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).

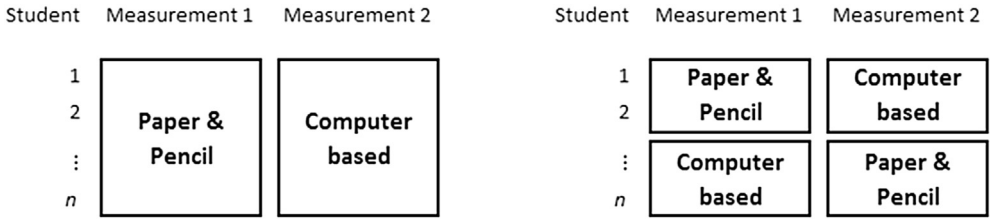

Fig. 1. The design used by Akdemir and Oguz (left) and a fully crossed design (right).