E3S Web of Conferences **274**, 12007 (2021) https://doi.org/10.1051/e3sconf/202127412007
STCCE – 2021

On the statistical indicators of the effectiveness of teaching methodologies

Ruslan Abitov[1 *] [0000-0003-4219-9815], *Maria* Nizamieva[1] [0000-0002-6984-7846], and *Elena* Konovalova[1] [0000-0003-2598-167X]

[1]Kazan State University of Architecture and Engineering, 420043, Zelenaya st., Kazan, Russia

Abstract. The article is devoted to the problem of carrying out a proper educational experiment. A critical analysis of the conventional approaches to an educational experiment is given. The author argues that the majority of methods, assessing the results of educational experiments, were borrowed from sociology and psychology which, in turn, led to the misinterpretation of the results of these experiments. The criticism of the author is primarily aimed at the incorrect use of central tendency measures and the selection of tests for checking the probability of significance of samples. The qualitative approach, based on the percentile values, was proposed as one of the most relevant results of an experiment. The lack of universal measure with could allow comparing results of multiple educational experiments and meta-analyses was argued. The term «effective learning hours» was coined. The methodology of defining «effective learning hours» and corresponding them to the levels of the acquisition was proposed.

Keywords: teaching methodology, effective leaning hours, educational experiment, objective assessment, measures of central tendency.

1 Introduction

Pedagogy, or, in other words, «Education science», although being a part of behavioral sciences, has always been based upon mathematical apparatus to prove educational hypotheses. It is obvious that for them to be mathematically proved we have to properly design the experiment that is to define:

- Size of control and experimental samples, accordingly using certain statistic methods that must take into account the level of the shift we want to detect during an experiment;
- Final design of an experiment in compliance with the «Principle of the sole difference» or, in other words, «Equality of all experimental conditions of the experiment except for the observed one»;
- Measures of the experimental data;
- Statistical measures which will show the efficiency or inefficiency of a proposed teaching methodology;
- Proper statistical tests that will mathematically prove the difference between samples.

* Corresponding author: rouslan.abitov@gmail.com

Proper education assessment measures first appeared with the creation of Classical Test Theory (CTT) [1-3]. However, this theory faced problems of portability and non-linearity of results. CTT also lacks adaptability of items, in other words, the methods of this theory do not allow filtering out inappropriate items as well as defining the difficulty of each of them [4]. To solve these issues Item Response Theory (ITR) was created by G. Rasch and other psychometrics statisticians [5-11].

Even though education researchers mostly use statistical methods, invented within the context of sociology and psychology, there still has been some misuse of measures (especially ones of central tendency such as arithmetic means, medians, etc.) along with misuse and misinterpretation of significance tests [12-14].

The aim of the paper is primarily reasoned by the fact that educational experiments pay increasingly more attention to the results that serve as a mathematical proof to justify a proposed pedagogical hypothesis. The main weaknesses of educational experiments are:

- Experimental methods and «recipes» were mostly inherited from psychology and sociology where using an arithmetic mean and other measures of central tendencies seem reasonable;
- Measures of the central tendencies as the main result run counter to the idea of education because in this case formally wins the only pedagogical technology that is aimed at an average student;
- Experimenters rarely understand that arithmetic reasonably describes only normal or at least symmetric distribution;
- The evaluation of the quality, repeatability, and portability of the study is often judged only by a sample size;
- A fairly common way of the representation of results is to group the results into high, medium, and low-level, with a corresponding chi-square test;
- Rare use of non-parametric tests;
- The absence of a universal unit of competence development results.

Thus we shall try to make a set of measures to ensure correct and objective students' assessment and evaluate the quality of training. We would also like to make a point that our propositions are rather aimed at methods of assessing an educational experiment – not at the designing of tests, although they use identical algorithms to ensure the objectiveness of assessment measures.

2 Methods

To develop an objective scale and to assess students' performance a concept of «guided learning hours» was used [15, 16]. This measure deals with the number of hours, required for acquiring a certain level of foreign language acquisition according to the Common European Framework of Reference for languages (CEFR) [17]. The theory of psychometrics and statistics in behavioral sciences served as a common basis to define certain measures that indicates the quality of teaching methodologies.

3 Results and discussion

It is no secret that the logic of the pedagogical research has been largely inherited from psychology and sociology, where most of the indicators fit into the normal distribution, and the methods of comparing arithmetic mean usually make sense. Even in cases where a sample is far from the normal distribution, but has a symmetrical form, the measures of arithmetic mean, median, mode (or multiple modes) are also fully justified to show its relevant characteristics. (Fig. 1).

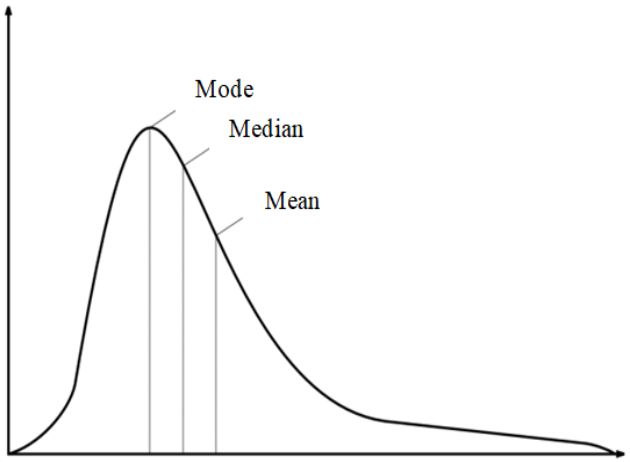

Fig. 1. Allocation of measures of central tendencies of asymmetric distribution.

However, to fully characterize the sample, it would be preferable to opt for not only the measure of one of the central tendency measures, but all of them, including the arithmetic mean, the median, the mode, along with standard deviations, and the confidence interval of both the arithmetic mean and the median.

However, if the maximum quality of education is to be obtained, then the most appropriate indicator is one of the lower percentiles of the sample. For example, the fifth percentile shows the guaranteed minimum level, acquired by 95 % of the sample, whereas the first one means that 99 % of students have higher performance than this value.

Let us look at the model situations of assessing students' performance (Fig. 2 and Fig. 3). In Fig. 2 (y-axis represents the number of students, whereas x-axis represents the level of performance) we can see a typical distribution before the beginning of a course where students have little-to-no knowledge.

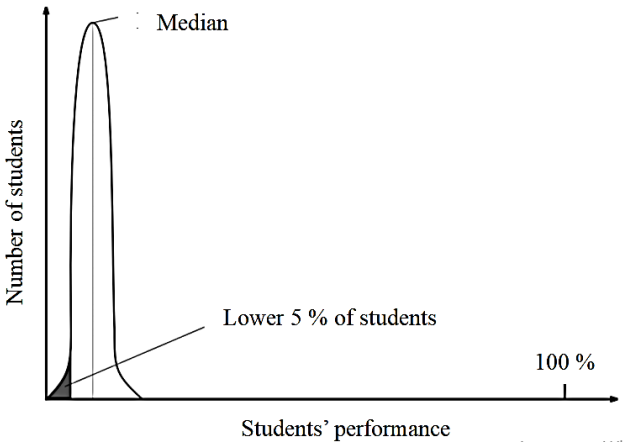

Fig. 2. Typical distribution of students' performance at the beginning of an educational experiment.

Fig. 3 shows the distribution of students' performance at the end of the experiment. Here we can say that the median of the sample came close to the value of 100%. Although showing the change of the median, as well as the other measures of central tendency, may seem to be very tempting to the majority of researchers, it would be the most honest to point out the shift of 5th percentile (the line that splits lower 5 % portion of students and the rest of the sample in Fig. 3) of the sample since this measure ensures that not less than 95 % of

students will not be below this value on completion of a course. The point we are driving at is if we rely upon the averages (mean, median, etc.) as the main indicators of research results, we will favor exclusively the middle range of an experiment sample. In other words, the lower range of the experiment sample can be completely disregarded by a teaching methodology, and yet we can present eye-pleasing results since the averages still has shown a significant shift. Thus experimental measures must show whether a new teaching methodology ensures the appropriate minimum number of students, and how it affected the lower portion of the sample.

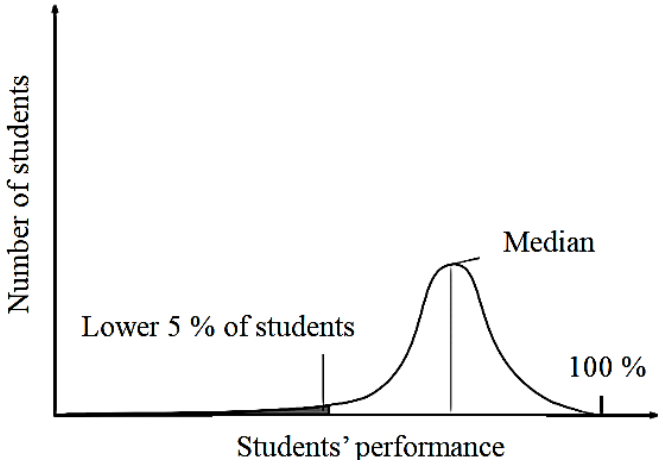

Fig. 3. Typical desired distribution of students' performance at the end of an educational experiment.

The next problem with the representation of pedagogical research results only through central tendencies are in contradiction with the basic idea of education – namely, the presenting of a higher average, which often does not reflect the processes of change within the sample.

For example, one of the most common ways of presenting results is through three levels of competence: high, medium, and low, further defined by verbal assessment descriptors. Fig. 4 shows us a hypothetical situations of two identical results. Often, at the final stage of an experiment, the average level change slightly or may not change at all. Which, in turn, raises the question, «Has the average level of competence even changed?» Thus, the results being completely identical, the same number of students can move either from low to high level (Fig. 4, the rightmost bar), or there also might be the situation where the same number of the students transferred both from low to medium level, and from medium to high (Fig. 4, central bar). However, these results are not the same. Thus, the most appropriate indicators of change in a sample are: midpoint, first and third quartile, as indicators of movement within the sample; and fifth or first percentile, or lower quartile as the main indicator of the quality of the training. Such indicators, along with the median, can provide a correct view of the sample and its internal changes. To show the shift within the sample Mann-Withney and Wilcoxon tests are justified [18-20]. The main advantage of nonparametric tests, especially of the T-criterion of Wilcoxon, is that they take into account the rank shift of all the items of a sample.

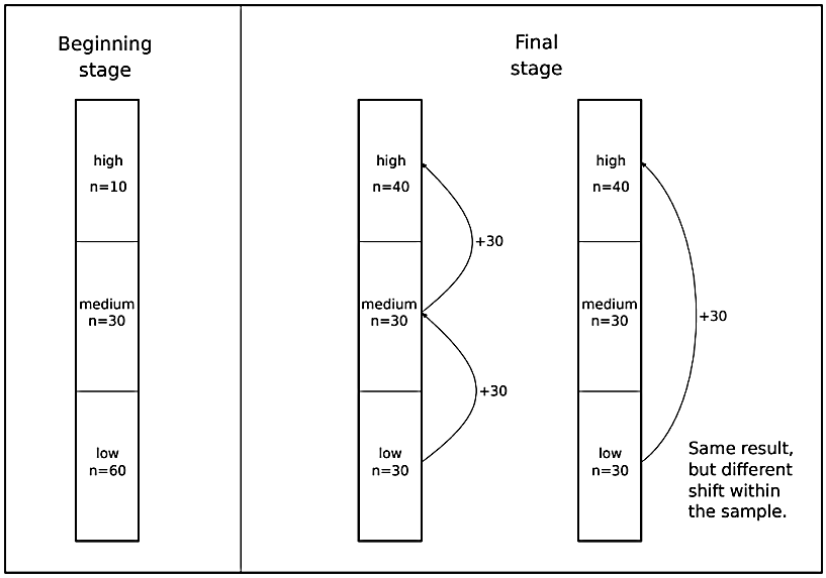

Fig. 4. The difference of shift within the sample.

Experimenters are quite often using parametric tests for checking homogeneity of samples, as well as the test for the homogeneity of variance (in case of t-test check). However, the negative results of these tests do not indicate that parametric tests are strictly prohibited. This only leads to the loss of the sensitivity of tests and, therefore, to errors of the second kind in many pedagogical experiments, especially on relatively small samples. It should also be remembered that the size is not so much a measure of portability and validity of the experiment as it is a parameter that determines how small changes in the experiment we want to indicate. The smaller changes we want to catch, the bigger a sample should be.

Since there is a lack of objective measures for an objective methods of testing teaching methodologies a new method is to be designed with a quantitative measure that would allow for a comparison between different competencies.

To address this challenge, two approaches seem to be the most adequate:

– The approach that takes into account the number of learning hours for a specific level of competence [15, 16];
– The labor productivity-based approach.

To determine the first indicator, it would be necessary to use the method of expert assessment to determine the effective hours of instruction in the following algorithm:

– Define the levels of proficiency by writing verbal descriptors for the most accurate description of the level;
– Carry out a primary pedagogical experiment that defines the 5th-percentile number of hours to reach all levels;
– Use the effective learning hours indicator for further researches and meta-researches.

The indicator of «effective learning hours» is a good objective indicator of the quality of education since:

– It shows a specific indicator of the training intensity;
– Effective learning hours can be a handy tool for comparing research and conducting meta-research.
– It indicates the effectiveness of learning by comparing the hours, spent on the guided training, and effective learning hours;

– Allows for determining the optimal number of hours of guided training for a competency.

4 Conclusion

In this paper, we tried to refine methods so that educational researches should be more accurate and representative. We don't insist on using lower percentiles as the main measures of quality of educational researches.

This paper, of course, does not cover all the aspects of the quality assurance of educational experiments, and there are a few aspects that we didn't touch in this work:

– The way how to introduce ITR to the concept of effective guided learning hours is still to be done;

– Effective guided learning hours, being a good objective measure for evaluating, may not be a latent characteristic.

Hopefully, we shall solve these problems and present a synthesis of effective guided learning hour concept and ITR in our further works.

References

1. M. Allen, W. Yen, Introduction to measurement theory (2002)
2. F. Lord, M. Novick, M Addison-Wesley, Statistical theories of mental test scores (1968)
3. M. R. Novick, *The axioms and principal results of classical test theory,* Journal of mathematical psychology, **3**, 1-18, DOI: 10.1016/0022-2496(66)90002-2 (1966)
4. R. K. Hambleton, R. J. Shavelson, N. M. Webb, H. Swaminathan, H. J. Rogers, Fundamentals of item response theory (1991)
5. F. B. Baker, The basics of item response theory (2001)
6. A. A. Bichi, R. Talib, *Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development,* International Journal of Evaluation and Research in Education, **2(7)**, 142-151, DOI: 10.11591/ijere.v7i2.12900 (2018)
7. W. Bonifay, L. Cai, *On the complexity of item response theory models*, Multivariate Behavioral Research, **4(52)**, 465-484, DOI: 10.1080/00273171.2017.1309262 (2017)
8. S. E. Embretson, S. P. Reise, Item response theory (2013)
9. I. Himelfarb, *A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating*, Journal of Chiropractic Education, **2(33)**, 151-163, DOI: 10.7899/jce-18-22 (2019)
10. F. Lord, Applications of item response theory to practical testing problems (1980)
11. G. Rasch, *An item analysis which takes individual differences into account*, British journal of mathematical and statistical psychology, **1(19)**, 49–57, DOI: 10.1111/bmsp.1966.19.issue-1 (1966)
12. W. Divale, M. Harris, D. T. Williams, *On the misuse of statistics: a reply to Hirschfeld et. al.*, American Anthropologist, **2(80)**, 379–386, DOI: 10.1525/aa.1978.80.2.02a00160 (1978)
13. T. J. Lamiell, Psychology's misuse of statistics and persistent dismissal of its critics (2019)
14. E. Park et al, *Correct use of repeated measures analysis of variance*, Korean J Lab Med, **1(29)**, 1-9, DOI: 10.3343/kjlm.2009.29.1.1 (2009)
15. https://support.cambridgeenglish.org/hc/en-gb/articles/202838506-Guided-learning-hours
16. L. Ahmad Afip, M. O. Hamid, P. Renshaw, *Common European framework of reference for languages (CEFR): insights into global policy borrowing in Malaysian higher*

education, Globalisation, Societies and Education, **3 (17)**, 378-393, DOI: 10.1080/14767724.2019.1578195 (2019)
17. Common European Framework of Reference for Languages: learning, teaching, assessment Cambridge: Cambridge University Press (2001)
18. S. Siegel, Nonparametric statistics for the behavioral sciences, (1956)
19. S. S. Sawilowsky, *Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney test for shift in location parameter,* Journal of Modern Applied Statistical Methods, **4**, 598-600, DOI: 10.22237/jmasm/1130804700 (2005)
20. M. W. Fagerland , L. Sandvik, *The Wilcoxon–Mann–Whitney test under scrutiny*, Statistics in medicine, **10 (28)**, 1487–1497, DOI: 10.1002/sim.3561 (2009)