On the Robustness of the Quizzing Effect under Real Teaching Conditions

Über die Robustheit des Abfrage-Effektes unter realen Unterrichtsbedingungen

Tilmann Betsch, Nancy Quittenbaum & Manfred Lüders

University of Erfurt

Tilmann Betsch
Department of Psychology
University of Erfurt
Nordhäuser Strasse 63
D-99089 Erfurt
Phone: +49-(0) 361 / 737 2221
E-mail: tilmann.betsch@uni-erfurt.de
Abstract

Lab experiments and field studies showed that studying improves short-term memory, whereas active retrieval improves long-term memory – a phenomenon known as the quizzing (or testing) effect. In a quasi-experimental field study with four elementary school classes (18 < n < 22) and a pretest – posttest (5 min after the intervention) – posttest (6 weeks later) design, we tested the robustness of the quizzing effect under real conditions involving verbal teacher-student interactions in geometry lessons on symmetry in elementary schoolers. Results showed that re-studying compared to active retrieval (quizzing) enhanced learning when measured directly after the lessons. This pattern revered when knowledge was measured six weeks later, demonstrating that the quizzing effect was robust. Moreover, long-term memory generally increased after six weeks. Limitations of the quasi-experimental approach are discussed.

Keywords: Testing effect, didactics of mathematics, direct instruction

Zusammenfassung


Schlüsselwörter: Test-Effekt, Didaktik der Mathematik, Direkte Instruktion
1. Introduction

A typical learning strategy of students is to repeat the material to be studied. For instance, when learning German vocabulary, a student might invest some time in repeating pairs of English and German words (e.g., “table – Tisch”). It is a fundamental assumption in associative learning that the joint activation of stimuli increases their association in memory. Successful retrieval of “Tisch” upon activation of the English word “table” should be a function of the number of instances in which the pair of words was repeated. Studying by merely repeating the material does indeed have a beneficial effect on retention, but mostly in the short term (Roediger & Karpicke, 2006).

An alternative strategy is quizzing (or testing), which, compared to mere studying, is cognitively more effortful, as it requires active memory search. To apply this strategy, the student may cover the German word, read the English stimulus word and attempt to recall its translation (e.g., “table – ?”). During the elaboration process, alternate retrieval routes can be established, thus promoting retrieval in the long term (Roediger & Karpicke, 2006).

Results from lab experiments show that quizzing outperforms re-studying with regard to long-term memory (Roediger & Butler, 2011, for an overview). In 2007, the U.S. Department of Education and the National Center for Education Research published a list of seven instructional techniques based on empirical evidence (Pashler et al., 2007). Quizzing was included in the list of recommendations for effective teaching methods, but included the caveat that the external validity of quizzing still awaits evaluation.

Most of the evidence stems from highly controlled lab settings that focused on the learning of declarative knowledge, which requires no deeper insight or understanding of procedural aspects. Moreover, researchers typically used university or college students as participants in the lab. It is an open question if quizzing effects are prevalent across knowledge domains, age groups, and learning conditions (cf., Pashler, et al., 2007; Carpenter et al., 2008). So far, only few researchers have studied quizzing in school with children below college age (e.g., Carpenter et al., 2009: eighth-graders, history; McDaniel et al., 2011: eighth-graders, science class; Roediger...
et al., 2011: sixth-graders, social studies). Most importantly, the quizzing manipulation in these studies was not embedded in a teacher’s intervention. McDaniel and colleagues, for example, manipulated quizzing and re-studying in a questionnaire (called a “quiz”) that was administered to students before and after a teaching lesson. Thus, re-studying and quizzing instructions were not given within a real teacher-student interaction but rather were supplementary to study units.

The aim of this research was to test the robustness of the quizzing effect in a real classroom setting, i.e., whether the effect is strong enough to survive in the field (Pashler et al., 2007, p. 19).

To this end, we subjected four classes to four experimental conditions rather than randomizing individuals. This procedure surely has its limitation. In the future it might be complemented by costlier designs using a larger sample of classes and allowing for multi-level analyses to disentangle individual from class effects. Thus the present study should be just considered a first step into this direction.

We studied the effects of quizzing on geometry learning. Prior research typically considered the learning of declarative knowledge. Geometry, however, is a more complex subject. Students must not only memorize concepts (e.g., the definition of symmetry) but also use this knowledge to categorize new objects (e.g., identifying symmetric objects) and apply production rules (e.g., draw a symmetry axis). As another novel aspect, we assessed quizzing effects at an earlier age (third-graders). Studying children, and especially at an early school age, is important because the effectiveness of instructional techniques might substantially differ between age groups (e.g., Cepeda et al., 2006).

We varied whether teacher’s interventions and questions target only studying (S) of the material (three times: SSS) or involve quizzing (Q) after one studying episode (SQQ). In line with prior findings, we predicted an interaction effect between intervention type (SSS, SQQ) and time of measurement\(^1\). Specifically, memory and performance in transfer tasks should be better in SSS groups than in SQQ groups, if measures are taken directly after the intervention. This pattern should reverse (SQQ > SSS) after delay (6 weeks). For exploratory reasons, we also varied lag between repetitions of the instructional interventions (Donovan & Radosievich, 1999, for a review on distribution of practice). Memory can benefit from increasing lags between learning episodes (Cepeda et al., 2006).

\(^1\) This prediction arises from prior evidence rather than from a specific theory. There is an ongoing debate on the theoretical foundation of the quizzing effect.
2. Experiment

2.1. Participants and design

We recruited four third-grade classes (83 children) from an elementary school in a middle-class community in Erfurt, Germany. A total of eighty children took part in the experiment (n = 21, 20, 19, 20 per class). Drop outs were due to illness. Classes did not systematically differ with respect to age (M = 112.19 [SD = 4.63], 110.40 [5.52], 111.16 [5.69], 110.7 [5.01], GM = 111.13 months; 2-factorial GLM: all F < 1) and sex (girls: 47.6%, 45%, 47.4%, 40%; phi = -0.26, p > .80). All children were native German speakers. Classes were randomly assigned to one of four conditions resulting from a 2 (quizzing: no – SSS vs. yes – SQQ) by 2 (lag: 2/7 min.) between-subjects design.

2.2. Maximizing control under field conditions

We aimed to test the robustness of the quizzing effect under real teaching conditions. As a drawback of any field research, experimental control may be limited. We were not able to randomly assign participants to conditions, as this would have destroyed the natural social unit in which teaching acts are regularly embedded – the school class. In addition to potentially systematic error variance between classes (e.g., differences in distraction, noise), classes may systematically differ with respect to ability, prior knowledge, and learning routines. Although the school’s policy prevented us from collecting data on individual measures other than our focal test, the principal provided information on class mean math grades from the last term certificate, which indicated that mathematical skills ranged in the upper level and were quite similar (1.9, 2.1, 2.1, 2.0). We controlled for specific prior knowledge in the domain of geometry by (a) administering our test before and after the intervention to individually assess improvement and (b) starting the research at a time at which the specific lessons had not yet been taught in any of the classes. Moreover, teachers agreed to not repeat any of the specific content during the period until our final measure was taken.

The greatest challenge was to maximize control during the intervention so that class differences in learning routines and all sources of error would be minimized. To that end, we used a teaching method that allowed us to satisfy standards of
methodological rigor. Specifically, we employed the method of direct instruction (e.g., Rosenshine & Stevens, 1986), which provides rules for turn taking, presenting learning content, structuring subsequent procedures, posing questions, providing feedback, and encouraging student participation. All verbal and many gestural activities must be carefully planned and practiced in advance. Empirical evidence indicates that the method of direct instruction is an efficient means for introducing a new area of study material (Hattie, 2009). Due to its high level of standardization, it allowed us to systematically vary quizzing and practice under real classroom conditions.

2.3. Material

2.3.1. Scripted direct instruction

The entire instructional intervention followed a carefully designed and pre-tested script\(^2\) for the learning-unit “symmetry” that was based on the school’s mathematics curriculum. The script comprised eight episodes, (1) definition of symmetry, (2) axis of symmetry, (3) position of centre line, (4) multiple centre lines, (5) number and position of multiple centre lines, (6) symmetrical objects complement, (7) asymmetric objects, and (8) distinguishing between symmetric and asymmetric objects. The script was identical in each condition except the instructions for quizzing and studying. In the studying condition, after a target question (e.g., “What are such positions of center lines called?”), the experimenter showed cards with the written answer and asked children to read aloud the answers. For quizzing, after the target question, the experimenter had students recall the correct answer without showing the written card. In the case of an incorrect answer, she gave the correct feedback (cf., Roediger & Butler, 2010, for studies on quizzing with feedback).

In all conditions, the entire script with the eight episodes was run three times. The first run was identical in all four conditions. In the following two runs, the script was unchanged in the SSS conditions. In the SQQ conditions, however, the presentation of the information and the wording of the task were changed to quizzing. In some instances, individual students were selected to perform the reading and recall task, respectively; on other occasions, the task was performed by the entire class. The selection of individuals and joint class actions were carefully pre-scheduled and counterbalanced over episodes and repetitions.

2.3.2. Test

\(^2\) Script can be obtained from the authors.
The test was administered three times (before (t1), directly after (t2), and 6 weeks after (t3) the intervention). It contained only items on the topic of symmetry axis. They were taken from textbooks and teacher manuals used in German schools and pretested for difficulty in a sample of third- and fourth-graders. Two members of the university’s pedagogic faculty were consulted as experts to ensure that only those items representative for math tests in elementary schools were selected. The goal of pretesting was to avoid ceiling and floor effects. Specifically, in the first round before the intervention, children should be able to solve some tasks in order to avoid frustration effects. The test, however, should also involve items that were so difficult that variance in performance could still be detected after the learning episodes. The final test comprised six tasks: (1) a cloze test (8 words to be filled in such as “symmetric”, “vertical”, etc.), (2) eight objects to identify symmetry and the number of center lines, (3) two objects to be symmetrically completed on one side, (4) reproduction of three symmetric capital letters that have two axes of symmetry, (5) one object to be symmetrically completed on two sides, and (6) a single choice item (German translation of “symmetric”). As an example, Figure 1 depicts the one-side completion task.

2.4. Procedure

The experiment was implemented as a real teaching unit (90 minutes with a 5 minute break after 45 minutes) in a math course on geometry (symmetry). Teachers agreed not to use any material on the subject prior to the experiment and until the final test was taken.

Two experimenters, female graduate students of teacher education (age 22, 23), served as teacher and supervisor, respectively. Both had successfully taken part in a teacher training program for the method of direct instruction (Lüders, 2010) and were trained in the script. In addition, the first experimenter, who always served as the teacher, was carefully trained to enact all four versions of the script using feedback from members of the faculty and videotaped training sessions. The supervisor was seated at the back of the room during all class sessions. She ensured that the teacher adhered to the script and time schedule by giving nonverbal feedback that could not be viewed by the students. The feedback was presented on cards containing time measures, counts of repetitions, prompts, etc. and was given whenever necessary. As another measure of control, we videotaped all interventions to check post-hoc for potential deviations from the script.
Seven days before the intervention (t1), the experimenter visited the classes, introduced herself, and administered the test under the presence of the class teacher. The instructional interventions were scheduled during the morning in all classes. After the first implementation of the script (S-version in all conditions), there was a 2 or 7 minute break, depending on lag condition. During this period, the students practiced a relaxation exercise under the supervision of the experimenter’s assistant. The break was followed by the second implementation of the script (either S- or Q-version, depending on experimental condition). In the break thereafter (again lasting 2 or 7 minutes, respectively), students participated in games. Then, the script was implemented a third time (either S- or Q-version). Altogether, the three runs lasted between 68 and 72 minutes total (excluding breaks). Five minutes after the third run, the test was administered a second time (t2). Note that the test was exactly the same as the one children completed prior to the experiment. Six weeks after the intervention, each class received the same test a third time (t3).

3. Results

Table 1 shows mean test scores (max. 44 points) for the three testing times (before, 5 min, and 6 weeks after the intervention). We checked for differences in prior knowledge at t1. Importantly, SSS groups ($M_{SSS} = 10.58$, $SD = 6.21$) did not significantly differ from SQQ groups ($M_{SQQ} = 9.35$, $SD = 5.33$) in an ANOVA ($F(1,76) = 0.86$, $p = .35$, $\eta^2 = .01$). In the same analysis, the effect for lag also failed to reach a conventional level of significance ($F(1,76) = 1.53$, $p = .22$, $\eta^2 = .02$). There was a tendency towards an interaction effect ($F(1,76) = 2.98$, $p = .09$, $\eta^2 = .03$), reflecting the observation that one class score slightly higher on the test at t1. Note, however, that this class was not assigned to the quizzing condition. Moreover, the intraclass–correlation coefficient (ICC) was low ($rho = .02$) indicating that variance in knowledge is not substantially determined by class differences.

We predicted an interaction effect between quizzing and time. We tested this prediction twice. First, we performed GLM analyses on the original test scores (Table 2) with time as a three-level, repeated measurement factor. Second, we computed improvement scores to compensate for variance in prior knowledge. Specifically, for each individual we subtracted the test score at t1 from the subsequent test scores at time t2 and t3. We performed a second GLM analysis with time as a two-level, repeated measurement factor involving the improvement scores at t2 and t3 as
dependent variables. Lag and quizzing were entered as between-subjects factors. In both analyses, our prediction was statistically substantiated by a medium-sized interaction effect between time and quizzing \((F(2, 152) = 3.79, p < .05, \eta^2 = .05; F(1, 76) = 4.9, p < .05, \eta^2 = .06)\). Figure 2 depicts the interaction effect on improvement scores. Whereas SSS yields higher improvement than quizzing SQQ at t2, this effect reverses at t3.

In the two GLM analyses, time had a strong effect \((F(2, 152) = 171.03, p < .01, \eta^2 = .69; F(1, 76) = 15.1, p < .01, \eta^2 = .16)\), indicating that knowledge improves over time. On the one hand, this shows that the teaching intervention was effective. On the other hand, it is somewhat unexpected that knowledge again improves from t2 to t3 without intervention, as evidenced by the significant effect on improvement scores in the second analyses.

We also obtained a medium-sized main effect for lag \((F(1, 76) = 6.31, p < .05, \eta^2 = .08; F(1, 76) = 6.2, p < .05, \eta^2 = .08)\), which reflects the finding that the two-minute lag between interventions improves learning more efficiently than the seven-minute lag. Finally, there was a tendency towards a lag-by-time interaction \((F(2, 152) = 3.03, p = .51, \eta^2 = .03; F(1, 76) = 2.01, p = .16, \eta^2 = .03)\). No other effect reached a conventional level of significance in the GLM analysis.\(^3\)

4. Discussion

We demonstrated that quizzing can be effectively implemented under real classroom conditions and that the effect is strong enough to significantly enhance long-term learning. Moreover, we showed that the quizzing effect generalizes beyond the reproduction of declarative knowledge. Geometry learning requires students to not only memorize concepts (e.g., definition of symmetry) but also establish and apply schemata in a generic fashion. Specifically, our test asked students to categorize new objects in terms of symmetry as well as apply production rules to complete figures (cf. Fig. 1). Note also that we examined participants from the youngest age group studied so far – 9-year-olds in elementary school.

\(^3\) GLM results for the insignificant effects: quizzing*lag*time interaction \((F(2, 152) = .36, p =.70, \eta^2 < .01; F(1, 76) = .34, p =.56, \eta^2 < .01)\); quizzing \((F(1, 76) = 1.38, p =.24, \eta^2 < .01; F(1, 76) = .01, p =.94, \eta^2 < .01)\); quizzing*lag interaction \((F(1, 76) = 1.43, p =.24, \eta^2 < .01; F(1, 76) = .47, p =.49, \eta^2 < .01)\).
As a typical feature of any field research, one must strike a balance between maintaining ecological validity and experimental rigor. With the decision to study children under naturalistic class conditions, we could not control error variance through individual randomization. Thus, the class itself provided another potential unit of analysis. Ideally, one would draw a large sample of classes and perform multi-level analyses to control for sources of variance stemming from class features. Such a procedure would require a large number of schools and hundreds of student participants. Hence, the present research is only a first step.

Note, however, that classes were highly homogenous with respect to age, gender proportions, prior math grades, language, and socio-economic background. Moreover, children demonstrated a common level of prior knowledge, as indicated by test results before the intervention. Only one class tended to perform better than the others at t1; however, this difference worked against our prediction. The interaction effect could not be caused by this variation, because it was also present when test scores were individually corrected for prior knowledge level. Importantly, results should not be evaluated in isolation from prior evidence obtained under strictly controlled lab conditions. Acknowledging these findings, it is very unlikely that the replication of the quizzing effect would be a product of an arbitrary combination of variables other than the ones manipulated.

As an unexpected finding, we obtained a general increase in test scores over time, although teachers did not repeat the topic “symmetry axis” until the last test was taken. However, students did receive training in other geometry domains (e.g., area of forms), which presumably yielded a consolidation of geometry knowledge. This might have resulted in a facilitation effect on the test, especially on the items that address transfer of knowledge (cf. Figure 1). Note, however, that the interaction qualified this main effect. Thus, the effect cannot be explained by a general consolidation in knowledge.

For exploratory reasons, we varied the lag between script repetitions (2 vs. 7 minutes). Based on the literature, one should expect increasing lags to covary with learning success (Cepeda et al., 2006). However, this is not what we found. Our lag manipulation had a main effect on learning performance, but in the opposite direction.

---

4 Due to the school’s catchment area, all children were middle class native speakers without migrant background.
Individuals showed generally higher learning scores in the short compared to the long interval condition. The long interval might have yielded a decrease in motivation in children because it extended the entire session by 15 minutes compared to short lag condition. Hence, the obtained effect could be due to variables other than a genuine time distribution effect.

Pashler and colleagues (2007) recommended quizzing as an instructional technique with one reservation. The authors pointed out that the overwhelming majority of evidence for the quizzing effect stems from highly controlled lab studies. Consequently, the authors urged more effort to be invested in field research to assess whether the effect is sufficiently robust to survive in actual classroom settings. Thus far, only a few researchers followed that call. Though approaching classroom conditions, none of these embedded the quizzing manipulation in a real teacher-student interaction in its “natural” context, the school classroom. Our study closed this gap and demonstrated that quizzing is a robust effect that was also beneficial under real teaching conditions.

5. Authors’ note

We cordially thank Jessica Schöneberg and Madeleine Berend, who served as experimenters, and Heather Fuchs for very helpful comments.
6. References


Table 1: Test scores

<table>
<thead>
<tr>
<th>Test score</th>
<th>SSS lag</th>
<th>SQQ lag</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>12.43 (7.01)</td>
<td>9.03 (5.31)</td>
<td>9.65 (5.46)</td>
</tr>
<tr>
<td>n = 21</td>
<td>n = 20</td>
<td>n = 19</td>
<td>n = 20</td>
</tr>
<tr>
<td>t2</td>
<td>24.52 (6.86)</td>
<td>19.32 (9.62)</td>
<td>15.93 (7.93)</td>
</tr>
<tr>
<td>n = 21</td>
<td>n = 20</td>
<td>n = 19</td>
<td>n = 20</td>
</tr>
<tr>
<td>t3</td>
<td>24.32 (7.19)</td>
<td>23.39 (8.29)</td>
<td>20.88 (5.32)</td>
</tr>
<tr>
<td>n = 21</td>
<td>n = 20</td>
<td>n = 19</td>
<td>n = 20</td>
</tr>
</tbody>
</table>

Note: Table displays mean values of test scores at the three times of measurement (t1 = before, t2 = 5 min after, t3 = 6 weeks after intervention). Standard deviations are in parentheses.
Figure 1: Test items of the one-side completion task.
Quizzing

Figure 2: Mean learning improvement scores and standard errors (vertical lines) on the two times of post-testing collapsed over lag conditions. Quizzing = SQQ, studying = SSS.