# A critique to Akdemir and Oguz (2008)

Albers, Casper J.; Boevé, Anja J.; Meijer, Rob R.

*Published in:*
Computers & Education

*DOI:*
10.1016/j.compedu.2015.07.001



*Publication date:*
2015

Computers & Education 87 (2015) 238–242

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

# A critique to Akdemir and Oguz (2008): Methodological and statistical issues to consider when conducting educational experiments

Casper J. Albers*, Anja J. Boevé, Rob R. Meijer

*Department of Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands*

ARTICLE INFO

*Article history:*
Received 29 January 2015
Received in revised form 29 June 2015
Accepted 1 July 2015
Available online 10 July 2015

*Keywords:*
Evaluation methodologies
Methodology in education
Statistics

ABSTRACT

In the paper "Computer-based testing: An alternative for the assessment of Turkish undergraduate students", Akdemir and Oguz (2008) discuss an experiment to compare student performance in paper-and-pencil tests with computer-based tests, and conclude that students taking computer-based tests do not underperform compared to students taking pen-and-pencil tests. In this letter, we indicate two severe methodological and statistical flaws in this paper. We show how, in general, such flaws can affect experimental research. Due to these flaws, the conclusions by Akdemir and Oguz are unfounded: one cannot reach these conclusions on basis of this design and analysis. We provide a set of guidelines and advices to avoid methodological problems when setting up an educational experiment.


## 1. Introduction

Recent cases of fraud in the social scientific community have sparked debates on healthy research practice in the social sciences (Sijtsma, 2015). It is our responsibility as researchers to learn from these mistakes and promote healthy research practice in the future (Martinson, Anderson, & De Vries, 2005). For this reason, the present paper comments on a study published in *Computers and Education* that is flawed both methodologically and statistically. In the present paper we will discuss these flaws with the aim to promote healthy research practice. In the paper "Computer-based testing: An alternative for the assessment of Turkish undergraduate students", Akdemir and Oguz (2008) discussed an experiment in which student performance was compared in paper-and-pencil (P&P) tests and computer-based (CB) tests. They concluded that students taking CB tests did not underperform compared to students taking P&P tests. We first shortly discuss the Akdemir and Oguz (2008) study, followed by a methodological and statistical critique. We then provide several recommendations concerning experimental design and analysis.

## 2. The Akdemir and Oguz-study

The purpose of the Akdemir and Oguz (2008) study was to investigate whether students performed equally well in CB tests and P&P tests. This is an important issue when implementing new technologies: students should not be disadvantaged when

* Corresponding author.
*E-mail addresses:* c.j.albers@rug.nl (C.J. Albers), a.j.boeve@rug.nl (A.J. Boevé), r.r.meijer@rug.nl (R.R. Meijer).

http://dx.doi.org/10.1016/j.compedu.2015.07.001

using a new mode of test administration (McDonald, 2002; see also Akdemir and Oguz, p. 1198–1199, for references to literature indicating possible disadvantages). The study of Akdemir and Oguz (2008) was conducted with a group of undergraduate students at a public university in Turkey. The authors reported that 47 students were randomly selected to participate in the study; there were 17 male students and 30 female students. All students completed a P&P test consisting of 30 multiple-choice questions on topics from a course studied by the students in the previous semester, thus the material that was tested was not part of the students current education program at the time of the study. Four weeks later, the same group of students again completed the test, but this time via the computer. (It was not clear from the Akdemir and Oguz (2008) paper whether it was exactly the same test that was administered at both moments, or two different tests on the same study material.) The average number of correct answers on the P&P test was 12.9 (SD = 2.1), and the average number of correct answers on the CB test was 13.6 (SD = 2.6). Using three separate one-way ANOVA's, the authors reported that there was no overall difference in test performance between modes, and there was no difference in test performance between modes for both sexes separately.

## 3. Methodological flaw

The causal effect of interest (difference in performance between modes of testing) was not isolated, but confounded with a potential practice effect since a crossed design, also known as crossover design, was *not* used in this study. All students in the Akdemir and Oguz (2008) study first participated in the paper-and-pencil test and some weeks later they participated in the computer-based test. Both tests were constructed on the basis of the same study material; this is visualized in Fig. 1 (left). The authors found that, on average, students scored better the second time they took the test (average scores 13.6 vs 12.9 at the first test). Due to the study design, it is impossible to distinguish whether this difference in performance is purely due to differences in testing mode (P&P vs CB) or due to a practice effect. The practice effect (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Kulik, Kulik, & Bangert, 1984) refers to the tendency to score higher on a repeated measurement of the same test. The effect of the test mode cannot be isolated from a practice effect and this may have different and unknown consequences. The reverse could also be true: the score on the first test may have been inflated because this test occurred sooner after the test material was taught and students therefore recollected more of the study material compared to the second test (remembering effect). If either or both of these two examples occurred, then the difference between the observed CB and P&P results would be underestimated. Thus, the practice effect leads to an increased score on the second test, whereas the remembering effect leads to an increased score on the first test. It is unknown which of these effects outweighs the other. Furthermore, the size of these effects may differ per student so that for some students the effect is an increase in the difference between CB and P&P, whereas for other students this difference would decrease.

It is impossible to determine to what extent practice and/or remembering effects occurred in the Akdemir and Oguz (2008) study on hindsight. These effects *may* have been virtually absent or they may have canceled each other out, but we simply do not know: we cannot rule out that the outcomes of Akdemir and Oguz (2008) were distorted. One of the main goals in designing experiments is to control such so-called confounding variables (cf. Fisher, 1935; VanderWeele & Shpitser, 2013), and this was not the case in the paper by Akdemir and Oguz. It is thus impossible to judge whether the results are reliable or not, which makes them, by definition, unreliable. For this reason, we will discuss in Section 5 how such confounding variables could have been controlled for.

A better strategy would have been to collect the data through some kind of randomized crossed design, as visualized in Fig. 1 (right). In this design roughly half of the students are randomly assigned to group A, and the other half of the students is randomly assigned to group B. Assume that the students in group A are first administered the P&P test followed by the CB test, and assume that the students in group B are administered these tests in reversed order. If a within-subjects study design is used with so-called parallel tests, the practice effect could also be investigated further by extending the design to include students randomly assigned to two paper-and-pencil tests as well as students randomly assigned to two computer-based tests. Note that this latter option is only available when both tests have been shown to be parallel, which is not always possible and requires sophisticated psychometric analysis of the test questions (Boekkooi-Timminga, 1990; Jöreskog, 1971). Most importantly, however, an appropriate design needs to be selected prior to the data collection.

It is important to consider the choice for a between-subjects or within-subjects design, remembering that the most important condition for drawing causal inference is random assignment to treatment conditions (Gerber & Green, 2012).

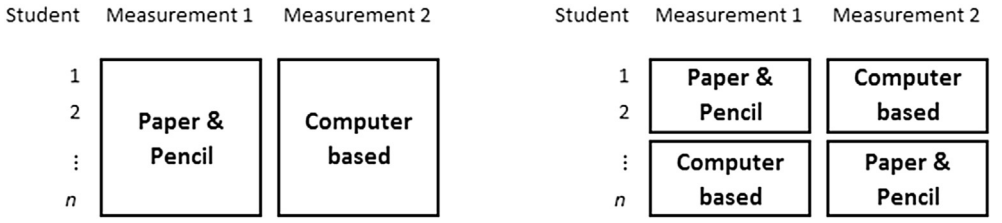


**Fig. 1.** The design used by Akdemir and Oguz (left) and a fully crossed design (right).

Random assignment partials out the effect of unobserved confounders. A between-subjects design, where students are randomly assigned to a CB or P&P based condition is sufficient to draw reliable causal inference on the difference between test modes. The advantage of a within-subjects design, however, is that individual differences are partialled out efficiently, which leads to higher power and, in turn, to a smaller required sample size.

## 4. Statistical flaw

In Akdemir and Oguz (2008) the significance of the difference between both groups was studied using consecutive ANOVAs. First, a one-way ANOVA was performed to study differences between P&P and CB tests; next two one-way ANOVAs were performed to study the difference between P&P and CB tests for males and females. In general, conducting a single two-way ANOVA is better than conducting in total three, separate ANOVAs on the same data because of capitalization on chance. However, it was fundamentally wrong to use an ANOVA in this case. The ANOVA, which is statistically equivalent to a *t*-test when two groups are involved, is a method for comparing *independent* samples, and the study of Akdemir and Oguz (2008) clearly had a *paired* or *repeated measures* sample: that is, the same students were measured multiple times. An ANOVA disregards any within-subject variations and, in case these variations occur, places them under the between-subject label.

Applying an independent samples procedure to a dependent samples context has serious consequences. First, incorrectly claiming that $n = 94$, a much larger sample size than the actual $n = 47$, has two consequences: (i) an increase in power and (ii) a decrease in *p*-values. Second, by ignoring within-subjects effects, the opposite occurs: (i) the power is deflated and (ii) *p*-values are inflated. It is unknown whether the effect of incorrectly doubling the sample size is larger than that of ignoring within-subject effects. Therefore the combined effect of this statistical flaw on the *p*-value can be either decreasing or increasing, as is shown in Fig. 2. This figure consists of two constructed data sets, with $n = 47$ and means and SDs comparable to those in Akdemir and Oguz (2008). For the figure on the left, the *p*-value for the independent samples approach is much higher than that of the paired approach: $p_{\text{indep}} = 0.174$ vs $p_{\text{paired}} < 0.001$ (assuming, like Akdemir and Oguz did, constant variance). Furthermore, the independent *t*-tests approach has much lower effect size ($d = 0.282$ vs $d_z = 0.803$) and power (0.39 vs 1.00) than the paired approach. Thus, the paired approach provides much more evidence for an effect than the independent samples approach. For the figure on the right, the opposite can be observed: the independent samples *p*-value is now smaller than the dependent samples *p*-value ($p_{\text{indep}} = 0.199$ vs $p_{\text{paired}} = 0.295$) and also for effect sizes ($d = 0.267$ vs $d_z = 0.154$) and power (0.36 vs 0.27) the independent samples approach shows (slightly) more evidence of an effect. Thus, it is impossible to tell whether the *p*-values reported in Akdemir and Oguz (2008) are too large or too small.

## 5. Conclusion and recommendations for setting up educational experiments

The Akdemir and Oguz (2008) study, claiming to have found evidence in favor of CB testing, has been cited now over twenty times in peer-reviewed articles. Furthermore, researchers have replicated this study (Jeong, 2014), in a different population, with the same flaws as the study by Akdemir and Oguz (2008). Because of the severity of these flaws, the conclusions in Akdemir and Oguz (2008) are unfounded: the paper did not find valid evidence in favor of CB testing, nor of the reverse. In the remainder of this section, we provide some practical recommendations that may help researchers that want to conduct a (pseudo-)controlled experiment. For a more in-depth analysis the reader may want to consult the statistical textbooks by Bailey (2008) and Casella (2008), or papers aimed at an educational audience by McGowan (2011), Cobb, Confrey, Lehrer, and Schauble (2003), and Krathwohl (1964).

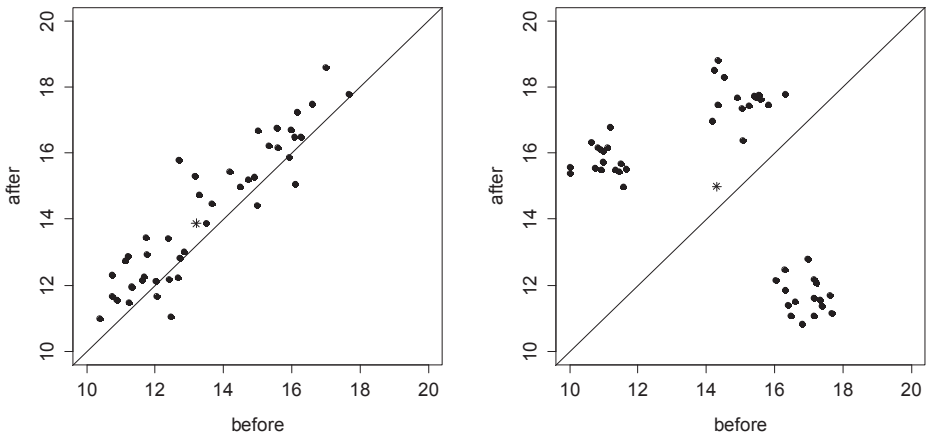


**Fig. 2.** Two constructed data sets with $n = 47$.

1. Control for confounding variables. When assignment of participants to groups can be experimentally manipulated, such as assigning students to the two test modes, the best way to control for confounding variables is by using randomized crossed designs. When assignment to groups of a variable cannot be manipulated, such as gender of the participant, the potential effect of such a variable can be minimized using block random assignment. This would entail randomly assigning all the males to both test modes, and separately assigning the females to both different test modes, so that eventually the CB and P&P groups roughly have the same proportion of men and women. Furthermore, measure the confounding variables such that their effect can be further accounted for in the statistical analysis. For a detailed explanation on the design and analysis of randomized crossed designs, see, for example, Gerber and Green (2012), Jones and Kenward (2003) and Senn (2002).
2. Before conducting an experiment, conduct a power analysis in order to check whether the research questions can be answered within the limits on sample size that time and money impose. Power is defined as the probability that, when there is an effect (thus, when $H_0$ is false), the statistical test indeed rejects $H_0$. It is better to find out that you do not have the required resources to adequately address your research question before than after you collected data. There are many easy-to-use online tools for computing statistical power, such as G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). In their study, Akdemir and Oguz had 47 participants, a difference between group means of 0.7, and group standard deviations of 2.1 and 2.6. Assuming a correlation of $r = 0.5$ between both test scores, the a posteriori power of the paired $t$-test or, equivalently, the repeated measures ANOVA, would be 50%. Thus, when flipping a fair coin in order to decide whether or not to reject $H_0$, the same success rate would have been achieved. When an a priori power analysis had been performed, these authors might have concluded that $n = 47$ is insufficient for their study. A power computation using G*Power shows that, for the paired samples $t$-test with $\alpha$ set to 0.05 and $n = 47$, one might expect to find effect sizes from $d_z = 0.483$ with 90% power and effect sizes from $d_z = 0.417$ with 80% power. Thus, their sample size was sufficient for finding medium and large effects but insufficient for finding small effects. If the aim of their study was to also find small effects ($d_z = 0.3$), then the sample size should have been no less than $n = 90$ (for 80% power) or $n = 119$ (for 90% power). Had the authors chosen for an independent sample design, then a sample size $n = 139$ is required for finding small effect sizes ($d = 0.3$) with at least 80% power.
3. Analyse all data in one model. Akdemir and Oguz (2008) first performed a one-way ANOVA to find overall differences between the CB and P&P-groups (Table 2 in Akdemir and Oguz (2008)), followed by one-way ANOVAs for male and female students (Tables 3 and 4 in Akdemir and Oguz (2008)) separately. Carrying out separate analyses leads to chance capitalization: The increased risk of a Type I error. By performing all analyses within the same model (e.g., a two-way ANOVA or regression model), this phenomenon is controlled for.
4. When you *expect* things to get complicated, call in a statistician and do this *before* you collect data. Especially when there is an expected hierarchical or longitudinal structure as in Akdemir and Oguz (2008) in the data, design and analysis can become complicated and a statistician can be a valuable member in the research team. In the words of statistician R.A. Fisher (1938), "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of".
5. Always remain critical with respect to the generalizability of your study. There are two important things to consider here – the generalization of the context of the study and the generalization to a broader population of interest. First, be careful when generalizing to the population level (cf. Manski, 2009). Since Akdemir and Oguz (2008) randomly selected students to participate in the study, the results of their study may generalize to the student population from which was randomly sampled.

Second, concerning the context of the study, it is important to remember that the environment in which (pseudo-) controlled experiments take place is controlled, whereas the context in which new technology is implemented may not be controlled to a similar extent. For instance, in the Akdemir and Oguz paper, students participated in a test that was not part of their degree program, and thus did not encounter the stress associated with high-stakes tests. An experiment that is conducted in a natural setting, but where the independent variable, in this case type of test, still can be controlled, is known as a field experiment (cf. Gerber & Green, 2012). The advantage of conducting field experiments is that they tend to have higher external validity, enabling generalization to the context in which a phenomenon actually occurs, which is generally not the case for controlled laboratory experiments. Designing good field experiments however, may be challenging for ethical reasons; the experimental manipulation may not have an adverse impact on those who consent to participate.

In this paper we outlined two important flaws encountered in the paper by Akdemir and Oguz (2008) and demonstrated how these flaws affected the validity of their results. We hope that our detailed explanations, as well as the list of practical recommendations, help educational researchers prevent similar flaws in future research.

## References

Akdemir, O., & Oguz, A. (2008). Computer-based testing: an alternative for the assessment of Turkish undergraduate students. *Computers & Education, 51*, 1198–1204. http://dx.doi.org/10.1016/j.compedu.2007.11.007.

Bailey, R. A. (2008). *Design of comparative experiments* (Vol. 25). Cambridge, UK: Cambridge University Press.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational and Behavioral Statistics, 15*(2), 129–145. http://dx.doi.org/10.3102/10769986015002129.

Casella, G. (2008). *Statistical design*. New York, USA: Springer Science & Business Media.

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational researcher, 32*(1), 9–13. http://dx.doi.org/10.3102/0013189X032001009.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. http://dx.doi.org/10.3758/BF03193146.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1938). Presidential address. *Sankhyā: The Indian Journal of Statistics*, 14–17. Retrieved from http://www.jstor.org/stable/40383882.

Gerber, A. S., & Green, D. P. (2012). *Field Experiments: Design, analysis and interpretation*. New York: W.W. Norton.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373. http://dx.doi.org/10.1037/0021-9010.92.2.373.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410–422. http://dx.doi.org/10.1080/0144929X.2012.710647.

Jones, B., & Kenward, M. G. (2003). Design and analysis of cross-over trials. In *Monographs on statistics and applied probability* (2nd ed., vol 98). London: Chapman & Hall.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109–133. http://dx.doi.org/10.1007/BF02291393.

Krathwohl, D. R. (1964). Experimental design in educational research. *Library Trends, 13*(1), 1964. Research Methods in Librarianship, 54-67. Retrieved from http://hdl.handle.net/2142/6159.

Kulik, J. A., Kulik, C. L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435–447. http://dx.doi.org/10.3102/00028312021002435.

Manski, C. F. (2009). *Identification for prediction and decision*. Cambridge, USA: Harvard University Press.

Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737–738. http://dx.doi.org/10.1038/435737a.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education, 39*(3), 299–312. http://dx.doi.org/10.1016/S0360-1315(02)00032-5.

McGowan, H. M. (2011). Planning a comparative experiment in educational settings. *Journal of Statistics Education, 19*(2), 1–18. Retrieved from www.amstat.org/publications/jse/v19n2/mcgowan.pdf.

Senn, S. (2002). *Cross-over trials in clinical research* (2nd ed.). Chichester: Wiley & Sons.

Sijtsma, K. (2015). Playing with Data—Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, 1–15. http://dx.doi.org/10.1007/s11336-015-9446-0.

VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *Annals of Statistics, 41*(1), 196. http://dx.doi.org/10.1214/12-AOS1058.