Journal of Statistics: Advances in Theory and Applications
Volume 3, Number 1, 2010, Pages 43-59

ON OPTIMAL PLANNING AND ANALYSING EMPIRICAL STUDIES - SEQUENTIAL TESTING AS A STRATEGY FOR REDUCING AVERAGE SAMPLE SIZE

DIETER RASCH and KLAUS D. KUBINGER
Division of Statistics
University of Agriculture Vienna
Austria

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna
Austria
e-mail: Klaus.kubinger@univie.ac.at

Abstract

After explaining seven steps in empirical research, we demonstrate the procedure by using an example from psychological testing. This practical problem can be solved traditionally by first determining the size of an empirical study, performing the study, and then analysing it. But, we show an alternative, sequential approach, where sampling and analysing data take turns. On average, the sequential approach results in smaller sample sizes than the traditional one. In our example for the traditional approach, 59 observations are needed in each of two groups (i.e., 118 in total), whereas the sequential approach required 62 observations in total. The sequential procedure should be used especially, when observations are expensive but recruiting times are short.

2010 Mathematics Subject Classification: 62.

Keywords and phrases: planning experiments and research studies, calculating sample size, type-I-and type-II-risk, requirements of precision, sequential testing.

Received February 27, 2010

1. Introduction

Many (social) scientific disciplines regularly refer to bio- or psychometric methods for empirical studies, but very often suffer in research from mistaken or at least inefficient use of statistical methods. We do not go into detail in this paper as concerns naïvely adhering to the pertinent presupposition of a normal distribution for the variable under discussion (cf. Rasch and Guiard [4]) or the disastrous use of asterisks in interpreting results from hypotheses testing (cf. Rasch et al. [5]). But, we will deal with the matter of inadequate choice of sample sizes or not planning experiments or research studies at all, that is, neglecting the importance of the type-II-risk at hypothesis testing, if H_0 is not to be rejected by a statistical test. As the optimal alternative, we suggest sequential testing, which, in average, leads to a considerably smaller sample size than a-priori calculations.

First of all, we remind readers of the approach of calculating the sample size in advance according to a fixed type-I-, type-II-risk, and a specific effect of content relevance. Secondly, we demonstrate the practicability of sequential testing, i.e., sampling data and testing hypotheses in turn by using special software. Thirdly, we illustrate sequential testing by using an empirical example.

2. Method

2.1. Planning an experiment or a research study

Though, as far as empirical research work takes place, there is no need to distinguish between an experiment and a research study for the following, we briefly remind readers that an experiment is distinguished by some manipulation of several groups of persons (several interesting factor levels) by the researcher, during which group allocation is randomized. In contrast, all other empirical research studies lack that randomization since the interesting groups are given from the very beginning (for more details concerning research studies and experiments, see Rasch [7]).

Any serious empirical research work must include the following seven steps:

1. Exact formulation of the problem.

2. Determining the precision requirements.

3. Selecting the statistical model for planning and analysing-ideally having already determined the method of statistical analysis.

4. Creating the (optimal) design of the (experiment or the) research study.

5. Performing the research study.

6. Statistical analysis of the observed data.

7. Interpretation of the results and conclusions.

Bear in mind that in most cases, the first three stages are not consecutive but should rather be considered as a package. For instance, precision requirements can only be determined once it has been decided how the data should be analysed.

Given the steadily increasing financial problems that research institutions face, the question is how to conduct empirical research as efficiently (cost-effectively) as possible. That is, we are searching for more elaborated statistical approaches for optimization of research planning and analysis. Sequential testing, which will be discussed later in more detail, offers such an approach; and it will be seen that this approach differs in some ways from the seven steps described above. In sequential testing, data are observed and analysed several times consecutively; every time, we either come to a final conclusion or have to continue sampling observations.

2.2. An example-integrating Turkish immigrants in German-speaking countries

Within the globalization age, questions arise dealing with different approaches for integrating immigrants with other ethnicities and cultures into local countries. As concerns children, a need for optimal school administration and application of advancement programmes is clear; this in turn requires comprehensive intelligence testing by school-psychologists. As there are several means of assessing a pupil's intelligence, these means should be evaluated. We will consider the special case of Turkish children that have immigrated into a German-speaking country.

There are two main ways to assess a pupil's intelligence: First, we can apply the Turkish edition of an intelligence test, using a psychologist trained well enough in Turkish to administer the test orally and to recognize whether the examinee's answer is right or wrong-bear in mind that within the given context, intelligence test-batteries are better administered individually, because then a child's working style can be taken into account and specific test materials (for instance puzzles) may be used. Second, we can apply the Turkish edition of the same intelligence test, but use a Turkish psychologist as an examiner. Given the one to two million Turkish children to whom this applies, the latter case would mean employing about 1000 Turkish school psychologists for at least 10 years (though such a large number may not yet exist and would thus need to be trained and recruited). In the former case, it would only be necessary to adequately train about 1000 already employed German school psychologists in Turkish. Therefore, an evaluation study may have immense financial consequences.

Exact formulation of the problem

The given question, sloppily formulated is: Do the two different ways of measuring Turkish children's intelligence differ with respect to the resulting test scores? To answer this question, we create an experimental design by using two different randomly chosen groups of pupils, to whom an intelligence test-battery is administered by the given means. Statistically speaking, there is one (experimental) factor with two levels and a single dependent variable, the test score. So, the general question of the study is:

- What variables and factors emerge and how are they scaled: nominal-, ordered-, or interval-/ratio-scaled?

The factor is nominal-scaled and the observed variable is interval-scaled, because intelligence tests are usually calibrated according to an interval scale, specifically according to the well-known Rasch model (1-PL model; cf., for instance, Kubinger [1]).

The next question:

● is the factor fixed or random?

can be easily answered. The factor's levels are deliberately selected and therefore fixed. We are only interested in the two described intelligence measuring means of Turkish pupils in German-speaking countries.

Next, the question

● what is the minimal difference (from expectations) that is of practical relevance?

must be carefully reflected. The sloppily formulated question "Do the two different ways of measuring Turkish children's intelligence differ with respect to the resulting test scores?" must first be reworded to form a null-hypothesis and an alternative hypothesis (cf. Lehmann [2]). Given $\boldsymbol{y}$, the variable test score, we consider two populations of children, who are administered a test by either a German (*G*) psychologist or by a Turkish (*T*) one; hence the random variable $\boldsymbol{y}$ is defined by the population means μ_G and μ_T. The null-hypothesis is then:

H_0: $\mu_G = \mu_T$, the alternative hypothesis:

H_1: $\mu_G \neq \mu_T$.

Now, the minimal difference (from expectations) that is of practical relevance, δ^*, must be determined. As the error of measurement of intelligence tests is often near 2/3 of the standard deviation, σ, we choose this value $\delta^* = 2\sigma / 3$ to be the critical one we would like to discover with a very high probability, if given. Otherwise, we will not worry about wrongly accepting the null-hypothesis and choose method *G* because of its lower cost, though the test scores of method *G* reach a lower (higher) value in average than those of method *T*.

Furthermore, the question

- which population should the results refer to?

Should be answered. We plan to focus on the regional population, which is at our disposal, that of Austria and Germany.

Determination of precision requirements

To test the given null-hypothesis, we furthermore have to determine

- α, the type-I-risk, i.e., the probability of rejecting the null-hypothesis erroneously,
- β, the type-II-risk, i.e., the probability of accepting the null-hypothesis erroneously.

If the null-hypothesis is rejected wrongly, 1000 Turkish psychologists have to be employed (to more truthfully measure the children's intelligence) or, if the null-hypothesis is accepted wrongly, the children would be inappropriately assessed, leading to an inadequate education, which would in the long run not use and advance the child's resources in proper way. As either of these errors would have immense negative economic consequences, we determine: $\alpha = .05$; $\beta = .05$. The difference of expected means of practical relevance has been already fixed at $\delta^* = 2\sigma / 3$, so that $\delta = \delta^* / \sigma = 2 / 3 = .67$. This means that as long as $\delta_0 > \delta = .67$, the probability of not discovering expected mean differences is at most $\beta = .05$.

Selecting the statistical model for planning and analysing

The pertinent statistical method for testing the given null-hypothesis is student's *t*-test. That is, for statistical modelling, we begin with a normally distributed variable $\boldsymbol{y}$. However, as Rasch and Guiard [4] demonstrated, the assumption of a normal distribution is hardly of relevance, because the *t*-test has proven to be very robust against almost every respective violation.

The traditional approach. So far, the given approach is well known to many researchers who apply statistics. However, many of these researchers ignore the type-II-risk and do not calculate the required sample size in advance, respectively. Consequently, an arbitrarily sample size is chosen and, if the null-hypothesis is accepted, the researcher knows nothing about his/her risk that the null-hypothesis is actually wrong. And, if he/she rejects the null-hypothesis, he/she very often becomes aware that because of the estimated effect size, the respective difference is of almost no practical relevance. We will therefore illustrate how to calculate the sample size in advance, in order to be aware of all the risks of wrong decisions. At any rate, the statistical model is one for a given (preferably equal) size of the two samples. That is, there are two random samples of given size *n* each; the respective variables are commonly suggested (modelled) to be normally distributed (most conveniently with the same variance).

The sequential testing approach. However, there is also an alternative approach. In sequential testing (cf. Wald [10]), we do not fix the sample sizes in advance and separate the phases of planning and analysis. Instead, we repeatedly sample subjects and test the hypotheses in turn. We will here illustrate this approach. Again, the statistical model is one for two random samples, the variables of which are assumed to be normally distributed with the same variance. However, we now use the sequential triangular test (Schneider [9]), which is based on student's two sample *t*-test and is equally robust. Of course, analogous precision requirements must be set $(\alpha = .05;\ \beta = .05,\ \delta = .67)$ before starting the sequential testing algorithm. After sampling the first subject's data (or after sampling a few initial subjects' data), the relevant analysis takes place; if necessary, further data must be sampled. This can be described as a sequence of "observe-analyse-observe-analyse ...," as long as no decision for or against the null-hypothesis is possible. That is, after each step of data-sampling analysis, this approach leads to one of the following decisions: (a) the null-hypothesis is accepted, (b) the null-hypothesis is rejected, (c) data sampling continues with a subject in the group of the first factor level, (d) data sampling continues with a subject in the group of the second factor level. The exorbitant advantage is that, the given precision is usually (on average) reached after testing a smaller sample than would be tested by using the traditional approach. And that again is a very important aspect of empirical studies. Bear in mind that, testing a child requires some organizational effort and roughly 10 hours invested by the psychologist per child. Hence, any child who does not need to be tested because a terminal decision is already possible reduces the cost of the study.

The (optimal) design of the research study

It is optimal to use the same sample size $n_G = n_T$ of Turkish pupils for the intended randomized groups, treated by using the different methods of intelligence testing. This is true, because in this way, the non-centrality parameter

$$NCP = \frac{\mu_G - \mu_T}{\sigma} \sqrt{\frac{n_G \cdot n_T}{n_G + n_T}} \tag{1}$$

of the distribution of the t-statistic is maximized under the alternative hypothesis; this means that for a given difference $\mu_G - \mu_T$, it will be more likely that the t-test results in significance.

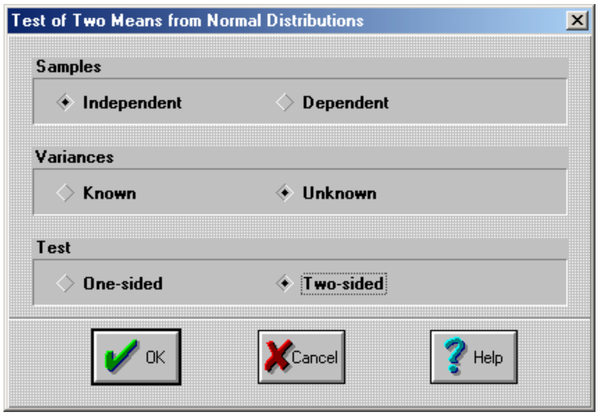
The sample size needed for $\alpha = .05, \beta = .05,$ and $\delta = .67$ may be calculated, for example, according to Rasch [3] as follows (we try to use an intelligence test, which is standardized into T-values, with $\mu = 50$ and $\sigma = 10,$ and thus, $\frac{\sigma^2}{\delta^2} = \frac{1}{0.4489}$):

$$n_G = n_T = n = \left\lceil 2\frac{1}{0.4489} \cdot [t(2n-2;\ 0.025) + t(2n-2;\ 0.05)]^2 \right\rceil, \tag{2}$$

where the square brackets open below mean: smallest integer larger or equal to the result within the brackets. We obtain the solution $n_G = n_T = n$ only by an iterative process. With the aid of the module MEANS from the planning software CADEMO[1], this size can be calculated, however, quite simple as follows: If we start the module MEANS, we get Figure 1, where our cases are already given. Pressing the button “OK” leads to Figure 2. In Figure 2, our necessary input is already completed. Again, pressing “OK” results in Figure 3, where the sample

[1] www.biomath.de

size for the required precision $\delta = .67$ and $s^2 = 1$ is shown as $n_G = n_T = 59$.



Test of Two Means from Normal Distributions

Samples

● Independent ○ Dependent

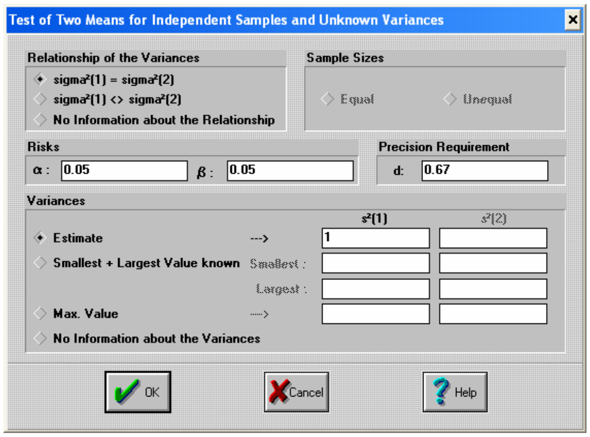
Variances

○ Known ● Unknown

Test

○ One-sided ● Two-sided

Figure 1. CADEMO module MEANS, test of two means from normal distributions



Test of Two Means for Independent Samples and Unknown Variances

Relationship of the Variances

● sigma²(1) = sigma²(2)
○ sigma²(1) <> sigma²(2)
○ No Information about the Relationship

Sample Sizes

○ Equal ○ Unequal

Risks

α : 0.05 β : 0.05

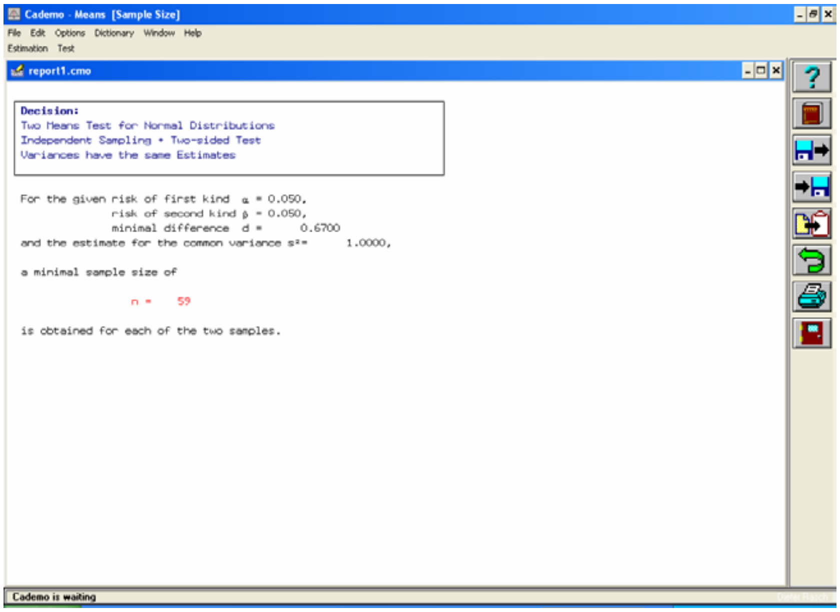
Precision Requirement

d: 0.67

Variances

		s²(1)	s²(2)
● Estimate	--->	1	
○ Smallest + Largest Value known	Smallest :		
	Largest :		
○ Max. Value	--->		
○ No Information about the Variances			

Figure 2. CADEMO module MEANS–input of the precision requirements


Cademo - Means [Sample Size]
File Edit Options Dictionary Window Help
Estimation Test
report1.cmo
Cademo is waiting

```
Decision:
Two Means Test for Normal Distributions
Independent Sampling • Two-sided Test
Variances have the same Estimates

For the given risk of first kind  α = 0.050,
              risk of second kind β = 0.050,
              minimal difference  d =     0.6700
and the estimate for the common variance s²=      1.0000,

a minimal sample size of

                n =    59

is obtained for each of the two samples.
```

Figure 3. CADEMO module MEANS-output of the minimum sample size

However, as indicated in the section before, we could also proceed by using sequential testing.

Performing the research study

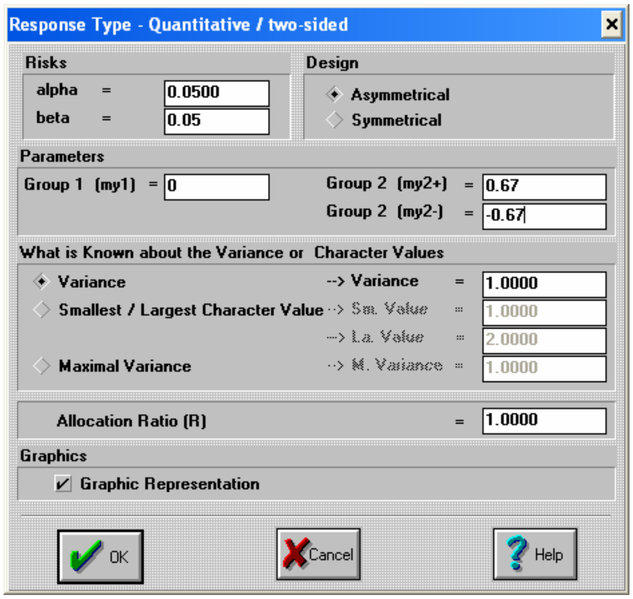
In performing our planned experiment, we have to sample the data.

Statistical analysis of the observed data

The method of statistical analysis determined in step 3 must then be applied.

In our example, we could follow the traditional approach and apply the *t*-test as indicated. However, we will here use the sequential testing approach, applying the triangular test.

Usually, one begins with two observations from each of the two groups and then gathering further observations if needed, alternating between groups 1 and 2. We use the module TRIQ from the software CADEMO. It offers both a numerical and a graphical output of the results as the sequential procedure takes place step by step. In the graphical output, there is either one (if the alternative hypothesis is one-sided) or two (if the alternative hypothesis is two-sided) intersecting triangles. Their shadowed area represents the situation that neither a decision for the null-hypothesis nor a decision for the alternative hypothesis is possible and data sampling has to continue. The triangular areas are as already indicated determined by a statistic based on the student's *t*-test. When the path representing the step by step calculation of this statistic leaves the triangle(s) for the first time, analysis stops and either the null- or the alternative hypothesis is to accept. All we have to do now is the following: From the introductory screen, we select "Design" and click on "Quantitative / two-sided"; this leads to the screenshot shown in Figure 4, where our precision requirements are already defined.



Response Type - Quantitative / two-sided

Risks

alpha = 0.0500

beta = 0.05

Design

● Asymmetrical

○ Symmetrical

Parameters

Group 1 (my1) = 0

Group 2 (my2+) = 0.67

Group 2 (my2-) = -0.67

What is Known about the Variance or Character Values

● Variance --> Variance = 1.0000

○ Smallest / Largest Character Value --> Sm. Value = 1.0000

--> La. Value = 2.0000

○ Maximal Variance --> M. Variance = 1.0000

Allocation Ratio (R) = 1.0000

Graphics

☑ Graphic Representation

Figure 4. CADEMO module TRIQ – input of the precision requirements

Now, we again use the introductory screen of TRIQ and click "Analysis", then choose "Quantitative/two-sided" (see Figure 5). "OK" leads to the window for entering the data, where we show the situation where 10 test scores for each of the groups have already been sampled (see Figure 6). Because of our calculation above, we do not believe that any terminal decision is realistic before sampling at least that much data. Applying the sequential triangular test leads to Figure 7, which proves our belief.

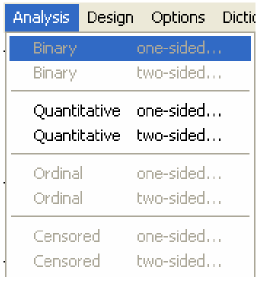
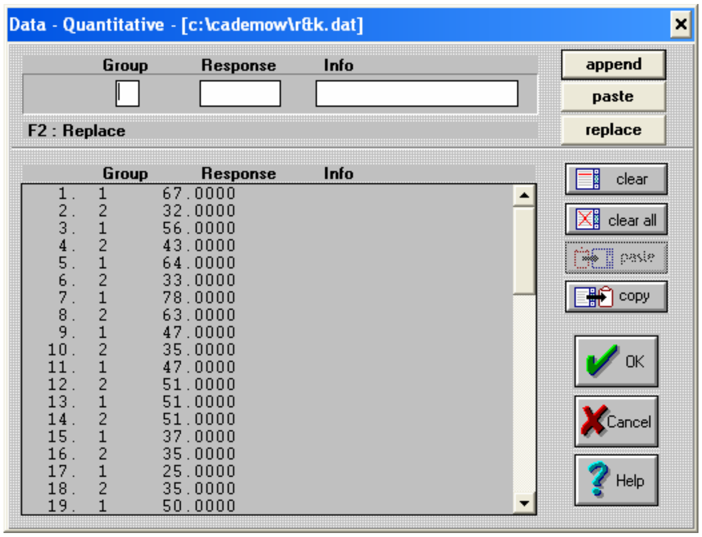
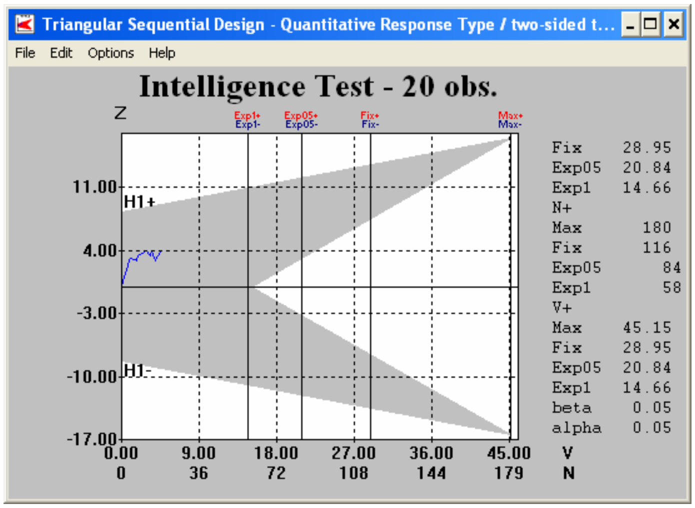


Figure 5. CADEMO module TRIQ-selection of the type of the data



	Group	Response	Info
1.	1	67.0000	
2.	2	32.0000	
3.	1	56.0000	
4.	2	43.0000	
5.	1	64.0000	
6.	2	33.0000	
7.	1	78.0000	
8.	2	63.0000	
9.	1	47.0000	
10.	2	35.0000	
11.	1	47.0000	
12.	2	51.0000	
13.	1	51.0000	
14.	2	51.0000	
15.	1	37.0000	
16.	2	35.0000	
17.	1	25.0000	
18.	2	35.0000	
19.	1	50.0000	

Figure 6. CADEMO module TRIQ-the first 19 sampled test scores (observation 20 equals 35)

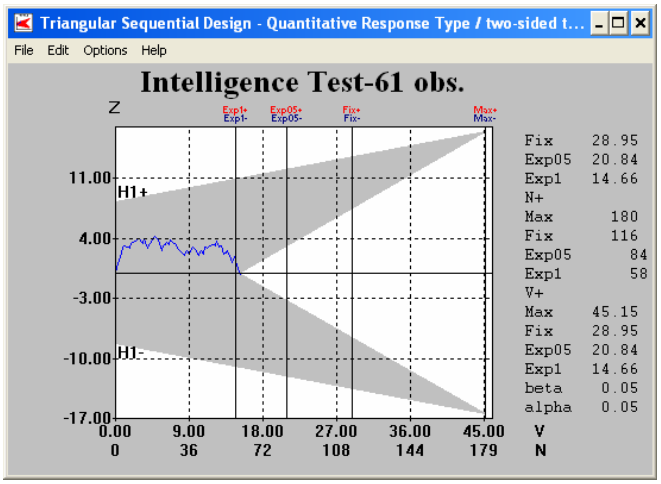


```
Fix     28.95
Exp05   20.84
Exp1    14.66
N+
Max       180
Fix       116
Exp05      84
Exp1       58
V+
Max     45.15
Fix     28.95
Exp05   20.84
Exp1    14.66
beta     0.05
alpha    0.05
```

Figure 7. The analysis of TRIQ after 20 observations in both samples altogether

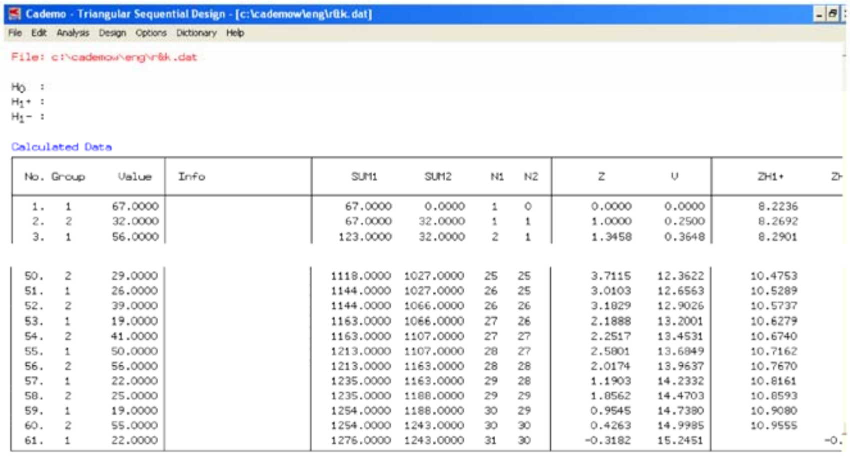
Hence, we sampled and tested 10 pupils more for each of the two groups, and then, after realizing that no terminal decision was still possible, we sampled alternately one pupil after the other and tested them step by step[2]. We illustrate the results for the case $n_G = 31$, $n_T = 30$ (given in Figure 8). While, the graphical output does for the first time not disclose unequivocally whether or not the analysis has to be stopped and the null-hypothesis accepted, the numerical output of CADEMO does (cf. Figure 9): We must sample at least one more pupil. If we actually take the next sampled test score of 37 in group 2 into account, then the analysis results in the final output shown in Figures 10 and 11: 31 and 31 pupils suffice. The null-hypothesis must be accepted.

[2] The data were originally used by Kubinger [1].


Triangular Sequential Design - Quantitative Response Type / two-sided t...
File Edit Options Help
Intelligence Test-61 obs.
Z
Exp1+ Exp1-
Exp05+ Exp05-
Fix+ Fix-
Max+ Max-
11.00
4.00
-3.00
-10.00
-17.00
H1+
H1-
0.00 9.00 18.00 27.00 36.00 45.00 V
0 36 72 108 144 179 N

```
Fix     28.95
Exp05   20.84
Exp1    14.66
N+
Max       180
Fix       116
Exp05      84
Exp1       58
V+
Max     45.15
Fix     28.95
Exp05   20.84
Exp1    14.66
beta     0.05
alpha    0.05
```

Figure 8. The analysis of TRIQ after 61 observations in both samples altogether


Cademo - Triangular Sequential Design - [c:\cademow\eng\r&k.dat]
File Edit Analysis Design Options Dictionary Help

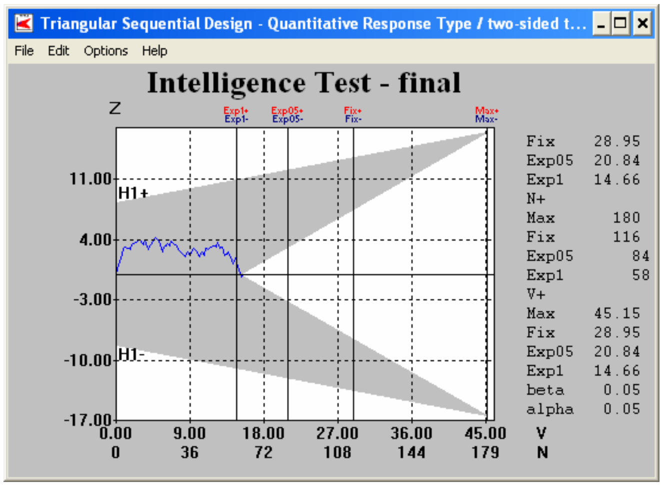
File: c:\cademow\eng\r&k.dat

H_0 :
H_1+ :
H_1- :

Calculated Data

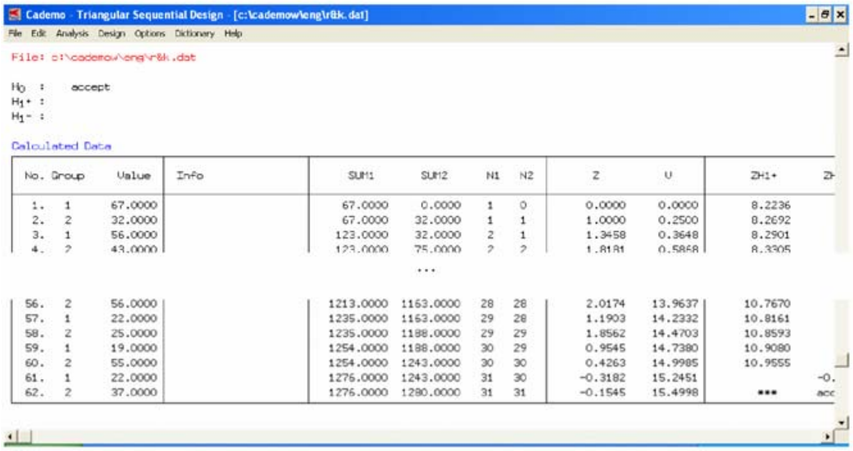
No.	Group	Value	Info	SUM1	SUM2	N1	N2	Z	V	ZH1+	Z
1.	1	67.0000		67.0000	0.0000	1	0	0.0000	0.0000	8.2236	
2.	2	32.0000		67.0000	32.0000	1	1	1.0000	0.2500	8.2692	
3.	1	56.0000		123.0000	32.0000	2	1	1.3458	0.3648	8.2901	
50.	2	29.0000		1118.0000	1027.0000	25	25	3.7115	12.3622	10.4753	
51.	1	26.0000		1144.0000	1027.0000	26	25	3.0103	12.6563	10.5289	
52.	2	39.0000		1144.0000	1066.0000	26	26	3.1829	12.9026	10.5737	
53.	1	19.0000		1163.0000	1066.0000	27	26	2.1888	13.2001	10.6279	
54.	2	41.0000		1163.0000	1107.0000	27	27	2.2517	13.4531	10.6740	
55.	1	50.0000		1213.0000	1107.0000	28	27	2.5801	13.6849	10.7162	
56.	2	56.0000		1213.0000	1163.0000	28	28	2.0174	13.9637	10.7670	
57.	1	22.0000		1235.0000	1163.0000	29	28	1.1903	14.2332	10.8161	
58.	2	25.0000		1235.0000	1188.0000	29	29	1.8562	14.4703	10.8593	
59.	1	19.0000		1254.0000	1188.0000	30	29	0.9545	14.7380	10.9080	
60.	2	55.0000		1254.0000	1243.0000	30	30	0.4263	14.9985	10.9555	
61.	1	22.0000		1276.0000	1243.0000	31	30	-0.3182	15.2451		-0.

Figure 9. The numerical output of TRIQ after 61 observations; no hypothesis is to be accepted



```
Fix     28.95
Exp05   20.84
Exp1    14.66
N+
Max       180
Fix       116
Exp05      84
Exp1       58
V+
Max     45.15
Fix     28.95
Exp05   20.84
Exp1    14.66
beta     0.05
alpha    0.05
```

Figure 10. The analysis of TRIQ after 62 observations in both samples altogether


Cademo - Triangular Sequential Design - [c:\cademow\eng\r&k.dat]
File Edit Analysis Design Options Dictionary Help

File: c:\cademow\eng\r&k.dat

```
H0  :    accept
H1+ :
H1- :
```

Calculated Data

No.	Group	Value	Info	SUM1	SUM2	N1	N2	Z	V	ZH1+	ZH
1.	1	67.0000		67.0000	0.0000	1	0	0.0000	0.0000	8.2236	
2.	2	32.0000		67.0000	32.0000	1	1	1.0000	0.2500	8.2692	
3.	1	56.0000		123.0000	32.0000	2	1	1.3458	0.3648	8.2901	
4.	2	43.0000		123.0000	75.0000	2	2	1.8181	0.5868	8.3305	
...											
56.	2	56.0000		1213.0000	1163.0000	28	28	2.0174	13.9637	10.7670	
57.	1	22.0000		1235.0000	1163.0000	29	28	1.1903	14.2332	10.8161	
58.	2	25.0000		1235.0000	1188.0000	29	29	1.8562	14.4703	10.8593	
59.	1	19.0000		1254.0000	1188.0000	30	29	0.9545	14.7380	10.9080	
60.	2	55.0000		1254.0000	1243.0000	30	30	0.4263	14.9985	10.9555	
61.	1	22.0000		1276.0000	1243.0000	31	30	-0.3182	15.2451		-0.
62.	2	37.0000		1276.0000	1280.0000	31	31	-0.1545	15.4998	***	acc

Figure 11. The numerical output of TRIQ after 62 observations; the null-hypothesis is accepted

Interpretation of the results and conclusions

Of course, any interpretation based not only on significance or non-significance, but also on relevant (estimated) effect size can be more easily given. In our example, we can conclude that both methods lead to the same average test score. As a consequence, we can choose either of the two methods under consideration – and of course, we will take the cheaper one of using German school psychologists sufficiently trained in Turkish.

3. Discussion

In our example, we accepted the null-hypothesis and we are, of course, confident that this decision is correct. But nevertheless, we are aware that the decision could also be wrong. By choosing a type-II-risk of 5 percent, we used a method that guarantees with a 0.95 probability that we do not overlook a practically relevant difference in expected means (i.e., a difference of two thirds of the standard deviation). This is a probability statement and therefore is valid only for the statistical method as a whole, but does not refer to a single study. That is, if we proceed with this probability in our research, then all conclusions based on an accepted null-hypothesis are right in about 95 % of all cases and wrong in about 5 %. Which is the case in the present study is completely unknown. And being wrong in the given empirical study means that, the methods under consideration differ by more than two thirds of the standard deviation. Differences smaller than two thirds of the standard deviation will occur more often; the corresponding probabilities (as a function of the true mean difference) of those cases are, however, not usually calculated.

We indicated with this example how two population means can be compared by using the classical approach, where the sample size is fixed in advance according to given precision requirements and a student's two-sample *t*-test is applied. This approach is exact but has a disadvantage. As demonstrated in the given example, the alternative approach of sequential testing by using Schneider's sequential triangular test needs considerably smaller samples to meet the same precision requirements (and this is also true on average). While in our example, the traditional approach demanded 59 plus 59 subjects, we actually only needed 31 plus 31. Indeed, our experience with a large number of empirical studies has shown that using the sequential triangular test never led to a larger sample size than that needed for the traditional approach.

For this reason, we recommend the demonstrated sequential approach. This is true especially, when observations are expensive and have a short recruiting time, and it also applies to various application fields in the social sciences, medicine, agriculture, and economics.

References

[1] K. D. Kubinger, Adaptives Intelligenz Diagnostikum-Version 2.2 (AID 2) samt AID 2-Türkisch [Adaptive Intelligence Diagnosticum, AID 2-Turkey included], Beltz, Göttingen, (2009).

[2] L. E. Lehmann, Testing Statistical Hypothesis, Wiley, New York, (1959).

[3] D. Rasch, Determining the optimal size of experiments and surveys in empirical research, Psychology Science 45 (2003), 3-48.

[4] D. Rasch and V. Guiard, The robustness of parametric statistical methods, Psychology Science 46 (2004), 175-208.

[5] D. Rasch, K. D. Kubinger, J. Schmidtke and J. Häusler, The misuse of asterisks in hypothesis testing, Psychology Science 46 (2004), 227-242.

[6] D. Rasch and K. D. Kubinger, Statistik für das Psychologiestudium-Mit Softwareunterstützung zur Planung und Auswertung von Untersuchungen sowie zu sequentiellen Verfahren [Statistics for the study of psychology-software supplication of planning and sequential procedures], Spektrum, München, (2006).

[7] D. Rasch, L. R. Verdooren and J. I. Gowers, Fundamentals in the Design and Analysis of Experiments and Surveys, Oldenbourg, München, (2007).

[8] D. Rasch, G. Herrendörfer, J. Bock, N. Victor and V. Guiard, Verfahrensbibliothek Versuchsplanung und-auswertung (2nd ed.), Electronical book with CD [Handbook of planning and analysis of experiments and surveys, electronic book with CD], Oldenbourg, München, (2008).

[9] B. Schneider, An interactive computer program for design and monitoring of sequential clinical trials, In Proceedings of the 16th International Biometric Conference (S. 237-250), Hamilton, New Zealand, (1992).

[10] A. Wald, Sequential Analysis, Wiley, New York, (1947).

[11] J. Whitehead, The design and analysis of sequential clinical trials, Ellis Horwood, Chichester, (1983).

■