see that participants assigned to the personalized condition would prefer more formal instructional texts (i.e., the Language variable), this preference is not very potent (*d* = -0.29). In fact, it would be significant in none of the individual experiments, were the tests corrected for multiple comparisons. Participants' learning outcomes in the personalized condition were also not worse, in general, than the learning outcomes of the formal condition participants. Generally, we found that the majority of participants were quite satisfied with the instructions they received and considered them "short", "succinct", "accurate", and/or "comprehensible"; no matter the condition. Therefore, on the one hand, participants in the personalized condition were unlikely influenced positively by the language style of instructional texts. On the other hand, if they were influenced negatively, this influence was probably rather small: both in cognitive terms (e.g., a higher distraction) and "affective" terms (e.g., a lower motivation). Otherwise, we should have witnessed a higher preference for a more formal style, more negative comments and worse learning results. Generally, the personalized way of address was unlikely "a big deal" for the participants.

As concerns perceived difficulty, it is noteworthy that some researchers have used this question as a derivation of Paas's (1992) question on self-reported effort, often used as an index of cognitive load (see de Jong, 2010, pp. 114 – 115; see also Brünken, Seufert, & Paas, 2010). If we accept this idea, this would mean that the total cognitive load of participants in both conditions was, on average, roughly the same (which is in agreement with the flow-based interpretation above). At the level of processes pertaining to the cognitive explanations of the personalization principle, i.e., familiarity and self as a reference point (see the left part of Figure 1), this would mean that either influences from these processes must have counterbalanced each other, or that influences from both processes were negligible. In addition, correlations between flow and perceived difficulty were negative (i.e., the higher the flow, the lower the perceived difficulty; medium to large range; see Table E3, E4). Assuming that perceived difficulty is indeed an index of cognitive load and, at the same time, flow is a proxy for the level of active cognitive participation, the negative correlation between flow and perceived difficulty would mean that those with a lower cognitive load allocated more resources to cope with this load and those with a higher load automatically devoted less resources to deal with this load. That would be an interesting consequence of the above interpretation and our findings. However, as de Jong (2010; p. 114-115, 117) argued persuasively, linking perceived difficulty with total cognitive load, or its components (cf. Ayres, 2006; DeLeeuw & Mayer, 2008), is problematic. Therefore, despite the appeal of the consequence of the flow+difficulty interpretation, this interpretation puts us on even shakier ground than the interpretation based only on flow levels. For this reason, we prefer the flow-only-based interpretation.

The question on perceived difficulty remains of practical interest though, because perceived difficulty seems to predict learning outcomes relatively consistently in studies on the personalization principle (Ginns, Martin, & Marsh, 2013) and beyond (e.g., DeLeeuw & Mayer, 2008). However, future research should link this item to a solid underlying theoretical construct.

The final question is why there was no overall influence of the personalization of instructional texts on learning outcomes for Czech learners even though its influence is well documented for US learners? In our opinion, Czech learners' preference for direct and formal statements in an educational setting (that was clearly apparent in Experiment 5) supports our original hypothesis that Czech learners are less used to polite/conversational forms of address in their traditionally more formal schooling system. Therefore, the positive effects of personalization (be they via social cuing, higher familiarity or priming the self-structure) have not materialized for Czech learners. If personalization had some of these effects for some of the Czech learners, these effects were, on average, counterbalanced by the negative effects of lower familiarity, distraction or slight aversion for other learners.

We will now consider how this explanation can also accommodate the small moderating effect for school level.

8.2 Moderating Effect of School Level

The data indicated the possibility that personalization was slightly beneficial for high school students but it slightly hindered learning for college audiences (as measured by transfer tests). We now offer three possible explanations for this effect.

First, this pattern of results is reminiscent of an expertise reversal effect (Kalyuga, 2007), where learning for expert learners is negatively influenced by a manipulation that has a positive impact on novice learners. This effect is typically explained as follows: the manipulation improves learning of novices, because it features components of external guidance, which reduces the learners' cognitive load. It hinders learning for experts, because they already possess prior schemata/mental models related to the topic and not only do the experts not need external guidance, but the guidance imposes an additional cognitive load on them.

We have not included high prior-knowledge participants in the sample and we have also found no interaction between prior knowledge and the style of instructional texts, which means that no true expertise reversal effect has been found. Nevertheless, our results concerning moderating effect for school level can be interpreted as an "expertise reversal" effect, if general scientific knowledge and general ability to acquire mental models are considered as proxies for prior knowledge: because these two primarily distinguish our high school participants ("novices") and our college learners ("experts"). This is not the first time a kind of expertise reversal effect has been found in the context of the personalization principle (see Stiller & Jedlicka, 2010; McLaren, DeLeeuw, & Mayer, 2011b; but see also McLaren, DeLeeuw, & Mayer, 2011a). Also, school level (Yeung, Jin, & Sweller, 1998) and general scientific knowledge (Lee, Plass, & Homer, 2006) have been considered as proxies to prior knowledge in the past.

Therefore, the first explanation of the moderating effect for school level would be as follows. Czech college learners have arguably higher general scientific knowledge and general ability to acquire mental models than Czech high school students. Thus, conversational styles of instructions would slightly increase cognitive load imposed on college learners; whereas, they would slightly decrease cognitive load imposed on high school learners. These changes to cognitive load could function through the processes discussed in Section 8.1, particularly via familiarity/distraction or priming of the self-structure (see the left part of Figure 1). Alternatively, the changes to cognitive load could be independent or partly independent from these processes.

Second, this pattern of results can also be interpreted from a motivational perspective (the right part of Figure 1). The line of reasoning would be that Czech high school participants would favor personalized instructional texts due to their youth, but Czech college students would dislike them because they are already grown-up. Therefore, Czech high school participants (unlike Czech college learners) would somewhat benefit from personalized instructional texts. However, findings from our Experiment 5 do not support this conjecture, because in that experiment, there were negligible differences in liking direct vs. polite statements between high school and college participants. Our affective data from Experiment 1 - 4 supports this conjecture neither, since there were generally negligible between-group differences in affective variables. We also obtained roughly the same number of negative comments on personalized instructional texts from high school participants as we did from college participants. We therefore do not prefer this interpretation.

Third, the results can be interpreted from the perspective of the on-going school reform in the Czech Republic. The reform has been gradually initiated during the past decade (see Straková & Simonová, 2013; Stolinská, 2012; MEYS, 2004; see also Pol & Rabušicová, 2003). It is primarily aimed at expansion of educational objectives, promoting problem-solving and critical-reasoning skills, promoting creative thinking and motivating for life-long learning, changing student-teacher interaction patterns to emphasize the role of the student, improving methods of formal assessment of students, etc. In fact, the pedagogical approaches and methods promoted by the reform tend to be similar to approaches and methods identified (and somewhat criticized) by Hirsch (1997) as being relatively prominent in US schools. The reform started with the primary education system. Our participants passed through the primary and secondary education levels mostly before this reform started or at the beginning, with high school samples arguably being influenced more than college samples. If we view the reform as an attempt to change the Czech schooling system from a "formal" model toward a more US-like "open" model, which is not totally unfounded, and if we posit that there is no overall effect of conversational style in a "formal" system (due to reasons mentioned in Section 8.1), but a positive effect in an "open" system, this would imply that there should be a higher positive effect for our high school learners compared to our college learners. The limitations of this interpretation are that the highest effect should have been found for Experiment 4, which was conducted last, and that the progress of the reform may actually be small (Straková & Simonová, 2013, p. 476) and its principles have been only partly accepted by teachers so far (Janík et al., 2016).

To conclude, we prefer the first interpretation based on the idea of "expertise reversal" effect. Capitalizing on this idea, one might further wonder what general scientific knowledge (or ability to acquire new mental models) the US college samples in previous studies had? Could it be that it was comparable to, or even lower than, general scientific knowledge in our high school samples? We have no direct information regarding the prior scientific knowledge of our participants (or those from other studies), but the majority of our high school students planned to pursue university studies (roughly 50% of yearly cohorts enroll in university studies in the Czech Republic) and thus they were probably above-average with respect to their same-age peers. At the same time, Czech young adults (16 – 24 years of age) score higher than US young adults in literacy proficiency, numeracy proficiency, and proficiency in problem solving in technology-rich environments (OECD, 2013; p. 72, 73, 82, 83, 93). Therefore, the possibility that our high school participants were comparable in terms of general scientific knowledge to US psychology college participants cannot be entirely excluded. Should this indeed be the case and if lower general scientific knowledge is really connected to benefits from instructions in conversational style, this would mean that part of the cross-cultural differences (on the top of factors discussed in Section 8.1) could be also attributed to differences in prior scientific knowledge.

That said, it should be remembered that the "expertise reversal" effect we found was small and unstable: moderate benefits from personalized instructional texts for the high school audience found in Experiment 2 and moderate detrimental effects from these instructions for the college audience in Experiment 1 were not found in Experiments 4 and 3, respectively. This "expertise reversal" effect was also found during an exploratory analysis rather than by means of hypothesis-driven research. Drawing any general conclusion pertaining to this effect is also complicated by the fact that, as far as we know, there are only a few other studies conducted with high school samples: two German studies with positive results (Dutke, Grefe, & Leopold, 2016; Schneider et al., 2015), one German study reporting an expertise reversal effect (Stiller & Jedlicka, 2010), one US study with null results (McLaren, DeLeeuw, & Mayer, 2011a), and one Flemish study with negative results (Clarebout & Elen, 2007). There are also only a small number of US studies with college participants in which participants' background was reported and, at the same time, it was in a discipline other than psychology (e.g., Wang et al., 2008).

To conclude, we consider "no detected main effect for personalized instructional texts in short animations for Czech learners" as the main result of this study and the lower familiarity, distraction or slight aversion (in the case of Czech learners) as factors likely contributing more to overall outcomes than the effects of prior scientific knowledge (or ability to acquire new mental models). Anyway, the present study warns us that the learning advantages of personalized instructional texts found in a certain language/cultural/schooling context (i.e., for participants with particular language experiences, who passed through a particular schooling system and/or who acquired in their schooling system a certain amount of scientific knowledge) may not automatically be found in a different context. Cross-cultural studies of the personalization principle would be vital. There are factors predicting proficiency in mental model acquisition, such as spatial abilities or mathematical skills (see Brom & Děchtěrenko, 2015), and general scientific knowledge can also be measured (e.g., Lee, Plass, & Homer, 2006). It would be useful if these factors were controlled for in recommended cross-cultural studies to separate the influences of different factors on overall outcomes. Also, investigating possible moderating effects of gender should be considered in future studies.

8.3 Positive Affect as a Predictor of Learning Outcomes

The secondary aim of this study was to investigate the relationship between several affective variables and learning outcomes; that is, the hypothetical link personalization → an increase in positive affect → learning. Despite no overall difference in learning outcomes between the two language styles of instructional texts, we can still research the link's second part. However, as concerns simple measures, except for perceived difficulty, no consistent relationship between any of the measured variables (interest, motivation, friendliness of materials, usefulness and self-assessed learning) and learning outcomes was found. Given past inconsistent results as concerns the ability of these variables to predict learning outcomes in studies on the personalization effect (summarized in Ginns, Martin, & Marsh, 2013), we have to reiterate Mayer et al.'s (2004) conclusion that either some of these measurement instruments does not work very well (in the present context) or the positive-affect-as-mediator-of-learning hypothesis has to be revised (see also Wang et al. (2008)). This does not mean, however, that these measures are necessarily invalid. For instance, some of them consistently correlate with each other, indicating a common denominator. The lack of a linear relationship between affective variables and learning outcomes can also disguise a more complex relationship (see Footnote (4)).

Using more complex measures of generalized affect (PANAS; Watson, Clark, & Tellegen, 1988) and flow levels (Flow Short Scale; Rheinberg, Vollmeyer, & Engeser, 2003) in Experiments 3 and 4 helped partially: we found a small effect of generalized positive affect on retention (but not on transfer), we also found a medium to large effect of flow levels on transfer, and a small, marginally significant, effect of flow levels on retention. In multimedia learning studies, somewhat more robust relationships (generally, medium to high range correlations) between generalized positive affect/flow and learning outcomes tended to be reported from longer (hours rather than minutes) interventions (van der Meij, 2013; Brom et al., 2014a; Brom et al., 2014b). The correlations were less consistent (and in small to medium range only) as concerns shorter interventions (this one; Um et al., 2012; Plass et al., 2014). We also found high to very high correlations between generalized positive affect and flow levels in Czech samples with longer interventions (Brom et al., 2014a, Brom et al., 2014b); but small to medium correlations in this study. One can thus wonder to what extent the positive affect/flow was induced by the animation and to what extent by a different aspect of the experience (such as undergoing the very experiment).

Overall, this indicates that either induced affect and flow play a somewhat limited role in learning from short multimedia interventions, or that PANAS and Flow Short Scale are more reliable in long studies. Still, it is notable that learning experiences in Exp. 3 and 4 induced positive affect and that actual positive affect and flow scores were not lower compared to the scores of Czech participants undergoing game-based learning experiences (see Footnote (13)).

In the future, complex instruments for measuring affective variables are definitely recommended. However, experiments indicating their validity in the context of multimedia design would also be useful.

8.4 Implications

The primary practical implication of this study stems from the detection of a boundary condition for the personalization principle. In the future, any attempt at implementing a personalized treatment in a new language/cultural/schooling context should be backed up by empirical demonstration of the alleged advantages of the principle or at least by a strong theoretical consideration of why instructional texts in a conversational style may be advantageous for the given category of learners.

On a theoretical level, this study raises the following question: can other principles of multimedia learning (Mayer, 2009) be language/culturally dependent? Also, the following questions and their implications deserve attention. How does familiarity operate at the level of cognitive processes and how does it influence cognitive load? At least two forms of familiarity are conceivable within the CATLM framework. The CATLM assumes prior knowledge will be “stored” in the learner’s long-term memory within “permanent” memory structures, which are either present in the memory of a particular learner or not. The first form of familiarity can be conceived in the following way. A) Present memory structures can be either activated or inactivated. B) Only activated memory structures afford advantages of prior knowledge posited by the CATLM. C) Certain knowledge structures can be activated by a familiar aspect of the learning material. D) These knowledge structures would not be activated by less familiar material. Taken together, these points would imply that because intrinsic cognitive load depends on the learning task’s complexity with respect to the learner’s prior knowledge (i.e., activated long-term memory structures), activating previously inactive memory structures would be a manipulation of the intrinsic load. This is an atypical intrinsic load manipulation; previously, this type of load was manipulated primarily by a direct change in the task’s complexity (e.g., segmenting the task) or by a direct change in the learner’s prior knowledge (e.g., pre-training the learner).

The second form of familiarity can be presupposed to operate at the level of input processing of the learning message: i.e., at the level of selecting and organizing the message’s elements in the learner’s working memory. The effectiveness of this process can depend on the learner’s previous experience with particular types of input information (such as with on-screen texts rather than spoken words or with a conversational rather than formal style). This form familiarity pertains more to changes to extraneous rather than intrinsic cognitive load.

Generally, familiarity with certain forms of instructional format is an important concept that would deserve being incorporated into the CATLM as an integral part – and also studied more often (for example, see Schneider et al., 2015; see also Moreno, 2010; p. 137).

This work also has several methodological implications. First, questions on friendliness, self-perceived learning, usefulness, interest, and motivation are intercorrelated and thus probably share a common factor. However, there are better instruments available, such as Flow Short Scale (Rheinberg, Vollmeyer, & Engeser, 2003), PANAS (Watson, Clark, & Tellegen, 1988) or QCM (Rheinberg, Vollmeyer, & Burns, 2001). Still, it would be useful to validate even these more complex inventories in the context of multimedia learning studies; especially studies with brief treatments.

Situational interest is another useful construct in this regard (e.g., Magner et al., 2014). Second, it would be useful if perceived difficulty were linked to a solid theoretical construct, because this question seems to have a predictive value as concerns learning outcomes. Third, related to the last point, several desirable inventories are presently lacking. Most importantly, it would be extremely useful to have valid and reliable measures of actually used cognitive resources, intrinsic load and extraneous load (cf. Brünken, Plass, & Leutner, 2003; Brünken, Seufert, Paas, & 2010; de Jong, 2010; DeLeeuw & Mayer, 2008). Without these, it is difficult to investigate underlying causes behind the principles of multimedia learning. Intrinsic and extraneous cognitive load measures may eventually be available (Leppink, Paas, van Gog, van der Vleuten, & Merriënboer, 2014), but research is needed to validate these items in various contexts, especially with non-psychology/social sciences students, because some of the items refer to psychological concepts that may not be familiar to a general audience (e.g. "I invested a very high mental effort in the complexity of this activity."). We are unaware of a measure of actually used cognitive resources, and the flow level may not be the best proxy for that.

8.5 Limitations

This study's primary objective was to investigate the personalization principle's boundary condition (using a large sample) rather than underlying causes behind the findings. Still, a lack of adequate measures that could help to clarify the individual contributions of the posited underlying causes of the personalization principle (summarized in a complex model in Figure 1) to the study's results can be viewed as the study's limitation. Without these measures (at least a measure of the two types of cognitive load and a measure of actually allocated resources for dealing with this load), attempts at explaining the underlying causes might be, by necessity, speculative only. This is actually a limitation of a substantial portion of this entire research field (see, e.g., de Jong, 2010). As concerns the two types of cognitive load, it could be possible to use the instruments of Leppink and colleagues (2014) in the future. An alternative would be usage of a dual-task paradigm or another objective measure (Brünken, Plass, & Leutner, 2003; Brünken, Seufert, Paas, & 2010), which would however substantially increase the cost of study per participant (probably implying a reduced sample size). However, for now, we are unaware of a direct measure of actually used cognitive resources (except of various, potentially problematic proxies).

Another limitation is a missing variable that would tap participants' perceived similarity between language style of instructional texts in animations and those in textbooks (or given by teachers). This could strengthen the argument that Czech learners are unfamiliar with too familiar/conversational/polite form of address in the context of a formal schooling system. We believe that the findings from Experiment 5 are persuasive enough, but they still provide only indirect evidence. An attempt to measure familiarity with the language style of instructional message has been done by Schneider et al. (2015), and their approach can be considered and extended in future studies.

A dissociation between the learning outcomes of the audiences in Experiments 1 and 2 (i.e., high school vs. college) is not apparent in Experiments 3 and 4. This may be considered as a limitation to generalizability of the findings. The difference between the first and the second pair of experiments is also apparent in the differing numbers of negative statements regarding the language style of instructional texts (lightning formation: 45; wastewater treatment: 14). Therefore, this difference might have real substance. It was most likely caused by a) recruiting slightly different samples (despite the recruitment process having been the same for Experiments 1 and 3, and for 2 and 4), or b) between-treatment difference in the way instructional texts were personalized (but see Appendix C) or c) an interaction between the animation topic, the language style of instructional texts and/or participants' personal characteristics. Anyway, this is not a limitation to the study's main point; rather, it supports it. The weakness and certain inconsistency of the results suggest that the Czech cultural and language background presents a boundary condition for the personalization principle.

There is also one less obvious, methodological limitation: Moreno and Mayer (2000) used in their original experiment a system-paced animation, but we used a self-paced one. Could that cause the different findings? We believe that this is not the case because the meta-analysis of the personalization effect studies (Ginns, Martin, & Marsh, 2013) suggested that there should not be a substantial difference between these two types of animations.

Experiment 5 indicated that Czech learners prefer being addressed using the formal form of the second person singular pronoun rather than the informal form. However, the instructional texts in our animations in Exp. 1 – 4 used the informal form. One can thus ponder as to whether the findings would differ, had we used the formal form. Because differences between the formal and informal forms of statements (in Exp. 5) were small compared to the differences between direct and polite statements, it is conceivable that just changing the form of the second person singular pronoun may not have a large impact. Rather than testing instructional texts using formal vs. informal forms of the second person singular pronoun, it would be worthwhile to address a more general question: what is the impact of different levels of personalization/formalness on instructional texts?

Finally, a limitation of Experiments 1 and 2 is that we have not used complex instruments for measuring affective variables therein. Complementing one- or two-item instruments with more complex inventories will be vital in future studies. A limitation of Experiment 5 is that we have used only translated statements from the study of Mayer and colleagues (2006). Addition of other statements could have provided us a broader view regarding the preferences of Czech learners toward polite or direct statements in educational settings.

To conclude, we believe that the study's limitations do not undermine its key point: certain language/cultural backgrounds present a boundary condition for the personalization principle.

References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*(5), 389-400.

Baddeley, A. D., Eysenck, M., & Anderson, M. C. (2009). *Memory*: Hove: Psychology Press.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods, 37*(3), 379-384.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Bol, N., van Weert, J. C., de Haes, H. C., Loos, E. F., & Smets, E. M. (2015). The Effect of Modality and Narration Style on Recall of Online Health Information: Results From a Web-Based Experiment. *Journal of medical Internet research, 17*(4), e104.

Brom, C., Bromová, E., Děchtěrenko, F., Buchtová, M., & Pergel, M. (2014a). Personalized messages in a brewery educational simulation: Is the personalization principle less robust than previously thought? *Computers & Education*, 72, 339-366.

Brom, C., Buchtová, M., Šisler, V., Děchtěrenko, F., Palme, R., & Glenk, L. M. (2014b). Flow, social interaction anxiety and salivary cortisol responses in serious games: A quasi-experimental study. *Computers & Education*, 79, 69-100.

Brom, C., Děchtěrenko, F. (2015). Mathematical Self-Efficacy as a Determinant of Successful Learning of Mental Models From Computerized Materials. *Proceedings of European Conference on Game-Based Learning* (pp. 89-97)

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53-61.

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load. *Cognitive load theory* (pp. 181-202): Cambridge University Press.

Clarebout, G., & Elen, J. (2007). In Search of Pedagogical Agents' Modality and Dialogue Effects in Open Learning Environments. *E-Journal of Instructional Science and Technology, 10*.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational psychology review, 3*(3), 149-210.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.): Hillsdale, NJ: Erlbaum.

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety*: Jossey–Bass, San Francisco, CA.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*(2), 105-134.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of educational psychology, 100*(1), 223-234.

Doolittle, P. (2010). The effects of segmentation and personalization on superficial and comprehensive strategy instruction in multimedia learning environments. *Journal of Educational Multimedia and Hypermedia, 19*(2), 159-175.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171-185.

Ginns, P., Martin, A. J., & Marsh, H. W. (2013). Designing instructional text in a conversational style: a meta-analysis. *Educational psychology review, 25*(4), 445-472.

Grice, H. P. (1975). Logic and conversation *Syntax and semantics 3: Speech arts* (pp. 41-58): New York, NY: Academic.

Hirsh, E. D. (1997). *The Schools We Need: And Why We Don't Have Them*: Anchor, New York.

Hussain, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011). Affect detection from multichannel physiology during learning sessions with AutoTutor *Artificial Intelligence in Education* (Vol. 6738, pp. 131-138): Springer.

Janík, T., Pešková, K., Janko, T., Spurná, M., Knecht, P. (2016) *Vnímání kurikulárních změn učiteli ZŠ: Shrnutí výsledků dotazníkové šetření.* [in Czech, "Perception of curricular changes by primary school teachers: Summary of survey findings."] A paper presented at the 24th Conference of Czech Educational Research Association. České Budějovice.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review, 19*(4), 509-539.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational Psychology Review, 23*(1), 1-19.

Kartal, G. (2007). How universal are e-learning design guidelines? Reconsidering the personalization principle *Proceedings of the 2nd International Conference on E-Learning* (pp. 269-275).

Kartal, G. (2010). Does language matter in multimedia learning? Personalization principle revisited. *Journal of educational psychology, 102*(3), 615-624.

Kiili, K. (2006). Evaluations of an experiential gaming model. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*, *2*(2), 187-201.

Kurt, A. A. (2011). Personalization Principle in Multimedia Learning: Conversational versus Formal Style in Written Word. *Turkish Online Journal of Educational Technology-TOJET, 10*(3), 185-192.

Lee, H., Plass, J. L., & Homer, B. D. (2006). Optimizing cognitive load for learning from computer-based science simulations. *Journal of educational psychology, 98*(4), 902-913.

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32-42.

Magner, U. I., Schwonke, R., Aleven, V., Popescu, O., & Renkl, A. (2014). Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. *Learning and Instruction, 29*, 141-152.

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.): Cambridge University Press.

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction, 29*, 171-173.

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A Personalization Effect in Multimedia Learning: Students Learn Better When Words Are in Conversational Style Rather Than Formal Style. *Journal of educational psychology, 96*(2), 389-395.

Mayer, R. E., Johnson, W. L., Shaw, E., & Sandhu, S. (2006). Constructing computer-based tutors that are socially sensitive: Politeness in educational software. *International Journal of Human-Computer Studies, 64*(1), 36-42.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of educational psychology, 90*, 312-320.

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011a). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education, 56*(3), 574-584.

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011b). A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies, 69*(1), 70-79.

McLaren, B. M., Lim, S.-J., Gagnon, F., Yaron, D., & Koedinger, K. R. (2006). Studying the effects of personalized language and worked examples in the context of a web-based intelligent tutor. *Lecture Notes in Computer Science: Vol. 4053. Intelligent tutoring systems* (pp. 318-328): Springer.

MEYS, Ministry of Education, Youth and Sports of the Czech Republic (2004). *Act No. 561/2004 Coll. of 24 September 2004 on Pre-School, Basic, Secondary, Tertiary Professional and Other Education*. http://www.msmt.cz/vzdelavani/skolstvi-v-cr/act-no-561-2004-coll-of-24-september-2004-on-pre-school?lang=1 (Accessed 2016-10-30)

Moreno, R. (2005). Instructional technology: Promise and pitfalls. *Technology-based education: Bringing researchers and practitioners together* (pp. 1-19): Information Age Publishing.

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of educational psychology, 92*(4), 724-733.

Moreno, R., & Mayer, R. E. (2004). Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology*, *96*(1), 165-173.

Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational psychology review, 19*(3), 309-326.

OECD (2013). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. OECD Publishing. http://dx.doi.org/10.1787/9789264204256-en (Accessed 2016-10-30)

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84*, 429–434.

Palečková, J. (1999). Performance assessment in the Czech Republic. *Studies in Educational Evaluation, 25*(3), 265-268.

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic Emotions and Student Engagement *Handbook of Research on Student Engagement* (pp. 259-282): Springer Science+Business Media.

Perry, L. B. (2005). The seeing and the seen: contrasting perspectives of post-communist Czech schooling. *Compare, 35*(3), 265-283.

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction, 29*, 128-140. doi: 10.1016/j.learninstruc.2013.02.006

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. Cambridge University Press.

Pol, M., & Rabusicova, M. (2003). The White Book Launched: On the Prospects of Education in the Czech Republic. ERIC document, nr. ED497460. Available online (http://eric.ed.gov/?id=ED497460).

Průcha, J. (2015) *Srovnávací pedagogika* [in Czech: "Comparative pedagogy"], 3rd. ed., Portál.

Reeves, B., & Nass, C. (1996). *The media equation:: How people treat computers, television, and new media like real people and places*: New York, NY: Cambridge University Press.

Reichelt, M., Kämmerer, F., Niegemann, H. M., & Zander, S. (2014). Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human behavior, 35*, 199-210.

Rey, G. D., & Steib, N. (2013). The personalization effect in multimedia learning: The influence of dialect. *Computers in Human behavior, 29*(5), 2022-2028.

Rheinberg, F. (2004). *Motivationsdiagnostik [Motivation diagnosis]*: Gottingen: Hogrefe

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen [QCM: A questionnaire to assess current motivation in learning situations]. *Diagnostica, 47*, 57-66.

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). Die Erfassung des Flow-Erlebens [in German]. In J. Steinsmeier-Pelster & F. Rheinberg (Eds.), *Diagnostik von Motivation und Selbstkonzept* (pp. 261-279): Hogrefe.

Rogatko, T. P. (2009). The influence of flow on positive affect in college students. *Journal of Happiness Studies, 10*(2), 133-148.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of personality and social psychology, 35*(9), 677-688.

Schneider, S., Nebel, S., Pradel, S., & Rey, G. D. (2015). Introducing the familiarity mechanism: A unified explanatory approach for the personalization effect and the examination of youth slang in multimedia learning. *Computers in Human behavior, 43*, 129-138.

Schworm, S., & Stiller, K. D. (2012). Does personalization matter? The role of social cues in instructional explanations. *Intelligent Decision Technologies, 6*(2), 105-111.

Son, J. Y., & Goldstone, R. L. (2009). Contextualization in perspective. *Cognition and Instruction, 27*(1), 51-89.

Stigler, J. W., & Perry, M. (1988). Mathematics learning in Japanese, Chinese, and American classrooms. *New Directions for Child and Adolescent Development, 41*, 27-54.

Stiller, K. D., & Jedlicka, R. (2010). A kind of expertise reversal effect: Personalisation effect can depend on domain-specific prior knowledge. *Australasian Journal of Educational Technology, 26*, 133-149.

Stolinská, D. (2012). *Interakce učitel-žák v proměnách primárního vzdělávání*. [in Czech: "Teacher-pupil interaction in the transformation of primary education"] Ph.D. thesis. Palacky University in Olomouc.

Straková, J., & Simonová, J. (2013). Assessment in the school systems of the Czech Republic. *Assessment in Education: Principles, Policy & Practice*, *20*(4), 470-490.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction, 4*(4), 295-312.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory.* New York: Springer.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: a meta-analysis. *Psychological bulletin, 121*(3), 371-394.

Šisler, V, Gemrot, J., Kokoška, S., Černá, M., Cuhra, J., Hoppe, J., Činátl, K., Pinkas, J., Brom, C. (2014). Czechoslovakia 38-89, Assassination 1.0., a computer game. Retrieved from http://cs3889.cz.

Tze Wei, L., Su-Mae, T., & Nuo Wi, T. (2014). The Role of Learners' Field Dependence and Gender on the Effects of Conversational versus Non-Conversational Narrations in Multimedia Environment. *Journal of Interactive Learning Research, 25*(2), 281-302.

Um, E. R., Plass, J. L., Hayward, E. O., & Homer, B. D. (2012). Emotional Design in Multimedia Learning. *Journal of educational psychology, 104*(2), 485-498. doi: 10.1037/a0026609

U.S. Department of Education, National Center for Education Statistics (2003). *Teaching Mathematics in Seven Countries: Results From the TIMSS 1999 Video Study*. NCES (2003-013). Washington, DC.

van der Meij, H. (2013). Motivating agents in software tutorials. *Computers in Human behavior, 29*(3), 845-857.

Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational psychology review, 18*(3), 239-253.

Wang, N., & Johnson, W. L. (2008) The politeness effect in an intelligent foreign language tutoring system. *Lecture Notes in Computer Science, vol. 5091. Intelligent Tutoring Systems* (pp. 260-280): Springer.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66*(2), 98-112.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063-1070.

Yeung, A. S., Jin, P., & Sweller, J. (1998). Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary educational psychology, 23*, 1-21.

Figures

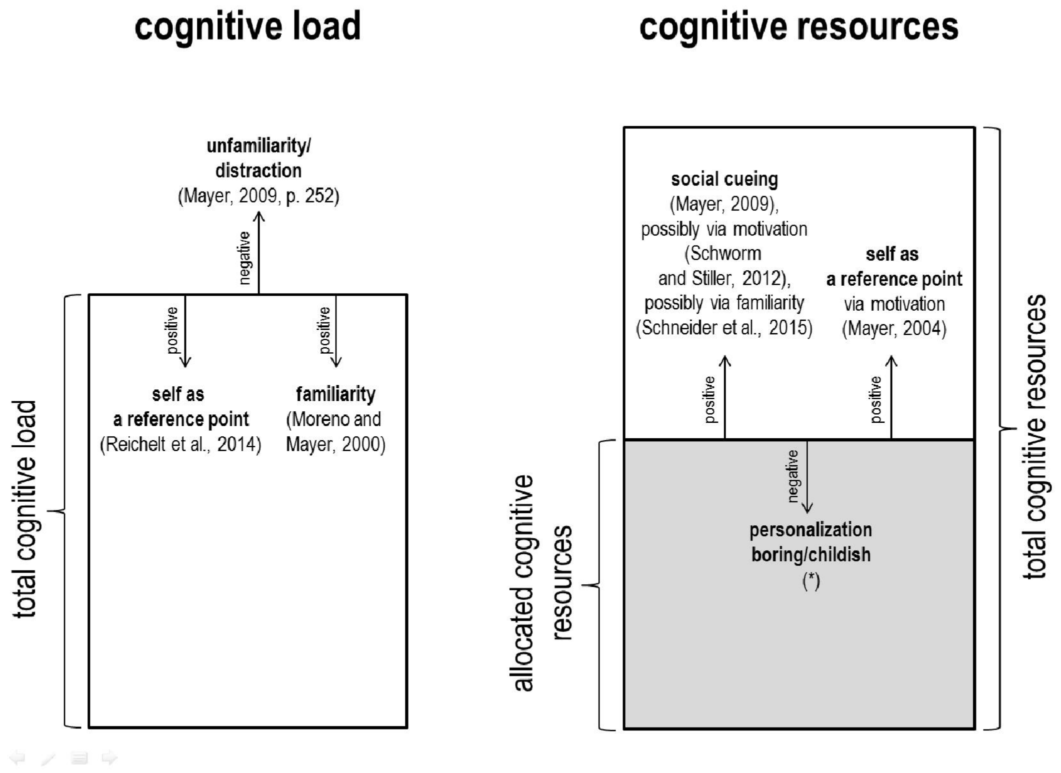
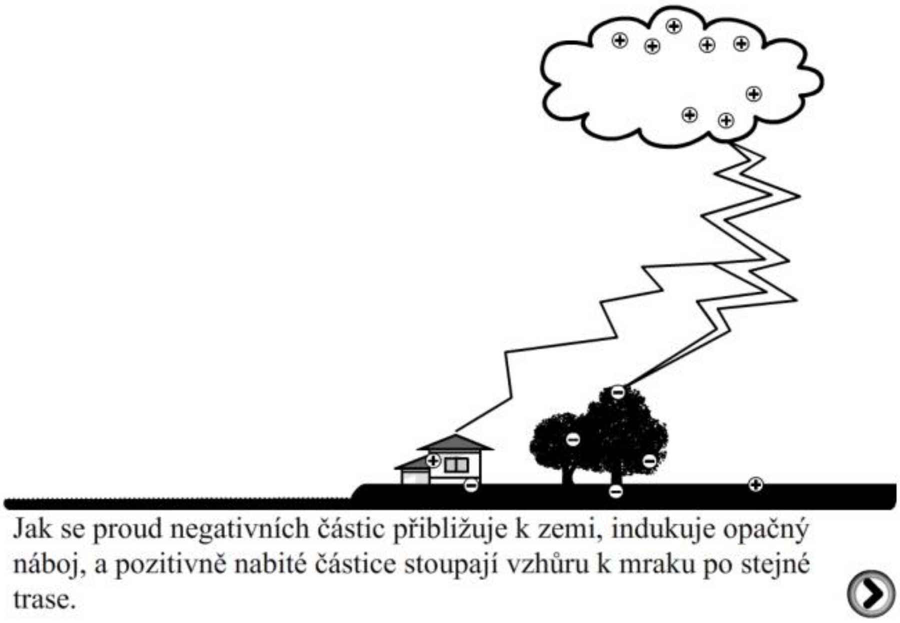
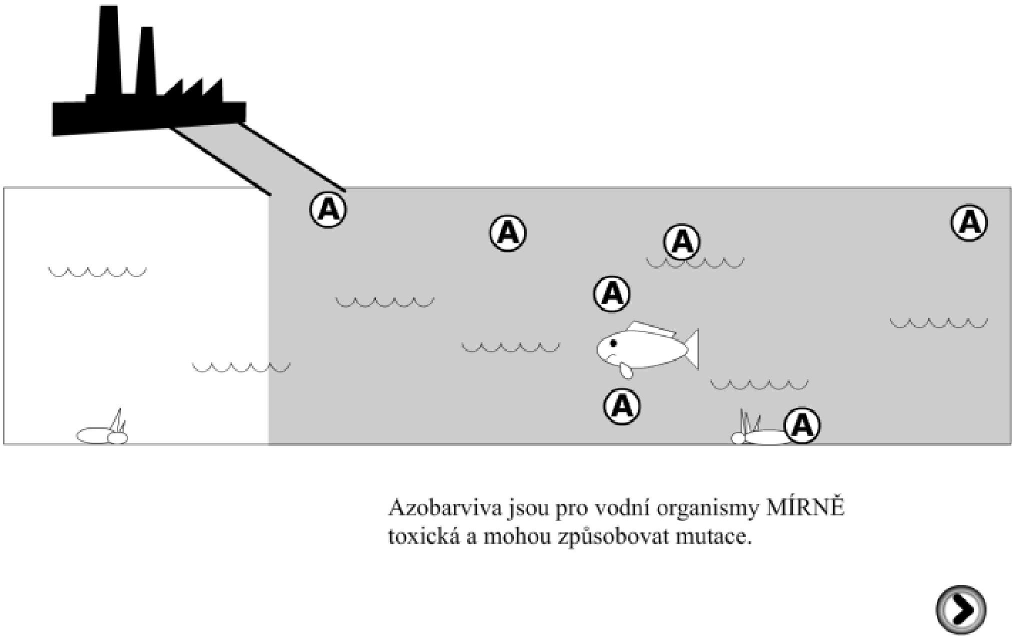


Figure 1. The theoretical model summarizing explanations behind the personalization principle. The right side of the figure shows explanations pertaining to changes in allocated cognitive resources. The left side shows explanations pertaining to changes in the total cognitive load imposed on the learner. Arrows denote positive or negative changes to cognitive load or allocated cognitive resources. Examples of works mentioning the given explanations are also shown (with the asterisk (*) denoting that the explanation in question has not been, to our knowledge, mentioned in the literature).



Jak se proud negativních částic přibližuje k zemi, indukuje opačný náboj, a pozitivně nabité částice stoupají vzhůru k mraku po stejné trase.

Figure 2. A screenshot from the lightning formation animation (instructions in Czech).



Azobarviva jsou pro vodní organismy MÍRNĚ toxická a mohou způsobovat mutace.

Figure 3. A screenshot from the biological wastewater treatment plant animation (instructions in Czech).

.

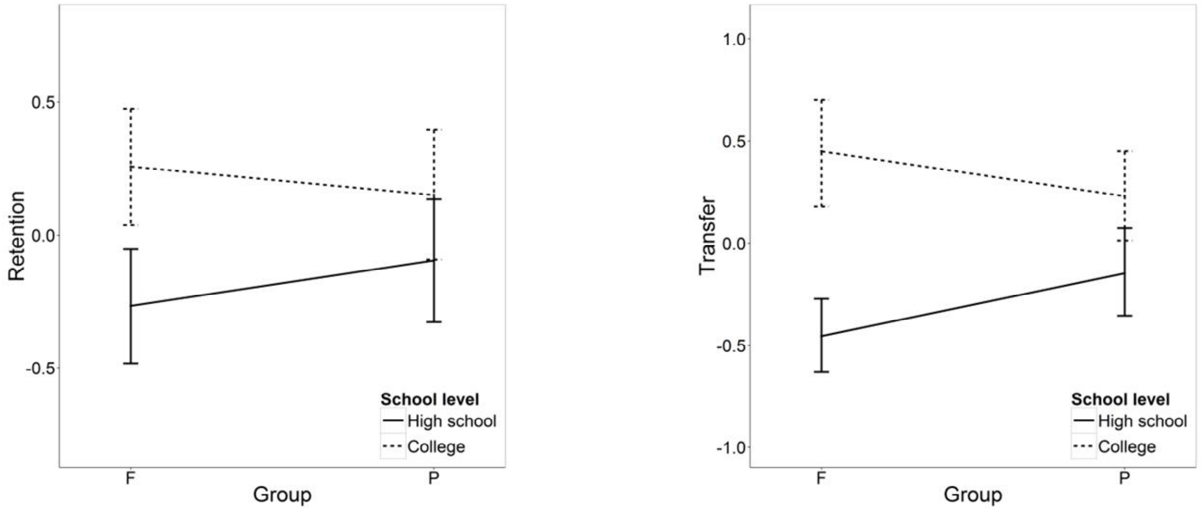


Figure 4. Mean values for each group and school level for the retention and transfer tests (merged data from Exp. 1 – 4). Whiskers denote SEM. The test scores are depicted after we z-transformed them separately for the first two experiments and for the last two experiments because the scores have different scales in the two animation treatments.

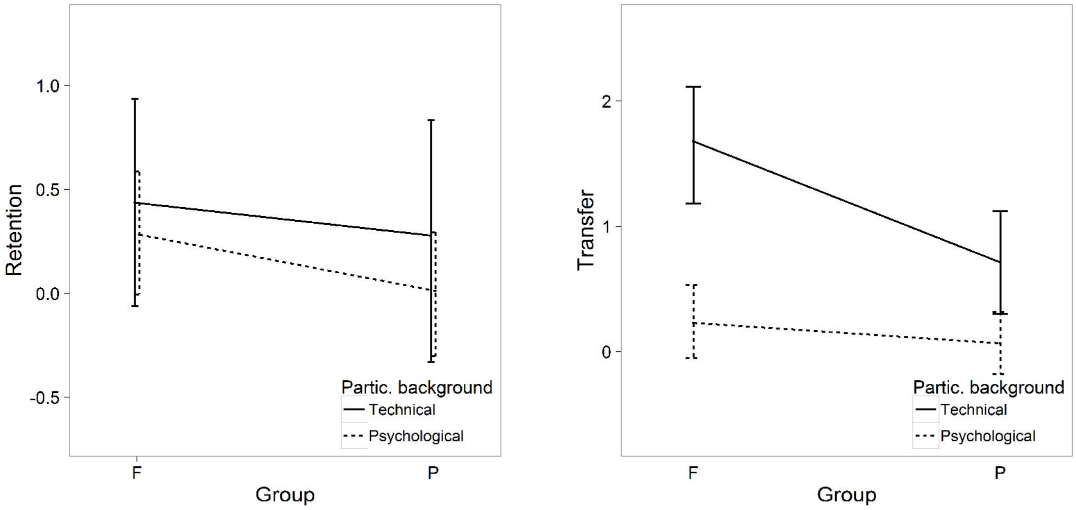


Figure 5. Mean values for each experimental group and participant subgroup for the retention and transfer tests (merged data from Exp. 1 and 3). Whiskers denote SEM. The test scores are depicted after we z-transformed them separately the two experiments because the scores have different scales in the two animation treatments.

Tables

Table 1

Summary of the four personalization principle experiments conducted in this study

Experiment	Sample	Animation	Cohen's *d*[a]
Experiment 1	College	Lightning formation	-0.45†
Experiment 2	High school	Lightning formation	0.48*
Experiment 3	College	Wastewater treatment plant	-0.04
Experiment 4	High school	Wastewater treatment plant	0.22

[a] A positive *d* means higher transfer test scores for participants in the personalized condition.

† $p < .10$. * $p < .05$. without a correction for multiple comparison.

Table 2

Experimental sample

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Animation	L. f.[a]	L. f.	W. t.[b]	W. t.
Sample	College	High school	College	High school
Age (SD)	22.18 (2.72)	17.32 (0.72)	22.05 (2.49)	17.11 (0.85)
n (Formal + Personalized)	27 + 30	37 + 36	37 + 37	37 + 37
Females	60 %	62 %	70 %	51 %
Excluded	9	7	5	6
Non-native speakers[c]	4	5	2	4
Sick during the exp.	2	1	-	-
Different age group	1[d]	-	2[e]	-
Very high prior knowledge	2	-	-	-
Not understanding instructions	-	1	-	-
Extremely tired at the beginning	-	-	1	1
Not answering knowledge test	-	-	-	1

[a] L. f.: Lightning formation

[b] W. t.: Wastewater treatment plant

[c] Neither Czech nor Slovak. Czechoslovakia was a federation of Czech and Slovak Republic until the end of 1992. The Slovak language is very close to the Czech language. Many Slovak students study in the Czech Republic, and it is generally no problem for Slovak university students to understand or even speak Czech fluently.

[d] 47 years old

[e] 38 and 45 years old

Table 3

Average times participants needed to complete the animation.

Experiment	Animation	P version	F version
Experiment 1	Lightning formation	5:07 (1:16)	4:51 (1:14)
Experiment 2	Lightning formation	4:42 (0:50)	4:43 (1:03)
Experiment 3	Wastewater treatment plant	6:41 (1:47)	6:01 (1:53)
Experiment 4	Wastewater treatment plant	6:12 (1:30)	5:59 (1:29)

Table 4

Means, SDs, effect sizes, 95% CI, and numbers of participants included in the analysis for the retention and transfer tests for Experiments 1 and 2

	Experiment 1				Experiment 2			
	F	P			F	P		
	M	*M*	*d*	N_F	*M*	*M*	*d*	N_F
	(SD)	(SD)	[CI]	N_P	(SD)	(SD)	[CI]	N_P
Retention	5.16	4.92	-0.20	27;	4.67	4.94	0.23	37;
	(0.92)	(1.41)	[-0.73; 0.33]	30	(1.22)	(1.16)	[-0.24; 0.69]	36
Transfer	8.05	6.91	-0.45†	27;	5.50	6.58	0.48*$^{,+}$	37;
	(2.79)	(2.23)	[-0.99; 0.09]	30	(1.98)	(2.54)	[0.00; 0.95]	36

† $p < .10$. * $p < .05$ without correction for multiple comparisons.

+ $p < .10$ when corrected for multiple comparisons by Benjamini and Hochberg procedure.

Table 5

Means, SDs, effect sizes, 95% CI, and numbers of participants included in the analysis for the retention and transfer tests for Experiments 3 and 4

	Experiment 3				Experiment 4			
	F	P			F	P		
	M	*M*	*d*	N_F	*M*	*M*	*d*	N_F
	(SD)	(SD)	[CI]	N_P	(SD)	(SD)	[CI]	N_P
Retention	11.75 (3.17)	11.66 (2.96)	-0.03 [-0.49; 0.43]	37; 37	9.68 (3.21)	10.07 (3.50)	0.12 [-0.35; 0.58]	37; 37
Transfer	7.66 (2.95)	7.55 (2.37)	-0.04 [-0.50; 0.42]	37; 37	5.49 (2.03)	5.98 (2.49)	0.22 [-0.25; 0.68]	37; 37

Note. None of the differences was significant.

Table 6

Means, SDs, effect sizes, 95% CI, and numbers of participants included in the analysis for affective variables and the Language preference variable for Experiments 1 and 2

	Experiment 1				Experiment 2			
	F	P			F	P		
	M	*M*	*d*	N_F	*M*	*M*	*d*	N_F
	(SD)	(SD)	[CI]	N_P	(SD)	(SD)	[CI]	N_P
Learning[a]	5.07 (1.23)	5.10 (1.16)	0.02 [0.51, 0.55]	27; 30	5.32 (0.82)	5.49 (0.93)	0.18 [-0.28, 0.65]	37; 36
Usefulness[a]	5.04 (1.32)	5.30 (1.39)	0.19 [-0.34, 0.73]	27; 30	5.62 (1.44)	5.67 (1.15)	0.03 [-0.43, 0.50]	37; 36
Interest[a]	4.80 (1.53)	5.18 (0.99)	0.30 [-0.23, 0.84]	27; 30	4.54 (1.52)	5.08 (1.24)	0.39† [-0.08, 0.86]	37; 36
Motivation[a]	5.59 (1.72)	5.37 (1.61)	-0.14 [-0.67, 0.40]	27; 30	5.92 (1.66)	6.06 (1.17)	0.10 [-0.37, 0.56]	37; 36
Difficulty[a]	2.30 (1.38)	2.10 (1.37)	-0.14 [-0.67, 0.39]	27; 30	2.11 (1.74)	2.33 (1.59)	0.14 [-0.33, 0.60]	37; 36
Friendliness[a]	5.48 (1.40)	5.93 (0.64)	0.42 [-0.11, 0.96]	27; 30	5.78 (1.11)	6.03 (0.88)	0.24 [-0.22, 0.71]	37; 36
Language[b]	3.05 (0.39)	2.71 (0.64)	-0.53† [-1.07, 0.01]	20; 21	3.05 (0.70)	2.72 (0.85)	-0.43† [-0.90, 0.05]	37; 36

[a] Scale 0 – 7. Higher values mean “more”.

[b] Preference for the style of language used in instructional materials. Scale 1 – 5.

† $p < .10$ (without correction for multiple comparisons).

Table 7

Means, SDs, effect sizes, 95% CI, and numbers of participants included in the analysis for affective variables and the Language preference variable for Experiments 3 and 4

	Experiment 3				Experiment 4			
	F	P			F	P		
	M	*M*	*d*	N_F	*M*	*M*	*d*	N_F
	(SD)	(SD)	[CI]	N_P	(SD)	(SD)	[CI]	N_P
Learning[a]	4.68	4.61	-0.05	37;	4.95	5.09	0.12	37;
	(1.39)	(1.49)	[-0.51, 0.42]	37	(1.28)	(1.12)	[-0.34, 0.59]	37
Usefulness[a]	5.46	5.59	0.09	37;	5.19	5.35	0.11	37;
	(1.59)	(1.30)	[-0.37, 0.56]	37	(1.56)	(1.40)	[-0.35, 0.57]	37
Interest[a]	5.59	5.82	0.23	37;	5.20	5.47	0.26	37;
	(1.10)	(0.93)	[-0.24, 0.69]	37	(1.26)	(0.81)	[-0.21, 0.72]	37
Motivation[a]	6.08	6.32	0.17	37;	6.19	6.24	0.04	37;
	(1.61)	(1.20)	[-0.29, 0.64]	37	(1.45)	(0.95)	[-0.42, 0.51]	37
Difficulty[a]	1.59	1.16	-0.38	37;	1.95	1.46	-0.35	37;
	(1.26)	(1.01)	[-0.85, 0.09]	37	(1.58)	(1.17)	[-0.82, 0.12]	37
Friendliness	6.22	6.16	-0.05	37;	6.24	6.27	0.04	37;
[a]	(1.16)	(0.90)	[-0.52, 0.41]	37	(0.72)	(0.65)	[-0.42, 0.50]	37
Language[b]	3.03	2.92	-0.29	37;	3.11	2.89	-0.47*	37;
	(0.37)	(0.36)	[-0.76, 0.17]	37	(0.46)	(0.46)	[-0.94, 0.00]	37
IM:interest[c]	19.88	20.88	0.18	34;	22.06	20.59	-0.30	32;
	(5.66)	(5.16)	[-0.29, 0.64]	34	(4.85)	(4.59)	[-0.76, 0.17]	34
IM:anxiety[d]	14.49	13.82	-0.15	35;	12.12	11.64	-0.12	34;
	(4.32)	(4.00)	[-0.62, 0.31]	34	(3.83)	(4.22)	[-0.58, 0.35]	36
PANAS1+[e]	27.30	28.19	0.14	37;	28.92	28.28	-0.10	36;
	(5.86)	(6.73)	[-0.32, 0.61]	36	(6.13)	(6.21)	[-0.57, 0.36]	36
PANAS1-[e]	14.30	14.81	0.14	37;	17.00	16.95	-0.01	37;
	(3.67)	(3.74)	[-0.33, 0.60]	37	(5.88)	(5.46)	[-0.47, 0.45]	37
PANAS2+[e]	30.49	31.86	0.20	37;	30.89	30.91	0.00	37;
	(6.03)	(7.42)	[-0.26, 0.67]	37	(6.48)	(7.15)	[-0.46, 0.47]	35
PANAS2-[e]	12.19	12.30	0.03	37;	14.59	13.33	-0.26	37;
	(3.69)	(2.92)	[-0.43, 0.50]	37	(5.96)	(3.33)	[-0.72, 0.21]	36
Flow[f]	55.32	57.32	0.27	37;	54.47	55.26	0.09	36;
	(8.21)	(6.22)	[-0.19, 0.74]	36	(8.02)	(8.17)	[-0.37, 0.56]	34
PANAS+	3.19	3.83	0.15	37;	2.00	2.59	0.14	37;
diff	(4.25)	(4.27)	[-0.31, 0.61]	37	(3.67)	(4.28)	[-0.32, 0.61]	36
PANAS-	-2.11	-2.51	-0.14	37;	-2.41	-3.81	-0.37	37;
diff	(2.92)	(2.86)	[-0.60, 0.32]	37	(3.59)	(3.90)	[-0.84, 0.10]	36

[a]Scale 0 – 7. Higher values mean “more”.

[b]Preference for the style of language used in instructional materials. Scale 1 – 5.

[c]Scale 5 – 35. Higher values mean “more”.

[d]Scale 3 - 21. Higher values mean “less anxiety”.

[e]Scale 10 – 50. Higher values mean “more”.

[f]Scale 21 – 74 after the transformation through T-norms. Higher values mean “more”.

*$p < .05$ (without correction for multiple comparisons).

Table 8

Instructional style preferences: Statement categories (except for (a) – irrelevant), number and percentage of statements in each category, representative examples, the participant's language style preference, and the participant's characteristics

Category	Numbers of statements and percentage of statements per participants[a]				Examples	Pref.[b]	Participant[c]
	Formal		Personalized				
	High school	College	High school	College			
	n = 74	*n* = 64	*n* = 73	*n* = 67			
All categories, except for (a)	17 (23%)	11 (17.2%)	34 (46.6%)	41 (61.2%)			
b. Distraction (negative-Mixed)[d]	1 (1.4%)	-	-	2 (3.0%)	*Phrases like "Let me tell you what happens..." bothered/distracted me, but the principle was well-described.*	2	LF-Col-P-m-24
c. For younger (negative-P)	-	-	7 (9.6%)	9 (13.4%)	*...the texts were written for little kids.*	2	LF-Col-P-f-21
					...sentences like "now you have your cloud" are perhaps unnecessary for high school students (more for 7th graders)	2	LF-Sch-P-f-17
d. Too simple (negative-Mixed)	6 (8.1%)	6 (9.4%)	2 (2.7%)	6 (9.0%)	*...in some cases there was way too little information.*	2	LF-Col-F-f-22
					Sometimes, but not very often, the text did not explain everything I expected.	3	WT-Sch-F-m-17
					Comprehensible, but I would welcome more details.	2	LF-Col-P-m-25
e. Too familiar/not formal enough (negative-P)	1 (1.4%)	-	9 (12.3%)	7 (10.4%)	*It was as if some clever old man were speaking.* [note: ironic, negative tone]	2	LF-Sch-P-m-18
					I didn't like the texts being in first person.	2	LF-Sch-P-m-18
					It would be good if the author did not try to make funny comments.	2	WT-Sch-P-f-17
					I didn't like the informal style and the fact information was delivered as a story.	1	LF-Col-P-f-20
					Too friendly/familiar.	2	LF-Sch-P-m-18
f. Excessive curtness	2 (2.7%)	2 (3.1%)	-	-	*...perhaps too textbook-ish.*	4	LF-Sch-F-f-17

(negative-F)							
g. Superfluous (negative-P)	-	-	5 (6.8%)	3 (4.5%)	*The texts contain a large amount of superfluous constructions for my taste, such as: "Imagine that…", "I feel cold..."*	1	LF-Col-P-m-23
					I would avoid "funny comments" used to lighten the teaching process, it comes across as embarrassing and it is not necessary for teaching high school students.	1	LF-Sch-P-f-17
h. Too complex (negative-Mixed)	6 (8.1%)	2 (3.1%)	1 (1.4%)	2 (3.0%)	*…very difficult for people, who are not familiar with the topic.*	4	WT-Sch-F-f-18
i. Specific praise (positive-P)	1 (1.4%)	1 (1.6%)	10 (13.7%)	12 (17.9%)	*Thanks to the story it was easier to understand the animation.*	3	LF-Col-P-f-19
					I liked the less formal language. It seemed interesting.	3	LF-Col-P-m-23
					…as if it were a fairy-tale; very informal.	4	LF-Col-P-f-20
					The texts were cute, funny. They amused me.	3	WT-Col-P-m-20
					I liked the texts' "personal" approach.	4	LF-Sch-P-f-17
					It was good that we were, in that moment, taken to the spot where the action occurs (the cloud). I could immediately envision it better.	2	LF-Sch-P-f-18
					I liked the … familiarity of the text ("And now I will tell you how the story ends.")	3	WT-Col-P-f-23

[a]Most participants made one or no statement, some participants made two statements.

[b]1: I would definitely prefer more formal language; 2: I would rather prefer more formal language; 3 – I was fine with the version of the texts used in the animation; 4 – I would rather prefer less formal language.

[c]LF: lightning formation animation, WT: wastewater treatment plant animation; Col: college, Sch: high school; P: personalized, F: formal; f: female, m: male; the number codes for age.

[d]P: the category contained statements from participants, of which over 90% were from the P group; F: the category contained statements from participants, of which over 90% were from the F group; Mixed: mixed group statements category (at least 25% of statements from either of the groups).

Table 9

Means and SDs for preferences regarding the tutor statements, including the original results by Mayer et al. (2006)

Statement	Ratings							
	Present Experiment 5				Mayer et al. (2006)			
	High-school		College		Negative politeness		Positive politeness	
	M[a]	*SD*	*M*[a]	*SD*	*M*[b]	*SD*	*M*[b]	*SD*
Press the Enter button. (formal)	4.05	1.00	3.78	1.21	1.50	0.71	2.02	1.20
Press the Enter button. (informal)	3.16	1.24	3.58	1.29				
I would like you to press the Enter button. (formal)	1.76	0.76	1.59	0.85	2.26	0.81	2.90	1.13
I would like you to press the Enter button. (informal)	1.54	0.74	1.50	0.69				
Do you want to press the Enter button? (formal)	1.84	1.17	1.65	0.92	4.23	0.82	3.07	1.12
Do you want to press the Enter button? (informal)	1.57	0.80	1.42	0.71				
I would now click the Enter button. (formal)	2.16	1.48	2.50	1.37	2.45	0.80	2.55	1.21
I would now click the Enter button (informal).	1.89	1.17	1.90	1.08				
Use the worked-out example to solve the equation. (formal)	3.97	0.96	3.98	1.06	1.53	0.81	2.39	1.24
Use the worked-out example to solve the equation. (informal)	3.46	1.24	3.66	1.25				
I would like you to use the worked-out example to solve the equation. (formal)[c]	2.16	1.09	2.09	1.12	2.86	1.01	3.53	1.04
I would like you to use the worked-out example to solve the equation. (informal) [c]	2.03	1.01	1.86	0.98				
Do you want to use the worked-out example to solve the equation? (formal)[d]	1.46	0.69	1.63	0.97	3.07	1.12	3.29	0.81
Do you want to use the worked-out example to solve the equation? (informal)[d]	1.49	0.80	1.38	0.65				
I would use the worked-out example to solve this equation. (formal)	1.84	1.26	1.91	1.15	2.55	1.21	3.03	0.93
I would use the worked-out example to solve this equation. (informal)	1.81	1.24	1.60	0.86				

[a]Scale 1 – 5. Higher values mean "I prefer more".

[b]Scale 1 – 5. Higher values mean "more polite". Original values (scale 1 – 7) were rescaled to match our scale 1 – 5.

[c] The original formulation was “I suggest that you use...” and Mayer and colleagues intended it to express a similar level of politeness as the formulation “I would like you to...”.

[d] The original formulation was “Did you use...” and Mayer and colleagues intended it to express a similar level of politeness as the formulation “Do you want to...”. However, the “did you” formulation turned out to be less polite than the “do you” formulation (Mayer et al., 2006; p. 40).

Table E1

Correlations between measures (Experiment 1)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Learning	-							
(2) Usefulness	.30*	-						
(3) Interest	.12	.51***	-					
(4) Motivation	-.12	.43***	.37**	-				
(5) Difficulty	-.39**	-.01	.03	.28*	-			
(6) Friendliness	-.20	.51***	.52***	.50***	.17	-		
(7) Retention	.29*	-.12	.03	-.09	**-.38****	-.06	-	
(8) Transfer	.21	-.11	-.20	-.13	**-.44*****	-.15	.36**	-

**p* < .05. ***p* < .01. ****p* < .001 (without correction for multiple comparisons).

Table E2

Correlations between measures (Experiment 2)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Learning	-							
(2) Usefulness	.42***	-						
(3) Interest	.47***	.56***	-					
(4) Motivation	.34**	.43***	.56***	-				
(5) Difficulty	-.34**	-.22†	-.19	-.31**	-			
(6) Friendliness	.32	.29*	.56***	.53***	-.26*	-		
(7) Retention	.18	.29*	.19	.25*	**-.13**	.01	-	
(8) Transfer	.19	.04	.12	.05	**.03**	-.02	.32**	-

†*p* < .10. **p* < .05. ***p* < .01. ****p* < .001 (without correction for multiple comparisons).

Table E3

Correlations between measures (Experiment 3)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(1) Learning	-																
(2) Usefulness	.55***	-															
(3) Interest	.42***	.48***	-														
(4) Motivation	.32**	.30*	.36**	-													
(5) Difficulty	-.32**	-.43***	-.41***	-.18	-												
(6) Friendliness	.20†	.14	.59***	.28*	-.12	-											
(7) IM:interest	.08	.16	.24†	.20†	-.07	.12	-										
(8) IM:anxiety	-.08	.07	-.34**	-.19	-.12	-.16	.07	-									
(9) PANAS1+	.00	-.02	.16	.23†	.06	.10	.49***	.14	-								
(10) PANAS1-	.02	-.20†	-.08	.11	.29*	-.13	-.16	-.57***	-.03	-							
(11) PANAS2+	.18	.08	.35**	.27*	-.06	.15	.41***	.03	.79***	.05	-						
(12) PANAS2-	-.17	-.26*	-.09	.10	.34**	-.08	.08	-.42***	.02	.67***	.05	-					
(13) Flow	.31**	.20†	.32**	.07	**-.34****	.12	.41***	.20†	.19	-.38***	.27*	-.33**	-				
(14) PANAS+ diff	.29*	.16	.30*	.07	-.13	.06	-.09	-.16	-.22†	.12	.43***	.03	**.16**	-			
(15) PANAS- diff	-.23*	-.04	-.01	-.02	.02	.07	.29*	.26*	.07	-.52***	-.01	.30*	.11	-.11	-		
(16) Retention	.15	.04	.16	-.15	**-.32****	.11	-.07	-.07	-.29*	-.08	-.14	-.20†	**.25***	**.19**	-.13	-	
(17) Transfer	.03	.05	.08	-.03	**-.27***	.00	.08	.34**	.03	-.26*	.04	-.20†	**.36****	.00	.10	.39***	-

†*p* < .10. **p* < .05. ***p* < .01. ****p* < .001 (without correction for multiple comparisons).

Table E4

Correlations between measures (Experiment 4)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(1) Learning	-																
(2) Usefulness	.48***	-															
(3) Interest	.22†	.39***	-														
(4) Motivation	.29*	.19	.31**	-													
(5) Difficulty	-.12	-.06	-.18	.04	-												
(6) Friendliness	.12	.41***	.27*	.39***	-.05	-											
(7) IM:interest	.18	.39**	.64***	.23†	-.21†	.13	-										
(8) IM:anxiety	-.09	-.04	-.06	.04	-.13	.08	-.07	-									
(9) PANAS1+	-.01	.22†	.37**	.05	-.26*	.17	.52***	.11	-								
(10)PANAS1-	-.09	-.01	.14	.13	.30*	.04	.13	-.59***	.01	-							
(11) PANAS2+	.23†	.39***	.49***	.20	-.20†	.21†	.53***	-.02	.82***	.14	-						
(12) PANAS2-	-.04	-.02	-.03	.14	.38***	.17	-.02	-.25*	.06	.75***	.13	-					
(13) Flow	.36**	.45***	.43***	.12	**-.51*****	.17	.44***	.08	.36**	-.18	.48***	-.22†	-				
(14) PANAS+ diff	.40***	.32**	.29*	.26*	.07	.09	.12	-.23†	-.15	.22†	.44***	.13	**.23†**	-			
(15) PANAS- diff	.05	-.05	-.26*	-.04	.07	.13	-.22†	.53***	.07	-.52***	-.03	.17	-.06	-.15	-		
(16) Retention	.01	.10	.14	-.06	**-.24***	.09	-.04	-.02	-.12	-.08	.05	-.18	**.06**	**.28***	-.11	-	
(17) Transfer	.11	.21†	.11	-.18	**-.36****	.13	-.04	.17	.14	-.23*	.18	-.19	**.27***	.10	.07	.63***	-

†*p* < .10. **p* < .05. ***p* < .01. ****p* < .001 (without correction for multiple comparisons).

Table F1

Two three-way ANOVAs (from Section 6.2.2) for transfer/retention as a dependent variable and the language style of texts (formal vs. personalized), animation type (lightning formation vs. wastewater treatment), and school level (high school vs college) as factors

Test	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer	language style	0.29	1,270	0.00	[0.00, 0.01]	-
	animation type	0.10	1,270	0.00	[0.00, 0.00]	-
	school level	31.92***	1,270	0.11	[0.04, 0.18]	college > high school
	language style:animation type	0.08	1,270	0.00	[0.00, 0.00]	-
	language style:school level	5.31*	1,270	0.02	[0.00, 0.06]	college: P < F; high school: P > F
	animation type:school level	0.58	1,270	0.00	[0.00, 0.02]	-
	language style:animation type:school level	1.90	1,270	0.01	[0.00, 0.04]	-
Retention	language style	0.12	1,270	0.00	[0.00, 0.01]	-
	animation type	376.98***	1,270	0.58	[0.49, 0.65]	wastewater > lighting
	school level	13.19***	1,270	0.05	[0.01, 0.10]	college > high school
	language style:animation type	0.06	1,270	0.00	[0.00, 0.00]	-
	language style:school level	0.69	1,270	0.00	[0.00, 0.02]	-
	animation type:school level	7.07**	1,270	0.03	[0.00, 0.07]	wastewater: college >> high school; lighting: college > high school
	language style:animation type: school level	0.00	1,270	0.00	[0.00, 0.00]	-

**p* < .05. ***p* < .01. ****p* < .001.

Table F2

Six one-way ANCOVAs (from Section 6.2.3) for transfer/retention as a dependent variable, the language style of texts (formal vs. personalized) as a factor, and a pretest score as a covariate (there were two pretests for the lighting formation animation). Interactions are also included.

Test	Animation type	Pretest type	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer	Wastewater treatment	-	lang_style	0.13	1,139	0.00	[0, 0.01]	-
			pretest	0.99	1,139	0.01	[0, 0.07]	-
			lang_style:pretest	0.29	1,139	0.00	[0, 0.03]	-
	Lighting formation	Meteorology	lang_style	0.06	1,124	0.00	[0, 0.01]	-
			pretest_meteo	0.10	1,124	0.00	[0, 0.01]	-
			lang_style:pretest_meteo	0.35	1,124	0.00	[0, 0.03]	-
	Lighting formation	Electricity	lang_style	0.00	1,121	0.00	[0, 0.00]	-
			pretest_electro	0.51	1,121	0.00	[0, 0.05]	-
			lang_style:pretest_electro	0.02	1,121	0.00	[0, 0.00]	-
Retention	Wastewater treatment	-	lang_style	0.01	1,139	0.00	[0, 0.00]	-
			pretest	0.22	1,139	0.00	[0, 0.02]	-
			lang_style:pretest	0.00	1,139	0.00	[0, 0.00]	-
	Lighting formation	Meteorology	lang_style	0.10	1,124	0.00	[0, 0.01]	-
			pretest_meteo	0.96	1,124	0.01	[0, 0.06]	-
			lang_style:pretest_meteo	1.39	1,124	0.01	[0, 0.06]	-
	Lighting formation	Electricity	lang_style	0.03	1,121	0.00	[0, 0.00]	-
			pretest_electro	0.61	1,121	0.00	[0, 0.05]	-
			lang_style:pretest_electro	0.00	1,121	0.00	[0, 0.00]	-

Table F3

Six two-way ANCOVAs (from Section 6.2.3) for transfer/retention as a dependent variable, the language style of texts (formal vs. personalized) and school level (high school vs. college) as factors, and a pretest score as a covariate (there were two pretests for the lighting formation animation). Interactions are also included.

Test	Animation type	Pretest type	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer	Wastewater treatment	-	lang_style	0.20	1,135	0.00	[0, 0.02]	-
			sch_level	20.61***	1,135	0.13	[0.04, 0.26]	college > high school
			pretest	3.42†	1,135	0.02	[0, 0.12]	pretest predicts transfer
			lang_style:sch_level	0.71	1,135	0.01	[0, 0.05]	-
			lang_style:pretest	0.25	1,135	0.00	[0, 0.02]	-
			sch_level:pretest	0.26	1,135	0.00	[0, 0.02]	-
			lang_style:sch_level:pretest	1.30	1,135	0.01	[0, 0.07]	-
	Lighting formation	Meteorology	lang_style	0.00	1,120	0.00	[0, 0.00]	-
			sch_level	11.42**	1,120	0.09	[0.01, 0.21]	college > high school
			pretest_meteo	0.40	1,120	0.00	[0, 0.04]	-
			lang_style:sch_level	7.30**	1,120	0.06	[0, 0.15]	college: F > P; high school: P > F
			lang_style:pretest_meteo	2.08	1,120	0.02	[0, 0.09]	-
			sch_level:pretest_meteo	1.80	1,120	0.01	[0, 0.08]	-
			lang_style:sch_level:pretest_meteo	0.44	1,120	0.00	[0, 0.04]	-
	Lighting formation	Electricity	lang_style	0.00	1,117	0.00	[0, 0.00]	-
			sch_level	14.58***	1,117	0.11	[0.02, 0.25]	college > high school
			pretest_electro	3.13†	1,117	0.03	[0, 0.12]	pretest predicts transfer
			lang_style:sch_level	7.18**	1,117	0.06	[0, 0.17]	college: F > P; high school: P > F
			lang_style:pretest_electro	0.41	1,117	0.00	[0, 0.04]	-
			sch_level:pretest_electro	8.52**	1,117	0.07	[0.01, 0.18]	pretest predicts transfer better for college than high school learners
			lang_style:sch_level:pretest_electro	0.30	1,117	0.00	[0, 0.03]	-

Retention	Wastewater treatment	-	lang_style	0.01	1,135	0.00	[0, 0.00]	-
			sch_level	12.37**	1,135	0.08	[0.01, 0.18]	college > high school
			pretest	1.23	1,135	0.01	[0, 0.07]	-
			lang_style:sch_level	0.53	1,135	0.00	[0, 0.03]	-
			lang_style:pretest	0.00	1,135	0.00	[0, 0.00]	-
			sch_level:pretest	0.34	1,135	0.00	[0, 0.03]	-
			lang_style:sch_level:pretest	3.96*	1,135	0.03	[0, 0.11]	pretest predicts retention for college sample in the F condition (*r* = .21, *p* = .11) and high school sample in the P condition (*r* = .29, *p* = .08), but not in the other cells (*r* = -.05, -.12, n.s.) (i.e., this does not correspond to an expertise reversal effect)
	Lighting formation	Meteorology	lang_style	0.05	1,120	0.00	[0, 0.00]	-
			sch_level	2.61	1,120	0.02	[0, 0.09]	-
			pretest_meteo	1.81	1,120	0.01	[0, 0.08]	-
			lang_style:sch_level	0.75	1,120	0.01	[0, 0.06]	-
			lang_style:pretest_meteo	0.84	1,120	0.01	[0, 0.06]	-
			sch_level:pretest_meteo	0.25	1,120	0.00	[0, 0.02]	-
			lang_style:sch_level:pretest_meteo	1.02	1,120	0.01	[0, 0.06]	-
	Lighting formation	Electricity	lang_style	0.03	1,117	0.00	[0, 0.00]	-
			sch_level	1.71	1,117	0.01	[0, 0.09]	-
			pretest_electro	1.22	1,117	0.01	[0, 0.08]	-
			lang_style:sch_level	1.65	1,117	0.01	[0, 0.08]	-
			lang_style:pretest_electro	0.16	1,117	0.00	[0, 0.02]	-
			sch_level:pretest_electro	1.57	1,117	0.01	[0, 0.08]	-
			lang_style:sch_level:pretest_electro	1.69	1,117	0.01	[0, 0.08]	-

†*p* < .10. **p* < .05. ***p* < .01. ****p* < .001.

Table F4

Four two-way ANOVAs (from Section 6.2.3) on subsamples of students with background in technical disciplines or in psychology. Dependent variables: transfer/retention; factors: the language style of texts (formal vs. personalized) and animation type (lightning formation vs. wastewater treatment).

Test	Participants' background	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer	Technical	language style	7.51*	1,26	0.22	[0.01, 0.48]	F > P
		animation type	0.40	1,26	0.02	[0, 0.15]	-
		language style:animation type	0.12	1,26	0.00	[0, 0.05]	-
	Psychology	language style	0.61	1,72	0.01	[0, 0.07]	-
		animation type	0.00	1,72	0.00	[0, 0.00]	-
		language style:animation type	0.59	1,72	0.01	[0, 0.08]	-
Retention	Technical	language style	0.22	1,26	0.01	[0, 0.09]	-
		animation type	39.12***	1,26	0.60	[0.36, 0.75]	wastewater > lighting
		language style:animation type	0.13	1,26	0.00	[0, 0.07]	-
	Psychology	language style	0.52	1,72	0.01	[0, 0.07]	-
		animation type	119.70***	1,72	0.62	[0.48, 0.74]	wastewater > lighting
		language style:animation type	0.04	1,72	0.00	[0, 0.00]	-

*$p < .05$. ***$p < .001$.

Table F5

Two two-way ANCOVAs (from Section 6.2.6) for transfer/retention as a dependent variable, animation type (lightning formation vs. wastewater treatment) and school level (high school vs. college) as factors, and perceived difficulty as a covariate. Interactions are also included.

Test	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer	perceived difficulty	16.26***	1,270	0.06	[0.01, 0.12]	perceived difficulty predicts transfer
	animation type	2.07	1,270	0.01	[0, 0.04]	-
	school level	29.84***	1,270	0.10	[0.04, 0.17]	college > high school
	perceived difficulty:animation type	2.41	1,270	0.01	[0, 0.04]	-
	perceived difficulty:school level	5.39*	1,270	0.02	[0, 0.06]	perceived difficulty predicts transfer better for college than high school learners (see Tables E2 and E4)
	animation type:school level	0.01	1,270	0.00	[0, 0.00]	-
	perceived difficulty:animation type:school level	4.15*	1,270	0.02	[0, 0.05]	perceived difficulty does not predict transfer for the lighting animation in the case of high school learners (Exp. 2; see Table E2).
Retention	perceived difficulty	15.17***	1,270	0.05	[0.02, 0.12]	perceived difficulty predicts retention
	animation type	342.30***	1,270	0.56	[0.47, 0.63]	wastewater > lighting (i.e., different scales for retention test scores)
	school level	11.04**	1,270	0.04	[0.01, 0.09]	college > high school
	perceived difficulty:animation type	5.78*	1,270	0.02	[0, 0.07]	perceived difficulty predicts retention slightly better for the wastewater animation than for the lightning animation
	perceived difficulty:school level	1.41	1,270	0.01	[0, 0.03]	-
	animation type:school level	4.08*	1,270	0.01	[0, 0.05]	wastewater: college >> high school lighting: college > high school
	perceived difficulty:animation type:school level	0.00	1,270	0.00	[0, 0.00]	-

**p* < .05. ***p* < .01. ****p* < .001.

Table F6

Four one-way ANCOVAs (from Section 6.2.6) for transfer/retention as a dependent variable, school level (high school vs. college) as a factor, and flow/positive affect as a covariate. Interactions are also included.

Test	Covariate	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer	Flow	flow	15.51***	1,141	0.10	[0.02, 0.20]	flow predicts transfer
		school level	18.43***	1,141	0.12	[0.03, 0.21]	college > high school
		flow:school level	1.16	1,141	0.01	[0, 0.05]	-
	Panas+	positive affect	0.28	1,139	0.00	[0, 0.02]	-
		school level	19.26***	1,139	0.12	[0.04, 0.24]	college > high school
		positive affect:school level	0.27	1,139	0.00	[0, 0.02]	-
Retention	Flow	flow	3.11†	1,141	0.02	[0, 0.09]	flow predicts retention
		school level	11.46**	1,141	0.08	[0.02, 0.17]	college > high school
		flow:school level	1.29	1,141	0.01	[0, 0.07]	-
	Panas+	positive affect	8.10**	1,139	0.06	[0, 0.14]	panas+ predicts retention
		school level	8.75**	1,139	0.06	[0.01, 0.15]	college > high school
		positive affect:school level	0.66	1,139	0.00	[0, 0.04]	-

†*p* < .10. ** *p* < .01. ****p* < .001.

Table F7

Two three-way ANOVAs (from Section 6.2.7) for transfer test or perceived difficulty as a dependent variable and the language style preference (more formal vs. less formal), animation type (lightning formation vs. wastewater treatment), and school level (high school vs. college) as factors

Outcome variable	Term	*F*	*df*	η_p^2	η_p^2 95 % CI	Description
Transfer test	language pref.	4.05*	1,62	0.06	[0, 0.21]	more formal > less formal
	animation type	0.09	1,62	0.00	[0, 0.02]	-
	school level	0.76	1,62	0.01	[0, 0.10]	-
	language pref.:animation type	3.92†	1,62	0.06	[0, 0.18]	lighting: more formal > less formal wastewaster: more formal >> less formal
	language pref.:school level	0.02	1,62	0.00	[0, 0.00]	-
	animation type:school level	0.03	1,62	0.00	[0, 0.00]	-
	language pref.:animation type:school level	0.32	1,62	0.01	[0, 0.05]	-
Perceived difficulty	language pref.	14.05***	1,62	0.18	[0.02, 0.39]	less formal > more formal
	animation type	2.29	1,62	0.04	[0, 0.15]	-
	school level	0.09	1,62	0.00	[0, 0.02]	-
	language pref.:animation type	0.35	1,62	0.01	[0, 0.06]	-
	language pref.:school level	0.76	1,62	0.01	[0, 0.11]	-
	animation type:school level	0.88	1,62	0.01	[0, 0.11]	-
	language pref.:animation type:school level	1.00	1,62	0.02	[0, 0.13]	-

†$p < .10$. ** $p < .05$. ***$p < .001$

Appendices

Appendix A

Perceived prior knowledge questionnaires ("Pretests")

Meteorology

1. Please, indicate your knowledge of meteorology on a scale of 1 (very good) to 6 (very weak).
2. Please, indicate what is TRUE in your case, on a scale of 1 to 4.
   - I regularly read weather maps in newspapers or on the internet (1 – never; 4 – daily).
   - I can thoroughly explain to a high school student what a cold front is (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student the difference between cumulus and nimbus clouds (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student what makes the wind blow (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student what this symbols means [symbol for cold front] (1 – definitely not; 4 – definitely yes).

Electro-Physiology

1. Please, indicate your knowledge of meteorology on a scale of 1 (very good) to 6 (very weak).
2. Please, indicate what is TRUE in your case, on a scale of 1 to 4.
   - I can thoroughly explain to a high school student what a Faraday cage is (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student what an ion is (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student what an electric arc is (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student how a capacitor works (1 – definitely not; 4 – definitely yes).
   - I can thoroughly explain to a high school student why there is a ground hole in an electrical outlet (1 – definitely not; 4 – definitely yes).

Biological Wastewater Treatment

1. Please, indicate your knowledge of biological wastewater treatment on a scale of 1 (very good) to 6 (very weak).

2. Please make a check mark next to each sentence that is TRUE in your case.
    - I can thoroughly explain to a high school student what an oxidation-reduction reaction is.
    - I have been on an excursion to a wastewater treatment plant in the past.
    - I can thoroughly explain to a high school student how aerobic bacteria relate to wastewater treatment.
    - We learned about the topic of biological wastewater treatment in school (for at least one full class hour).
    - I know if a wood-decay fungus can live in water and I can explain why.
    - I can thoroughly explain to a high school student what bacterial film is.
    - I can thoroughly explain to a high school student what diazotization is.
    - I know what pH values acids and bases have.
3. Have you ever learned about the topic of wastewater treatment (even partially)? If so, when and where? ....

Question 1 was awarded 0 (very weak) to 5 (very good) points. Each item checked in Question 2 was awarded two points. An additional maximum 4 points could be awarded for answering Question 3 (provided the answer was not a duplicate response to an answer from Question 2).

Appendix B

Feedback Questionnaire

The questions were adjusted accordingly for the wastewater treatment plant animation.

- Learning 1: After watching the animation, how would you rate your knowledge of the process by which lightning is created? (*1 – very good*; *8 – very weak*).
- Learning 2: Do you feel that you learned something today about the process by which lightning is created? (*1 – very much*; *8 – very little*).
- Usefulness: How useful was the personal animation that you saw today in learning about the process by which lightning is created? (*1 – very much*; *8 – very little*).
- Interest 1: How interesting was today's lesson on the process by which lightning is created (without filling in the questionnaires and tests)? (*1 – very much*; *8 – very little*).
- Interest 2: How much fun for you was today's lesson on the process by which lightning is created (without filling in questionnaires and tests)? (*1 – very much*; *8 – very little*).
- Difficulty: How difficult was it for you to learn from today's animation on the process by which lightning is created? (*1 – very much*; *8 – very little*).
- Motivation: Would you gladly use similar animations (with different content) in your studies? (*1 – definitely yes*; *8 – definitely no*).
- Friendliness: Was the animation friendly for you? (*1 – very much*; *8 – very little*).
- Texts: What do you think about the texts you read in the animation? (*open-ended*)
- Style preference: Would you prefer that the texts be written in a more formal or less formal language? (*1 – I would definitely prefer more formal language*; *2 – I would rather prefer more formal language*; *3 – I was fine with the version of the texts used in the animation*; *4 – I would rather prefer less formal language*; *5 – I would definitely prefer less formal language*)

Appendix C

Animations' Instructions

Personalized additions are inserted in [brackets], deletions are ~~crossed out~~.

Lightning Formation – English

These expository texts are from Moreno & Mayer (2000; Appendix A).

"[Let me tell you what happens when lightning forms. Suppose you are standing outside, feeling the warm rays of the sun heating up the earth's surface around you.] Cool moist air moves over a warmer surface and becomes heated. The warmed moist air near the earth's surface rises rapidly. As the air in this updraft cools, water vapor condenses into water droplets and forms a cloud. [Congratulations! You have just witnessed the birth of your own cloud!] [As you watch, you tilt your head skyward. Your] ~~The~~ cloud's top extends above the freezing level, so the upper portion of [your] ~~the~~ cloud is composed of tiny ice crystals. [Brr! I'm feeling cold just thinking about it!] Eventually, the water droplets and ice crystals become too large to be suspended by updrafts. As raindrops and ice crystals fall through [your] ~~the~~ cloud, they drag some of the air in [your] ~~the~~ cloud downward, producing downdrafts. When downdrafts strike the ground, they spread out in all directions, producing the gusts of cool wind [you] ~~people~~ feel just before the start of the rain. [If you could look inside your cloud, you could see a neat pattern:] ~~Within the cloud,~~ the rising and falling air currents cause electrical charges to build. The negatively charged particles fall to the bottom of the cloud, and most of the positively charged particles rise to the top. [Now that your cloud is charged up, I can tell you the rest of the story:] A stepped leader of negative charges moves downward in a series of steps. It nears the ground. A positively charged leader travels up from objects [around you] such as trees and buildings. The two leaders generally meet about 165 feet above the ground. Negatively charged particles then rush from [your] ~~the~~ cloud to the ground along the path created by the leaders. It is not very bright. As the leader stroke nears the ground, it induces an opposite charge, so positively charged particles from the ground rush upward along the same path. This upward motion of the current is the return stroke. It produces the bright light that [you] ~~people~~ notice as a flash of lightning."

Biological Wastewater Treatment Plant – English

[Now I will tell you,] ~~Now it will be explained,~~ how biological wastewater treatment works. [Imagine you are standing on the bank of a muddy river. Next to the river is a] ~~A~~ textile mill [that] dyes fabrics and releases azo dyes into the river. Azo dyes are slightly toxic for aquatic organisms and can cause mutation. [The town where you live is near the factory.] The town also dumps its sewage into the river. This waste contains nutrients. [Look into the river]. The nutrients cause blue-green algae and seaweed to grow. Water filled with blue-green algae is not conducive to life for aquatic organisms. [In order for us to clean the water and remove] ~~It order to clean the water and get rid of~~ both the azo dyes and the nutrients, [we will build] ~~it is necessary to build~~ a water treatment plant. [Your] ~~The~~ water treatment plant [will have] ~~has~~ two sections. In the first section there [will be] ~~are~~ bacteria. The bacteria are adapted to a specific composition of the wastewater (temperature, nutrient content, pH). Were the composition of the wastewater to change, the bacteria would stop working or would die. [Were we to live among the bacteria, you would see how they] ~~The bacteria~~ consume the nutrients. That is how we use them to clean the wastewater. But bacteria can "stuff" themselves on nutrients and stop working. In such a case ~~it is~~ [you must] ~~necessary to~~ take them out of the purifier and starve them. On the other hand, should the bacteria run out of nutrients, they will feed on the azo dyes. The by-product of their metabolism is thereafter aromatic compounds. Aromatic compounds are extremely toxic for aquatic organisms. [Brr! I get the chills just thinking about it! (I will remind you azo] ~~(Azo~~ dyes were just slightly toxic.) [In order for us to] ~~In order to~~ remove the dangerous aromatic compounds and possible azo dyes, [we put] ~~there are~~ fungi in the second section of the treatment plant. Fungi require a different pH than the one found in wastewater to do their work. Wastewater is basic (pH 9); fungi need an acidic environment (pH 5). [You can change the] ~~The~~ pH in wastewater ~~is changed~~ by adding a certain amount of acid. This reduces the water's pH and the water becomes acidic. Provided the fungi have sufficient random food, their metabolism corrects the water's pH to neutral (pH 7). Were the water to remain acidic that would bother the aquatic organisms. [And now that both parts of your treatment plant have been built, I will tell you the rest of the story.] If the fungi have enough nutrients, they will consume them. It is only when the fungi do not have nutrients, because the bacteria consumed them in the first section of the treatment plant, that the fungi live on both azo dyes and aromatic compounds. If everything has worked correctly, the water that passes out of the treatment plant is clean. It contains no nutrients, no azo dyes and no aromatic compounds. And it has the right pH. [The river in front of you is crystal clear and aquatic] ~~Aquatic~~ organisms do well ~~in such water~~ [in it]. And that's all.

Appendix D

Initial Motivation Questionnaire

The following items are according to Vollmeyer & Rheinberg (2006). Every item had 7-point Likert scale *disagree – agree*. The original dimensions are: (A) – anxiety; (P) – probability of success; (I) – interest; (C) – challenge.

Anxiety:

- When I think about the task, I feel somewhat concerned. (A)
- I'm afraid I will make a fool out of myself. (A)
- I think I won't do well at the task. (P) (reverse-coded)

Interest:

- The topic of biological waste-water treatment seems to be very interesting to me. (I)
- I am eager to see how I will perform in the today's task. (C)
- I'm really going to try as hard as I can on this task. (C)
- While doing this task I will enjoy discovering how a biological waste-water treatment plan works. (I)
- I would work on this task even in my free time (if I have the instructional animation). (I)

Appendix E

Correlational Analysis

We report correlations between affective variables and test scores in Tables E1 –E4. Correlations between affective variables and test scores are generally instable. However, some correlations between affective variables themselves are relatively stable and in medium to high range, which indicates that they likely tap common underlying constructs (e.g., interest × usefulness, usefulness × learning, motivation × interest). Correlations discussed in the main text are in boldface.

--- Insert Tab. E1 around here ---

--- Insert Tab. E2 around here ---

--- Insert Tab. E3 around here ---

--- Insert Tab. E4 around here ---

Appendix F

For the sake of completeness, we report ANOVAs and ANCOVAs from Section 6.2 in full detail (Tables F1 – F6). Significant findings are interpreted in the "Description" columns.

--- Insert Table F1 around here ---

--- Insert Table F2 around here ---

--- Insert Table F3 around here ---

--- Insert Table F4 around here ---

--- Insert Table F5 around here ---

--- Insert Table F6 around here ---