



Klar N, Donner A (2002). The impact of EF Lindquist's text "Statistical Analysis in Educational Research" on cluster randomization.

© Allan Donner PhD, Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario, Canada, N6A 5C1. E-mail: donner@biostats.uwo.ca

In the 17th century [Van Helmont](#) made one of the earliest known proposals to 'cast lots' (randomize) to make a fair test of the effects of bloodletting and purging in people with fevers and pleurisies. There is debate about whether the original Latin should be interpreted to mean that lots would be cast to decide which individual patients or which of two groups of patients would be allocated to bloodletting and purging and which to a group treated without resorting to these methods. If Van Helmont intended that lots should be cast to decide which of two groups of patients would be treated with bloodletting and purging, then the controlled trial he proposed can be seen as an early example of what we now refer to as a 'cluster' randomized trial. The clinical trial described by [Amberson et al](#) in 1931 provides a clear description of cluster randomization: after dividing 24 patients into two pair-matched groups of 12, a single coin flip was used to assign all patients in a group to receive either the experimental treatment or a control treatment.

Cite as:

Klar N, Donner A (2002). The impact of EF Lindquist's text "Statistical Analysis in Educational Research" on cluster randomization. JLL Bulletin: Commentaries on the history of treatment evaluation (www.jameslindlibrary.org).

Cluster randomized trials - in which the units of randomization might be communities, worksites, schools, or families, for example - started becoming increasingly common at the end of the 20th century ([Donner and Klar 2000](#)). Almost invariably, however, the actual use of this design has come well before the recognition by investigators of the associated analytic implications. A key feature of cluster randomized trials is that responses of people in the same cluster tend to be correlated, or equivalently, the variation among observations in different clusters exceeds the variation within clusters. Failure to adjust standard statistical methods for within-cluster dependencies will result in underpowered studies with spuriously elevated type I errors.

Everet F. Lindquist (1901-1978), an educational researcher, was one of the first people to anticipate the profound statistical implications of cluster randomization. Lindquist joined the College of Education at the University of Iowa as a research assistant in 1925, and subsequently developed the Iowa Tests of Basic Skills and the American College Testing Program Tests, which are used by school students throughout the United States. The textbook that Lindquist published in 1940 - *Statistical Analysis in Educational Research* - contains the earliest clear exposition of the analytic implications of cluster randomized trials of which we are aware ([Lindquist 1940](#)). [Readers interested in the more recent history of cluster randomization may wish to consult Donner and Klar ([2000](#)), from which aspects of this commentary have been abstracted.]

Lindquist was influenced by Fisher's ideas on sound practice in experimental design (Fisher 1935), stating in the preface to his book that "The writer's primary purpose...has been to translate Fisher's expositions into a language and notation familiar to the student of education..." A recurring point made in Lindquist's text is that the availability of test scores from large numbers of students has falsely led many researchers in education to ignore 'recent' developments in small sample theory. He further points out (page v) that "In taking this attitude, we have overlooked the very significant fact that most of our samples, however large in terms of numbers of individual observations, are not simple random samples, but consist of relatively homogeneous and intact subgroups, such as the pupils in a single school or under a single teacher."

He expands this discussion in Chapter 1 (p 21-24), providing individual-level and school-level analyses of data collected on 3646 students enrolled in 24 schools. These analyses showed that methods that ignore the natural variation in response among schools tend to overestimate the precision of the estimated average test score by a factor of four! To ensure the validity of statistical inferences Lindquist favored analyses conducted at the school level, noting (page 24) that "...the size of the sample is dependent, not upon the number of individual observations, but upon the number of intact groups or subsamples of which the total sample is constituted."

Chapter 2 describes methods for the analysis of categorical outcome data while Chapter 3 attempts to introduce 'current' notions related to small sample theory to the reader. These chapters serve mostly to develop the tools needed

in later chapters when considering appropriate analyses for school-based randomized trials. For instance, Lindquist illustrates Pearson's chi-square test of homogeneity using data for the number of correct and incorrect answers to a test question posed to students from ten different schools. A conclusion drawn from the statistically significant heterogeneity in responses among these schools is that the data cannot be analyzed as if obtained from a simple random sample of students. Thus Lindquist again emphasizes (page 46) that "... the important consideration is not the number of pupils involved...but the number of schools represented".

The merits of several school-based experimental designs in measuring the precision of the observed effect of intervention are reviewed in Chapter 4. These designs are presented in the context of a hypothetical study comparing average scores on a spelling test given to fourth grade students assigned to one of two teaching methods, denoted by A and B. In the first and simplest experimental design students are randomly assigned either to the method A classroom or to the method B classroom, with each class taught by a different teacher.

As noted by Lindquist (page 81) the absence of replication means that it is not possible to "...discriminate between the error in results due to the teachers and the real differences due to the methods."

Lindquist goes on to compare the efficiency of two additional designs where replication is used so that valid estimates of variance may be obtained. The first of these (Design III) is a cluster randomized trial. Lindquist warns that standard statistical analyses that ignore the effects of clustering will result in false declarations of statistical significance. These errors may be avoided, he points out by adopting the cluster level analyses described in a previous chapter. To quote Lindquist (page 82) at length...

Design III: Experiment conducted in 10 schools. Five schools, selected at random, use Method A, the other five use method B. Results are evaluated by pooling scores on criterion test in a single distribution for all pupils under each method; by computing the standard error of the mean of each distribution according to formula (I) (page 12), and from these computing the standard error of the difference in means. Difference declared "significant" if three times its standard error.

Comments: This is a design which has actually been employed quite often in educational research. The estimate of error is invalid, even with reference only to errors of the first kind, since the samples are not random samples of pupils. ... The estimate which assumes random sampling of pupils seriously underestimates the error and the results may therefore appear "significant" even though really due to error.

Lindquist's ideas were not initially well received (McNemar 1940; Glass and Stanley 1970 [pp. 501-509]; Oakley 2000), with educational researchers still debating the need to account for the effects of clustering 40 years later (e.g. Barcikowski 1981; Hopkins 1982). It is interesting to speculate as to why the importance of Lindquist's work took so long to be appreciated. An initial difficulty almost certainly was that his work predated the rigorous development of statistical methods for the analysis of correlated outcome data. It was not until 1942 that Hansen and Hurwitz (1942) considered the effects of clustering on binary outcome data arising from complex surveys, while the earliest consideration of these issues in the context of Gaussian outcome data was by Walsh (1947). An additional difficulty faced by Lindquist was that he tended to present his work in textbooks rather than in separate articles (Feldt 1979). Consequently the novelty and importance of his ideas were at risk of being lost among pages devoted to more introductory concepts.

Oakley's excellent historical examination of experimental sociology serves as a reminder that the difficulties faced by Lindquist in putting forward his ideas were not unique (Oakley 2000, Chapter 8). She points out that there was a general move away from using experimental methods in educational research due, in part, to the dearth of studies finding statistically significant intervention effects.

Perhaps as a result of all these obstacles, the ideas put forward by Lindquist had little discernable impact on the quality of medical research, at least in the short run. In fact, Mainland (1952) provides what might be the first attempt in the medical literature to distinguish randomized trials by the unit of allocation. Indeed, in spite of some largely isolated examples, it was not until a brief but seminal article by Cornfield (1978) that these ideas were brought to wide attention to researchers in the health sciences.

References

Amberson JB, McMahon BT, Pinner M (1931). A clinical trial of sanocrysin in pulmonary tuberculosis. *American Review of Tuberculosis* 24:401-35.

Barcikowski RS (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics* 6: 267-285.

- Cornfield J (1978). Randomization by group: a formal analysis. *American Journal of Epidemiology* 108:100-102.
- Donner A, Klar N (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.
- Feldt LS (1979). Everet F. Lindquist (1901-1978). A retrospective review of his contributions to educational research. *Journal of Educational Statistics* 4: 4-13.
- Fisher RA (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Glass GV, Stanley JC (1970). *Statistical Methods in Education and Psychology*. Englewood Cliffs: Prentice-Hall.
- Hansen MH, Hurwitz WN (1942). Relative efficiencies of various sampling units in population inquiries. *Journal of the American Statistical Association* 37:89-94.
- Hopkins KD (1982). The unit of analysis: group means versus individual observations. *American Educational Research Journal* 19:5-18.
- Lindquist EF (1940). *Statistical analysis in educational research*. Boston: Houghton Mifflin.
- Mainland D (1952). *Elementary medical statistics; The principles of quantitative medicine*. Philadelphia: WB Saunders.
- McNemar Q (1940). Book review of Lindquist EF. *Statistical analysis in educational research*. *Psychological Bulletin* 37:746-748.
- Oakley A (2000). *Experiments in Knowing. Gender and Method in the Social Sciences*. New York: The New Press.
- Van Helmont JA (1662). *Oriatrike, or physick refined: The common errors therein refuted and the whole art reformed and rectified*. London: Lodowick-Loyd, p 526.
- Walsh JE (1947). Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics* 18:88-96.

[Home](#)[Contents](#)