Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

# Testing Effects in University Teaching: Short-answer Questions are Beneficial, Whereas Multiple-Choice Questions are not

Sven Greving, M.Sc.
Prof. Dr. Tobias Richter

# Abstract

Practicing retrieval instead of restudying is a powerful strategy to retain learned content (the *testing effect*). The robust testing effect found in laboratory experiments has spawned a growing body of research in educational contexts. However, existing research in educational contexts has often combined testing with additional didactical measures that hampers the interpretation of testing effects. In these contexts, it is furthermore unclear whether practicing multiple-choice questions is equally effective as practicing short-answer questions. We aimed to examine the testing effect in its pure form by implementing a minimal intervention design in a university lecture. In two field experiments we compared answering short-answer questions and multiple-choice questions to reading summarizing statements about core lecture content. Participants in both studies visited lectures and practiced lecture content according to their condition at the end of each lecture. Retention for lecture content was tested by means of a surprise criterial test at the end of the semester. In Experiment 1 we investigated whether the testing effect was affected by the retention interval. A positive testing effect emerged for short-answer questions that targeted information that participants could easily retrieve from memory. This effect was independent of the time of test. However, the results indicated no testing effect for multiple-choice questions. Experiment 2 was designed to replicate these finding in another lecture. Furthermore, means to increase the benefits of practicing multiple-choice questions were applied: Instead of selecting the single correct answer among four options, participants were required to rate every answer option independently as true or false. We replicated a positive testing effect for short-answer questions, however in Experiment 2 retrievability did not affect the testing effect. Again, the results indicated no testing effect for multiple-choice questions. Additionally, Bayesian analyses revealed evidence in favor of the null hypothesis indicating that practicing multiple-choice questions had no effect on retention. These results suggest that short-answer testing but not multiple-choice testing may benefit learning in higher education contexts.

# The Testing Effect

- Robust finding that testing of learned information increases retention more than re-studying (Karpicke & Roediger, 2007)
- Growing body of reseach that demonstrates testing effects in educational contexts (for meta-analyses, see Adesope, Trevisan, & Sundararajan, 2017; Schwieren, Barenberg, & Dutke, 2017)
- Most effective: Solvable but challenging retrieval practice (Pyc & Rawson, 2009)

# Methodological Issues

- However, in many studies in educational contexts methodical issues arise (see Greving & Richter, 2018 for details)
  - Lack of randomization
  - Lack of proper control condition
  - Feedback
  - Graded practice tests
  - Open label studies

# Experiment 1 - Rationale

(Greving & Richter, 2018)

- Are there substantial testing effects for short-answer(SA) and multiple-choice(MC) practice tests?
- Keeping methodological issues at minimum:
  - Real educational context (lecture)
  - Random assignement of conditions at end of each lecture session(SA-testing, MC-testing, or Restudy)
  - Surprise criterial tests at 1,12, and 23 weeks after last lecture

# Experiment 1 - Analyses

(Greving & Richter, 2018)

- Generalized linear mixed models (Subjects and items as random effects)
- Outcome: Pobability of answering correctly in criterial tests
- Important predictors:
  - Practice condition
  - Retrievability at end of lecture session

  (i.e., item difficulty; hard, medium, or easy)

# Experiment 1 - Results

## (Greving & Richter, 2018)

| | Short-Answer Testing | | Multiple-Choice Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***p*** | **β (*SE*)** | ***p*** |
| Intercept | -0.34 (0.25) | .173 | 0.07 (0.29) | .803 |
| Testing | 0.44 (0.24) | .033[a] | -0.42 (0.24) | .078 |
| Low retrievability | 0.03 (0.25) | .917 | -0.31 (0.25) | .219 |
| Medium retrievability | 0.09 (0.23) | .692 | -0.35 (0.27) | .184 |
| Testing x Low retrievability | -0.60 (0.28) | .016[a] | 0.17 (0.27) | .534 |
| Testing x Medium retrievability | -0.66 (0.35) | .030[a] | 0.06 (0.37) | .872 |
| $N_{Items}$ | 77 | | 77 | |
| $N_{Participants}$ | 92 | | 91 | |

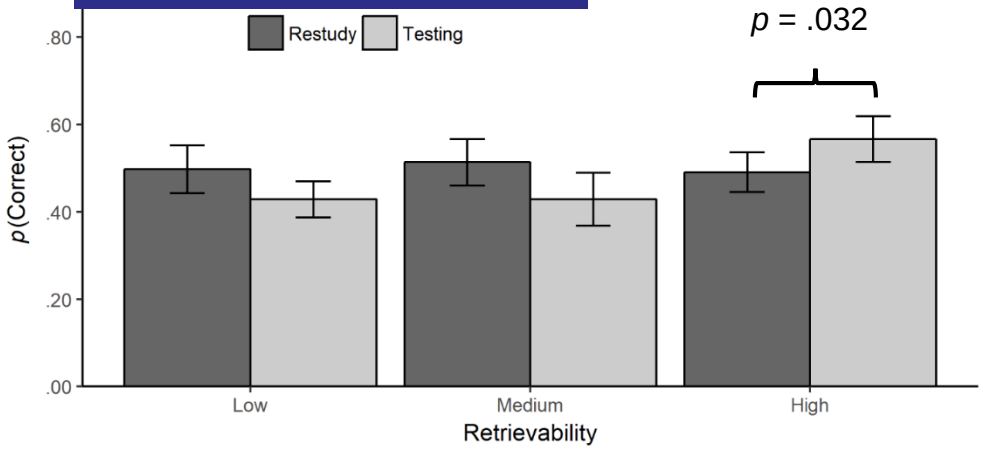
[a] p-values refer to one-tailed tests for β > 0. Other p-values refer to two-tailed tests.

# Experiment 1 - Results
## (Greving & Richter, 2018)

| | Short-Answer Testing | | Multiple-Choice Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***p*** | **β (*SE*)** | ***p*** |
| Intercept | -0.34 (0.25) | .173 | 0.07 (0.29) | .803 |
| Testing | 0.44 (0.24) | .033[a] | -0.42 (0.24) | .078 |
| Low retrievability | 0.03 (0.25) | .917 | -0.31 (0.25) | .219 |
| Medium retrievability | 0.09 (0.23) | .692 | -0.35 (0.27) | .184 |
| Testing x Low retrievability | -0.60 (0.28) | .016[a] | 0.17 (0.27) | .534 |
| Testing x Medium retrievability | -0.66 (0.35) | .030[a] | 0.06 (0.37) | .872 |
| $N_{Items}$ | 77 | | 77 | |
| $N_{Participants}$ | 92 | | 91 | |

[a] p-values refer to one-tailed tests for β > 0. Other p-values refer to two-tailed tests.

Interaction Pattern for SA-Testing



# Experiment 1 - Discussion
(Greving & Richter, 2018)

- Testing effect for highly retrievable short-answer questions
- No testing effect for multiple-choice questions
- Why only short-answer questions?
  - Multiple-Choice mere recognition task that does not require challenging retrieval? (Glover, 1989)
  - Multiple-Choice most effective when deeper processing of all options (Little, Bjork, Bjork, & Angello, 2012) → Multiple Response (Experiment 2)

# Experiment 2 - Results

| | Short-Answer Testing | | Multiple-Response Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***p*** | **β (*SE*)** | ***p*** |
| Intercept | 0.28 (0.36) | .444 | 0.28 (0.32) | .389 |
| Testing | 0.52 (0.41) | .099[a] | -0.57 (0.39) | .139 |
| Low retrievability | -1.23 (0.54) | .022 | -0.91 (0.50) | .068 |
| Medium retrievability | -0.18 (0.44) | .685 | -0.53 (0.48) | .283 |
| Testing x Low retrievability | 0.56 (0.58) | .166[a] | 1.07 (0.57) | .031[a] |
| Testing x Medium retrievability | -0.05 (0.60) | .467[a] | 0.80 (0.67) | .115[a] |
| $N_{Items}$ | 29 | | 29 | |
| $N_{Participants}$ | 43 | | 43 | |

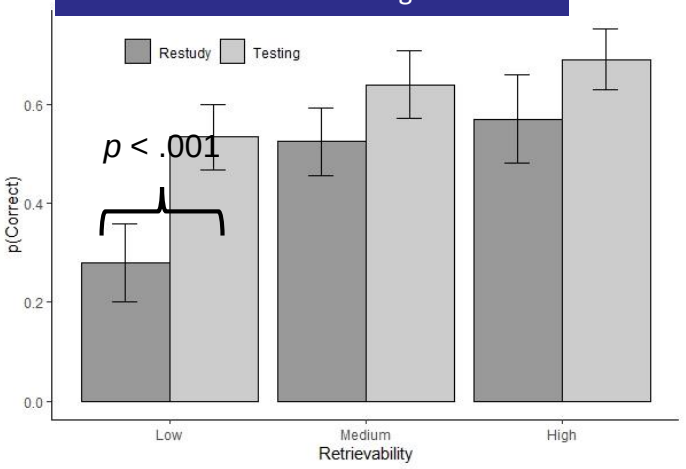
[a] p-values refer to one-tailed tests for β > 0. Other p-values refer to two-tailed tests.

# Experiment 2 - Results

| | Short-Answer Testing | | Multiple-Response Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***p*** | **β (*SE*)** | ***p*** |
| Intercept | 0.28 (0.36) | .444 | 0.28 (0.32) | .389 |
| Testing | 0.52 (0.41) | .099[a] | -0.57 (0.39) | .139 |
| Low retrievability | -1.23 (0.54) | .022 | -0.91 (0.50) | .068 |
| Medium retrievability | -0.18 (0.44) | .685 | -0.53 (0.48) | .283 |
| Testing x Low retrievability | 0.56 (0.58) | .166[a] | 1.07 (0.57) | .031[a] |
| Testing x Medium retrievability | -0.05 (0.60) | .467[a] | 0.80 (0.67) | .115[a] |
| $N_{Items}$ | 29 | | 29 | |
| $N_{Participants}$ | 43 | | 43 | |

[a] p-values refer to one-tailed tests for β > 0. Other p-values refer to two-tailed tests.

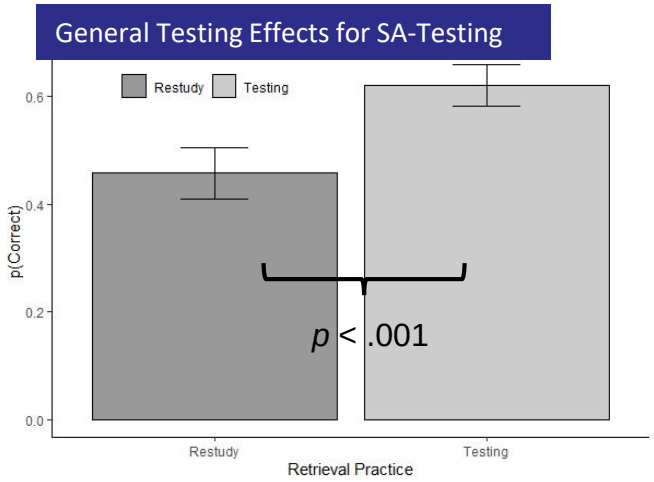
Effect Pattern for SA-Testing



# Experiment 2 - Results

| | Short-Answer Testing | | Multiple-Response Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***p*** | **β (*SE*)** | ***p*** |
| Intercept | 0.28 (0.36) | .444 | 0.28 (0.32) | .389 |
| Testing | 0.52 (0.41) | .099[a] | -0.57 (0.39) | .139 |
| Low retrievability | -1.23 (0.54) | .022 | -0.91 (0.50) | .068 |
| Medium retrievability | -0.18 (0.44) | .685 | -0.53 (0.48) | .283 |
| Testing x Low retrievability | 0.56 (0.58) | .166[a] | 1.07 (0.57) | .031[a] |
| Testing x Medium retrievability | -0.05 (0.60) | .467[a] | 0.80 (0.67) | .115[a] |
| $N_{Items}$ | 29 | | 29 | |
| $N_{Participants}$ | 43 | | 43 | |

[a] p-values refer to one-tailed tests for β > 0. Other p-values refer to two-tailed tests.

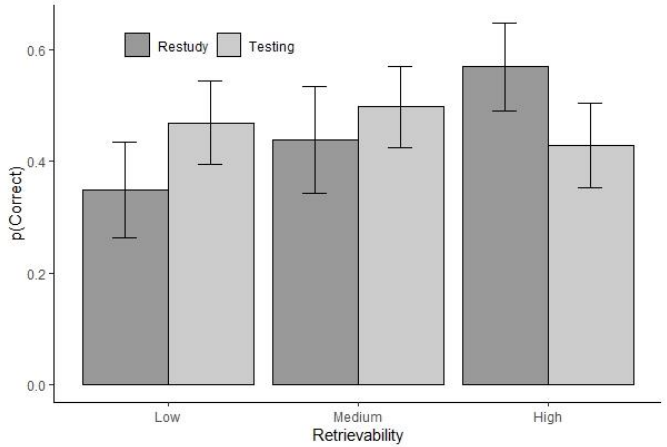


# Experiment 2 - Results

| | Short-Answer Testing | | Multiple-Response Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***p*** | **β (*SE*)** | ***p*** |
| Intercept | 0.28 (0.36) | .444 | 0.28 (0.32) | .389 |
| Testing | 0.52 (0.41) | .099[a] | -0.57 (0.39) | .139 |
| Low retrievability | -1.23 (0.54) | .022 | -0.91 (0.50) | .068 |
| Medium retrievability | -0.18 (0.44) | .685 | -0.53 (0.48) | .283 |
| Testing x Low retrievability | 0.56 (0.58) | .166[a] | 1.07 (0.57) | .031[a] |
| Testing x Medium retrievability | -0.05 (0.60) | .467[a] | 0.80 (0.67) | .115[a] |
| $N_{Items}$ | 29 | | 29 | |
| $N_{Participants}$ | 43 | | 43 | |

[a] p-values refer to one-tailed tests for β > 0. Other p-values refer to two-tailed tests.

Effect Pattern for MR-Testing



# Experiment 2 - Discussion

- Testing effect for short-answer questions irrespective of retrievability
  - In direct comparison, hard items perform best
- Again, no testing effect for multiple-response questions
- Is there evidence against multiple-choice/response testing effects?

# Additional Bayesian Analyses Across Experiments

| | Short-Answer Testing | | Multiple-Choice/Response Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***BF*$_{10}$** | **β (*SE*)** | ***BF*$_{10}$** |
| Intercept | 0.10 (0.19) | 0.21 | 0.49 (0.23) | .2.41 |
| Testing | 0.31 (0.19) | .3.46[b] | -0.37 (0.20) | < 0.01[b] |
| Medium retrievability | 0.04 (0.24) | 0.24 | -0.33 (027) | 0.54 |
| Low retrievability | 0.00 (0.25) | 0.26 | -0.31 (0.26) | 0.59 |
| Testing x Medium retrievability | -0.59 (0.36) | 2.20 | 0.06 (0.38) | 0.32 |
| Testing x Low retrievability | -0.57 (0.28) | 1.37 | 0.16 (0.27) | 0.39 |
| $N_{Observations}$ | 2397 | | 2461 | |

[b] $BF_{10}$-values refer to tests against the Null-Hypothesis of $\beta < 0.2$. Other $BF_{10}$-values refer to testing against $\beta = 0$.

# Additional Bayesian Analyses Across Experiments

| | Short-Answer Testing | | Multiple-Choice/Response Testing | |
|---|---|---|---|---|
| **Parameter** | **β (*SE*)** | ***BF*$_{10}$** | **β (*SE*)** | ***BF*$_{10}$** |
| Intercept | 0.10 (0.19) | 0.21 | 0.49 (0.23) | .2.41 |
| Testing | 0.31 (0.19) | .3.46[b] | -0.37 (0.20) | < 0.01[b] |
| Medium retrievability | 0.04 (0.24) | 0.24 | -0.33 (027) | 0.54 |
| Low retrievability | 0.00 (0.25) | 0.26 | -0.31 (0.26) | 0.59 |
| Testing x Medium retrievability | -0.59 (0.36) | 2.20 | 0.06 (0.38) | 0.32 |
| Testing x Low retrievability | -0.57 (0.28) | 1.37 | 0.16 (0.27) | 0.39 |
| $N_{Observations}$ | 2397 | | 2461 | |

[b] $BF_{10}$-values refer to tests against the Null-Hypothesis of $\beta < 0.2$. Other $BF_{10}$-values refer to testing against $\beta = 0$.

3.46 times more likely that data occured under an Hypothesis, stating **a positiv effect of testing**

333 times more likely that data occured under an Hypothesis, stating **no effect of testing**

# Summary

- Across experiments:
  - Evidence for testing effects of short-answer testing
    - However, retrievability should be factored in and investigated more
  - Evidence against testing effects for multiple-choice/response testing

# References


- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*, 659–701. https://doi.org/10.3102/0034654316689306
- Glover, J. A. (1989). The „testing“ phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.
- Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*. https://doi.org/10/gfkwvm
- Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. https://doi.org/10.1016/j.jml.2006.09.004
- Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344. https://doi.org/10.1177/0956797612443370
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447. https://doi.org/10.1016/j.jml.2009.01.004
- Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*, 179–196. https://doi.org/10.1177/1475725717695149