METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Tutorial

Group Sequential Designs Applied in Psychological Research

Klemens Weigl[abc], Ivo Ponocny[d]

[a] *Department of Psychology, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany.* **[b]** *Human-Computer Interaction Group, Technische Hochschule Ingolstadt, Ingolstadt, Germany.* **[c]** *Institute for Health Services Research and Clinical Epidemiology, School of Medicine, Philipps-Universitaet Marburg, Marburg, Germany.* **[d]** *Department for Sustainability, Governance, and Methods, MODUL University Vienna, Vienna, Austria.*

Methodology, 2020, Vol. 16(1), 75–91, https://doi.org/10.5964/meth.2811

Received: 2017-08-02 • **Accepted:** 2019-07-15 • **Published (VoR):** 2020-04-06

Corresponding Author: Klemens Weigl, Ostenstraße 25, 85072 Eichstätt, Germany

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Psychological research is confronted with ever-increasing demands to save resources such as time and money while assuring high ethical standards. In medical and pharmaceutical research, group sequential designs have fundamentally changed traditional statistical testing approaches featuring only one analysis at the end of a single-stage study. They enable early stopping at an interim stage, after a group of observations, for efficacy or futility in case of an overwhelmingly large or small effect, respectively. Otherwise, the trial is continued to the next stage. On average over many studies time and money are saved and more ethical trials are facilitated by diminishing the risk of patients' exposure to inferior treatments. We provide an easy-to-use tutorial for psychological research replete with easily understandable figures highlighting the core idea of different group sequential designs, a workflow chart, an empirical real-world data set, and the annotated R code. Finally, we demonstrate the application of early stopping for efficacy.

Keywords

group sequential designs, interim analyses, workflow chart, R Code, tutorial

Psychological researchers often conduct scientific studies under time pressure, while having to deal with financial and organizational constraints. Additionally, in all psychological investigations they are obliged to assure the highest possible ethical standards. For several decades, the very same challenges have been tackled by biostatisticians in medical and pharmaceutical research by the introduction of *interim analyses*. The alluring benefits of the statistical and methodological framework of interim analyses are saving resources such as time and money, and reducing organizational effort while

meeting high ethical standards. They have been thoroughly investigated and positively proven in medical and pharmaceutical research, over many scientific studies. These innovative statistical procedures of interim analyses have tremendously extended traditional frequentist confirmatory statistical testing approaches with *only one* statistical analysis at the end of the trial. When interim analyses are applied, data are analyzed at intervals and statistical testing is performed after a pre-specified number of $K \geq 2$ stages. In case of a sufficiently large effect at interim, the trial is stopped for efficacy; in case of an overwhelmingly small effect it is stopped for futility. Otherwise, the trial is continued to the next consecutive stage.

The genesis of interim analyses starts with fully sequential tests introduced by Wald (1943, 1945, 1947), wherein a statistical test is performed after each observation. The high practical burden of such frequent statistical testing limits the feasibility of these procedures. A great improvement of this fully sequential testing approach and a major impetus for group sequential testing came from Pocock (1977) and O'Brien and Fleming (1979) through the development of *group sequential designs* (also referred to as *group sequential methods*). Statistical testing is performed after an a priori fixed number of participants per group and per stage, yielding *equally-sized* stages (also referred to as *equally-spaced* information fractions). Further well-known approaches have been introduced by Haybittle (1971) and Peto et al. (1976; referred to as the *Haybittle-Peto* approach; rarely applied), and Wang and Tsiatis (1987).

Interim analyses facilitate more ethical practices. Superior treatments and therapies can be identified at an earlier stage and applied more widely, while harmful ones can be stopped prematurely. Given the manifold strengths of interim analyses, which are continually being refined by biostatisticians in academia, industry, and statistics agencies around the world, they open up numerous applications for the scientific field of psychological research.

Unfortunately, the Publication Manual of the American Psychological Association (2010) describes interim analyses in only one, potentially misleading, sentence: "If interim analysis and stopping rules were used to modify the *desired* sample size, describe the methodology and results." (p. 30). This sentence is quite similar and only slightly extended in the newer edition of the Publication Manual of the American Psychological Association (2020): "If interim analysis and stopping rules were used to modify the desired sample size, describe the methodology and results *of applying that methodology*." (p. 84). In a strict sense, this description is inaccurate, because interim analysis and stopping rules are not used to modify the desired sample size. Rather, they are applied to reduce, on average, the number of allocated participants per group within a scientific trial, and therefore save resources such as time and money, while enabling more ethical psychological studies.

As yet, the only primer for the application of group sequential designs in psychological research was published by Lakens (2014) in social psychology. We go beyond an

exemplified application of these highly sophisticated statistical procedures. In doing so, we provide a tutorial based on (i) easily understandable figures (see Figure 1 and Figure 2) and accompanying explanations of the core idea of group sequential designs, (ii) a workflow chart (see Appendix B) featuring all of the important steps for a concise application, (iii) the briefly annotated *R* code for further usage (Weigl & Ponocny, 2020), (vi) a demonstration of how to apply sample size estimation based on the sample size inflation factor (IF), (v) an elucidation of early stopping for efficacy on an illustrative example for a two groups comparison, and (vi) a real-world data set from psychological research for teaching purposes (Weigl & Ponocny, 2020). These six aspects have not been addressed by any other tutorial in psychological research, yet such a tutorial on the important methodological and statistical aspects of applying group sequential designs in psychological research is urgently necessary.

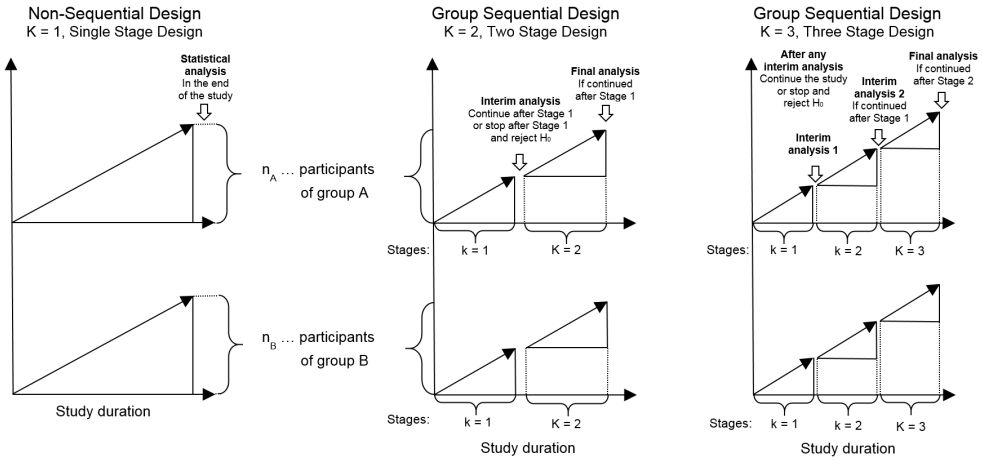
This article is organized as follows. The section "Group Sequential Designs" and Figure 1 provide a simple introduction to the core idea of these statistical procedures. Section "Different Group Sequential Approaches" outlines the most important group sequential designs of the approaches by Pocock (1977), O'Brien and Fleming (1979), Haybittle (1971) and Peto et al. (1976), as well as Wang and Tsiatis (1987). After a graphical comparison of the outlined approaches in Figure 2, an illustrative example explains how to design a group sequential psychological study, perform sample size estimation using the IF, and conduct group sequential testing and early stopping for efficacy, by applying the workflow chart in Appendix B. The last section discusses the potential for future applications in psychological research. In Appendix A, an introduction to recommended software solutions is provided. The annotated *R* code for the simulation of group sequential designs and for sample size estimation using the IF as well as the data set are provided on PsychArchives (data: Weigl & Ponocny, 2020a, code: Weigl & Ponocny, 2020b).

Group Sequential Designs

The core idea of group sequential designs is *repeated significance testing*, with one test following each group of observations. For didactic reasons, this idea is illustrated in Figure 1 for group sequential designs with $K = 2$ and $K = 3$ stages, with each compared to a non-sequential design for a two groups comparison. As indicated in both group sequential graphs, after each interim analysis the decision has to be made of whether to stop or continue the ongoing psychological study, depending on whether or not the observed effect (e.g., treatment difference) of the collected data exceeds the respective rejection boundary, i.e., the Z-boundary or the corresponding *nominal* significance level (i.e., the adjusted significance level at stage k) of the *a priori* chosen group sequential approach. The data can be collected continually in an accumulative process, or all data can be assessed at once (e.g., in a psychological group testing situation).

Figure 1

Non-Sequential Design and Group Sequential Designs With Two and Three Stages



The statistical idea behind repeated significance testing is based on a specific numerical recursive integration formula introduced by Armitage, McPherson, and Rowe (1969), and McPherson and Armitage (1971), which addresses the independent increment structure of the underlying process of data accumulation. However, repeated observation of the data increases the family-wise Type I error rate, if not controlled for. To avoid Type I error inflation (false positives) and to obtain a level α testing procedure, all the approaches of group sequential designs adjust the *Z*-boundaries and the nominal α levels at each stage yielding appropriately adjusted decision regions (i.e., continuation or rejection regions, see Figure 2, and the workflow chart in Appendix B).

Jennison and Turnbull (2000), and Wassmer and Brannath (2016) provide further statistical background to group sequential designs (e.g., formulas, theorems, and mathematical proofs). Hence, the additional statistical background is not discussed here because of the didactic mission of the present article.

Different Group Sequential Approaches

Repeated significance testing increases the probability of obtaining false-positive results if the Type I error rate α is not adjusted according to the number of repeated significance tests. For performing interim analyses there are several different approaches on how to split the Type I error probability α, which leads to different rejection boundaries and group sequential tests with different stopping rules. As already mentioned, the

most well-known group sequential approaches were developed in the scientific field of medical and pharmaceutical statistics. Pocock (1977) coined the term *group sequential methods* with his major contribution of a group sequential test for repeated significance testing after equally sized groups of observations with $n_1 = ... = n_K$ for two-sided testing scenarios. Haybittle (1971) and Peto et al. (1976), O'Brien and Fleming (1979), and Wang and Tsiatis (1987) developed different approaches, all with specific advantages. These can all be either formulated and applied via the Z-statistic (by transforming the statistic of interest into a Z-statistic following the standard normal distribution), or the corresponding significance level approach (using the p value obtained by the statistic of interest). Hence, in making statistical decisions at interim stage k or at final stage K, either the adjusted Z-boundaries or the nominal significance levels of the chosen group sequential approach for the respective stage can be applied. The significance level approach may be more familiar to psychological researchers because of their affinity for expressing statistical results in terms of p values[1]. To date, no gold standard has emerged among the various group sequential approaches as the approach-of-choice. Therefore, it is possible to choose any group sequential approach prior to commencement of the psychological study. However, the most widely applied group sequential designs are the approaches by O'Brien and Fleming (1979), and Wang and Tsiatis (1987). As long as Δ, the power parameter to adjust the rejection boundaries of the Wang and Tsiatis design, is not set to $\Delta = 0.5$ (which would approximate the boundaries of the Pocock, 1977, approach, whereas $\Delta = 0$ yields the boundaries of the O'Brien and Fleming, 1979, approach), both approaches provide monotonically decreasing rejection boundaries. The approach of O'Brien and Fleming enables testing nearly at full level α at the last stage K (see Figure 2).

Comparison of the Approaches

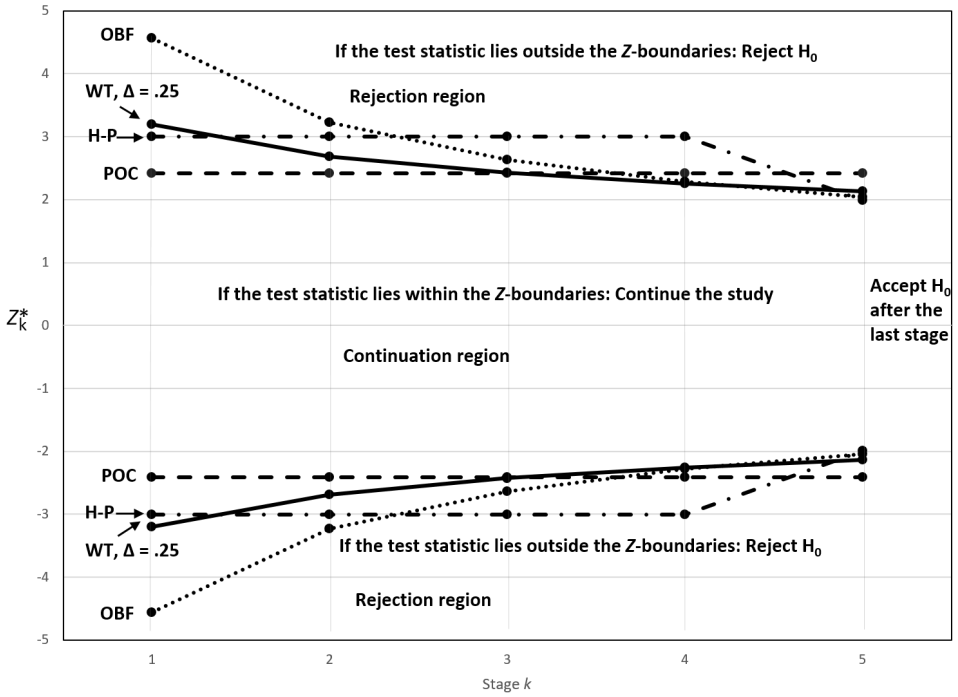
In Figure 2, the group sequential approaches of Pocock (POC), O'Brien and Fleming (OBF), Wang and Tsiatis with $\Delta = 0.25$ (WT), and Haybittle-Peto (H-P) are depicted. Though in practice interim analyses are mostly applied only with $K = 2$ and $K = 3$ stages, for illustrative purposes, $K = 5$ stages have been chosen to visualize the differences of the rejection boundaries of these approaches.

Comparing the boundaries of the Pocock (1977) approach with those of O'Brien and Fleming (1979) reveals a higher probability of rejecting the null hypothesis at an earlier stage for the approach by Pocock. This property changes at later stages, such that H_0 is rejected more easily if the O'Brien and Fleming approach is applied. Hence, Pocock's test is more liberal at earlier stages and more conservative at later stages than the approach of O'Brien and Fleming. The specific advantage of the Wang and Tsiatis (1987) approach is the possibility to adjust the power parameter Δ, whereby $\Delta = 0.25$ yields intermediate

1) The p value is the probability of obtaining a result at least as extreme as the one actually observed if the null hypothesis H_0 is true.

Figure 2

Comparison of Group Sequential Designs



Note. *Z*-boundaries of Pocock (POC), O'Brien and Fleming (OBF), Wang and Tsiatis with Δ = 0.25 (WT, Δ = 0.25), and Haybittle-Peto (H-P) for $K = 5$ a priori planned stages, respectively.

rejection boundaries between the boundaries of Pocock and of O'Brien and Fleming. The approach of Haybittle (1971) and Peto et al. (1976) slightly exceeds the overall Type I error rate at level α. Therefore, the constant for the rejection boundary at the last stage can be adjusted to provide a group sequential test with Type I error rate precisely at level α. However, if the maximum number of K stages is small, early stopping and rejecting H_0 is rather unlikely. Hence, Haybittle (1971), and Peto et al. (1976) is not recommended. Nevertheless, it is depicted for didactic reasons.

Method

We demonstrate the application of a group sequential design on real-world data from psychological research. For this purpose, we follow the workflow chart (see Appendix B).

Participants

The data were collected during a study on living conditions (Ponocny, Weismayer, Dressler, & Stross, 2015, 2016) which was designed as a pilot study for a detailed quality of life assessment, combining questionnaires (administered in both online and paper-based versions), interviews, and diary data. Because of the in-depth personal interview component, the sample was chosen from 10 different locations in Austria, ranging from urban to rural communities. The respondents were randomly selected from local population registers, telephone books, or commercial address lists - dependent on the particular data availability - and sent a paper-pencil questionnaire, with the additional option to complete it online. In one participating town, the questionnaire was included in the local community newspaper. In total, 1454 persons responded to the questionnaire; the completion rate could not be assessed exactly for organisational reasons, but is not much more than 5%. Therefore, self-selection effects cannot be ruled out, although demographic indicators do not point to essential biases apart from an overrepresentation of women (60%).

Instruments and Materials

The questionnaire about various quality-of-life aspects, in particular subjective information, was constructed based on qualitative interviews about good and bad circumstances in life, which had previously been conducted as part of the same study. This procedure sought to ensure the relevance of items for subjective experience, and that the language used reflected how respondents judge and think. In particular, statements were generated about the perception of the personal sense of life - such as "I live in harmony with myself", "I had to accept things as they are", or "my problems cast a shadow on my life" - which were then presented to participants. Since the questions were updated over the course of the study based on ongoing analyses, participants responded, in part, to different item selections. The selection chosen for this demonstration allows for the analysis of 611 cases with a common complete item set, which was aggregated to a single positivity score consisting of 17 items. However, because of the didactic mission, we only selected the first 176 cases of the $N = 611$ (88 women and 88 men; these sample sizes were computed according to the sample size inflation factor of IF = 1.034; see Section "Workflow Chart: 3. Perform Sample Size Estimation Using the Sample Size Inflation Factor (IF)"). Participants were instructed to tick those items with which they perceived a feeling of agreement, in which case the value 1 was assigned, and to leave the other items blank; unchecked items were scored as 0. Items 1, 3, and 8 were positively coded. All of the negatively coded items were recoded before all 17 items were summated. In our illustrative example, the range of the overall sum score was between 4 and 16 over all 176 cases.

Procedure

One aim of the study of living conditions was to collect as many responses as possible (restricted only by budget) in order to accumulate a rich data set which also represents smaller subpopulations that had not been explicitly considered (i.e., by oversampling). Therefore, interim testing was not considered during the process of data accumulation. However, the character of the data set (and its sample size) perfectly allows for a simulation which applies the assumption that economic and time resources are flexible, but should be conserved to the extent practicable. In this case, interim testing could have been suggested and planned a priori. Moreover, the simulation based on the real data set shows what the result would have been and what benefits in terms of resource use could have been realized through early identification and stopping for efficacy at an earlier stage *k*.

Hypothesis of Interest

Before we select the statistical model, we specify the hypothesis of interest. Given our data set, we are interested in the dependent variable, *perception of personal sense of life*, which was tested on the grouping variable, *gender* (women (w) vs. men (m)). We want to test the null hypothesis H_0: $\mu_{(w)} = \mu_{(m)}$: Women and men have roughly the same positive *self-rating* of the variable perception of the personal sense of life. We test the null hypothesis against the *two-sided*[2] alternative H_1: $\mu_{(w)} \neq \mu_{(m)}$: Women and men have a different self-rating of the variable perception of personal sense of life.

Our hypothesis of interest is based on the scientific background that women and men do not generally show consistent systematic differences regarding their self-ratings in life satisfaction or happiness, see, for example, Dolan, Peasgood, and White (2008); Meisenberg and Woodley (2015); or Diener, Suh, Lucas, and Smith (1999). In this context, there is no reason to assume large effect sizes, but planning the sample size based on the assumption of at least a medium effect seems reasonable if smaller effects are not considered as relevant. On the other hand, it has been shown that standard scales often fail to depict very important circumstances in life (Ponocny et al., 2016): a problem which may be overcome by the score reflecting global emotional perceptions of life. This raises the question of whether men and women show differences when asked in the manner of the items involved in the score construction.

2) Due to the widely used notations of *one-* and *two-sided* testing in the literature of interim analyses addressing *one-* and *two-tailed* testing, respectively, we also use these terms throughout this article.

Workflow Chart

1. Select the Statistical Model

We are interested in the difference between two independent means (two groups comparison of women and men) of the dependent variable and perform Student's independent two sample *t*-test. The test statistic θ is the *t*-statistic estimated by the data sampled from both groups. Furthermore, we choose the Type I error rate α as two-sided overall significance level $\alpha_{(\text{two-sided})} = .05$ (yielding a one-sided overall significance level $\alpha_{(\text{one-sided})} = .025$), and Type II error rate $\beta = .1$ for 90% power at the expected medium standardized effect size $d = .5$, in accordance with Cohen (1969), and Cohen (1988).

2. Choose One Group Sequential Approach and Design the Psychological Study

After the identification of the statistical model, we arbitrarily choose the group sequential approach by Wang and Tsiatis (1987; see workflow chart in Appendix B). Thereby, we select $K = 2$ stages and simulate the group sequential design with the already specified quantities $\alpha_{(\text{two-sided})} = .05$, $\beta = .1$ (yielding 90% power), $\Delta = .25$ (yielding intermediate rejection boundaries between the boundaries of Pocock, 1977, and O'Brien and Fleming, 1979) with the statistical software *R* (R Core Team, 2019) and the *R* package gsdesign (Anderson 2016; see *R* code).

3. Perform Sample Size Estimation Using the Sample Size Inflation Factor (IF)

Though a priori sample size estimation is different for fixed sample tests (with *no* interim analysis) and for group sequential designs, it is recommended in both settings.

First, we use the *R* package *pwr* (Champely, 2018) and compute the exact sample size for Student's independent two sample *t*-test. Thereby we assume $\alpha = .05$ (two-sided), $\beta = .1$ (for 90% power), and a medium effect size of $d = 0.5$, which yields $n = 85.03$ in each group and a total sample size of $N_{(\text{total})} = 170.06$ for the classical fixed sample design with $K = 1$ stage and no interim analyses. G*Power, Version: 3.1.9.4 (Faul, Erdfelder, Buchner, & Lang, 2009), the popular software for sample size estimation, should not be applied in the context of group sequential setting. Though the exact sample size for Student's independent two sample *t*-test is also precisely (internally) estimated, the software only provides the already *rounded* sample size $n_{(\text{r.})}$ (r. denotes *rounded*; i.e., G*Power provides $N_{(\text{total,r.})} = 172$). Hence, in certain cases the sample size may be artificially increased, which, when it is multiplied by the sample size inflation factor (IF; see below), would result in a slightly overpowered study.

Second, we multiply $N_{(\text{total})} = 170.06$ by the IF 1.034 (simulated by the *R* code) for the chosen Wang and Tsiatis (1987) design with $\Delta = .25$, $\alpha = .05$ (two-sided), and $\beta = .1$, and obtain $N_{(\text{total,adj.})} = 175.84$ which yields $N_{(\text{total,adj.,r.})} = 176$ and $n_{\text{A1}} = n_{\text{B1}} = 44$ for Stage 1 and $n_{\text{A2}} = n_{\text{B2}} = 44$ for Stage 2 for the two groups A and B, respectively.

Results

In many cases, if the statistical assumptions are not a priori known, nonparametric statistical procedures are recommended and planned in a study protocol. However, for didactic reasons and demonstrative purposes we retrospectively applied Student's independent two sample *t*-test and the group sequential Wang and Tsiatis (1987) design on already sampled psychological data. After data sampling for Stage 1 with $n_{A1} = n_{B1} = 44$ women and men, we perform an interim analysis.

After Stage 1: Stopping for Efficacy

The interim analysis after Stage 1 revealed a sufficiently large effect ($t(86) = -2.71$, $p_{1(\text{one-sided})} = .00407$, Cohen's $d = 0.58$, achieved power = 76%). If the significance level approach is applied, the one-sided p value $p_{1(\text{one-sided})} = .00407$ is smaller than the nominal p value $p_{(\text{nominal})} = .0077$ for $\alpha = .025$ (Note: (1) p_1 refers to the p value after Stage 1; (2) The statistical test decision of one-sided testing and the comparison with the respective one-sided nominal significance level is numerically identical to two-sided testing and the comparison with the two-sided p value; i.e., $p_{1(\text{two-sided})} = .0081 < .01535$ for $\alpha = .05$). Therefore, the study is stopped for efficacy after Stage 1 and the null hypothesis is rejected. The results show a more positive self-rating in the variable *perception of personal sense of life* for men $M = 12.68$, $SD = 2.67$ than for women $M = 11.23$, $SD = 2.36$. This finding allowed for a reduction of the a priori estimated sample size by $N = 88$ participants, due to stopping for efficacy after Stage 1.

Discussion and Conclusions

In the present article, we highlight the great potential of interim testing for psychological research, and we outline how to apply group sequential designs. Additionally, we provide an easily understandable figure of the core idea behind group sequential designs, a workflow chart, the annotated *R* code, and supply a real-world data set from psychological research. Moreover, we demonstrate the application of sample size estimation based on the sample size inflation factor (IF) and apply early stopping for efficacy to the real-world data set for a two groups comparison. This application illustrates the potential savings in costs and organizational effort through group sequential testing, due to the option of early stopping in case of demonstrated efficacy. In this case, the interim analysis would have saved 50% of the sample size, in absolute numbers a considerable reduction of 88 from the initially planned 176. The observed effect size (Cohen's $d = 0.58$) was actually larger than the initially planned $d = 0.50$. This was not predictable from literature or prior experience since results were not available for gender differences regarding the items as used, and nor was such a large difference indicated by more general results. On the other hand, the observed difference of 1.45 points on a 17-point scale is not sufficiently

large that it would be easily identifiable through a small-scale exploration. Therefore, the described application and the benefit obtained by interim testing can be considered fully realistic. In fact, it is even likely that the planning of an actual study would have been based on an even smaller effect size specification, which would have led to an even larger initial sample size planning, and that the savings realized by interim testing would have been greater.

As interim analyses have changed the game in medical and pharmaceutical research, we are convinced they will increasingly enter psychological studies. Apart from reducing overly long study durations and unnecessarily large sample sizes, related to effort, time, and money, they may lead to more ethical psychological studies by - on average - mitigating the effects of assigning participants to inferior psychological interventions. Though such consequences are often less drastic or visible in psychological research than in medical and pharmaceutical studies, many studies nevertheless involve suboptimal interventions, trainings, treatments, educational measures, etc. Furthermore, lower resource demands may enable additional studies which could not be conducted otherwise, or facilitate the acceptance of a study within a given institution, or promote the use of higher quality samples if researchers are less driven to resort to easily accessible yet less appropriate participants.

A key condition of applying interim testing strategies in psychological research is thorough a priori planning and specification before the study commences, which is less common in psychology than in medical and pharmaceutical research. But this demand to strictly follow a study protocol which is formulated and announced before conducting a study may cause researchers engaged in interim testing to become acquainted with more stringent standards. Otherwise, the exploratory hunting or fishing for significance at some unplanned interim stages would cause Type I error inflation because of uncontrolled multiple testing, in addition to the violation of scientific and ethical principles. Since approval by ethics committees or institutional review boards is becoming ever more common, this would be an appropriate occasion to announce group sequential designs as well. The ultimate objective of group sequential designs is not to meet formal requirements, however, but to protect respondents from avoidable burdens and to save researchers' resources.

When will group sequential designs be most effective? If the effect sizes observed are as expected, classical sample size planning will deliver reasonable sample sizes and results in order to achieve a power of, for example, 80%. In contrast, early stopping after an interim analysis will be particularly likely if the effect size is underestimated, as in our demonstration example, or is estimated with caution because of substantial uncertainty. The latter will apply to many cases where population effect parameters are difficult to predict, which is rather the rule than the exception in psychology, at least as soon as one deals with innovative topics or with innovative item material like in our empirical data set. Therefore, group sequential tests could be seen as a compromise between striving

for results and sample sizes with the desired power, and efforts to reduce the use of resources and the encumbrance on participants when the results are more marked than modelled by careful or pessimistic pre-assumptions as demonstrated.

In conclusion, given their innovative conceptualization and ongoing refinements, we expect these powerful methods to markedly invigorate psychological research and further social science fields with currently modest usage of group sequential designs.

Funding: The authors have no funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: We applied the SDC approach for the sequence of authors. We are very grateful to Prof. Dr. Gernot Wassmer and Prof. Dr. Werner Pölz, both experts in the field of biostatistics, for important statistical advice. We are indebted to the former editor-in-chief, Prof. Dr. Peter Lugtig, and two anonymous referees, for their valuable and insightful comments which substantially improved this tutorial for the intended audience of psychological researchers.

Data Availability: The data for this article is available for scientific use (for access options see Supplementary Materials).

Supplementary Materials

For this article the following supplementary materials are available:

- *Data* (Weigl & Ponocny, 2020a): The data is an extract from the MODUL Studies of Living Conditions (Ponocny et al., 2015, 2016) and must not be used for any other purpose (except with the authors' permission) than (1) tracking and reproducing the results in this tutorial, and (2) teaching and demonstrating interim analyses.
- *R Code* (Weigl & Ponocny, 2020b): The applied *R* code is available as an *R* script file with the real-world data. Furthermore, we briefly annotated this code for didactic reasons so it can be easily applied with *R* Studio (RStudio Team, 2019) in psychological research.

Index of Supplementary Materials

Weigl, K., & Ponocny, I. (2020a). *Supplementary materials [data] to: Group sequential designs applied in psychological research.* PsychOpen. https://doi.org/10.23668/psycharchives.2784

Weigl, K., & Ponocny, I. (2020b). *Supplementary materials [code] to: Group sequential designs applied in psychological research.* PsychOpen. https://doi.org/10.23668/psycharchives.2785

References

American Psychological Association. (2010). *Publication manual* (6th ed.). Washington, DC, USA. https://psycnet.apa.org/record/2009-16118-000

American Psychological Association. (2020). *Publication manual* (7th ed.). Washington, DC, USA. https://doi.org/10.1037/0000165-000

Anderson, K. (2016). gsDesign: Group sequential design (*R* package version 3.0-1). Retrieved from https://CRAN.R-project.org/package=gsDesign

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General), 132*, 235-244. https://doi.org/10.2307/2343787

Casper, C., & Perez, A. O. (2018). ldbounds: Lan-DeMets method for group sequential boundaries (Based on FORTRAN program ld98; *R* package version 1.1-1.1). Retrieved from http://CRAN.R-project.org/package=ldbounds

Champely, S. (2018). pwr: Basic functions for power analysis (*R* package version 1.2-2). Retrieved from https://CRAN.R-project.org/package=pwr

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York, NY, USA: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, USA: Lawrence Erlbaum.

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*(2), 276-302. https://doi.org/10.1037/0033-2909.125.2.276

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology, 29*(1), 94-122. https://doi.org/10.1016/j.joep.2007.09.001

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *The British Journal of Radiology, 44*(526), 793-797. https://doi.org/10.1259/0007-1285-44-526-793

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. London, United Kingdom: Chapman and Hall/CRC.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*(7), 701-710. https://doi.org/10.1002/ejsp.2023

McPherson, C. K., & Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society, 134*, 15-25. https://doi.org/10.2307/2343971

Meisenberg, G., & Woodley, M. A. (2015). Gender differences in subjective well- being and their relationships with gender equality. *Journal of Happiness Studies, 16*(6), 1539-1555. https://doi.org/10.1007/s10902-014-9577-5

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*, 549-556. https://doi.org/10.2307/2530245

Pahl, R. (2018). Groupseq: A GUI-based program to compute probabilities regarding group sequential designs (R package version 1.3.5). Retrieved from https://CRAN.R-project.org/package=GroupSeq

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., . . . Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer, 34*(6), 585-612. https://doi.org/10.1038/bjc.1976.220

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*, 191-199. https://doi.org/10.1093/biomet/64.2.191

Ponocny, I., Weismayer, C., Dressler, S., & Stross, B. (2015). *The MODUL study of living conditions* (Technical Report). Vienna, Austria: MODUL University. Retrieved from https://www.modul.ac.at/uploads/files/user_upload/Technical_Report_-_The_MODUL_study_of_living_conditions.pdf

Ponocny, I., Weismayer, C., Dressler, S., & Stross, B. (2016). Are most people happy? About the meaning of life satisfaction ratings. *Journal of Happiness Studies, 17*(6), 2635-2653. https://doi.org/10.1007/s10902-015-9710-0

R Core Team. (2019). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

RStudio Team. (2019). Rstudio: Integrated Development Environment for R. Retrieved from http://www.rstudio.com/

Vandemeulebroecke, M. (2009). adaptTest: Adaptive two-stage tests (*R* package ver- sion 1.0). Retrieved from http://CRAN.R-project.org/package=adaptTest

Wald, A. (1943). *Sequential analysis of statistical data: Theory* (A report submitted by the Statistical Research Group, Columbia University to the Applied Mathematics Panel, National Defense Research Committee).

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*, 117-186. https://doi.org/10.1214/aoms/1177731118

Wald, A. (1947). *Sequential Analysis*. New York, NY, USA: Wiley.

Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics, 43*, 193-199. https://doi.org/10.2307/2531959

Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Cham, Switzerland: Springer.

Wassmer, G., & Pahlke, F. (2019). rpact: Confirmatory adaptive clinical trial design and analysis (R package version 2.0.2). Retrieved from https://CRAN.R-project.org/package=rpact

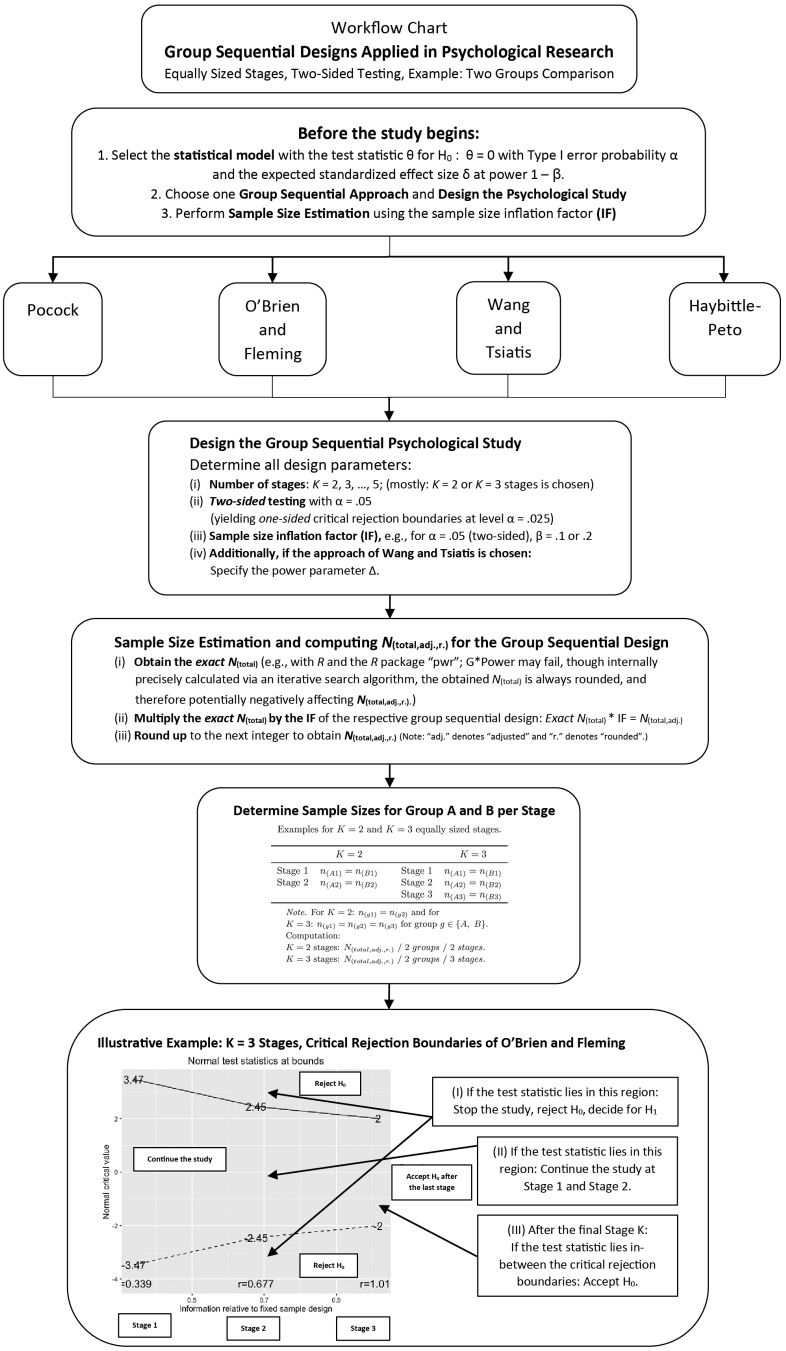
Appendices

Appendix A: Software

There are several different software solutions on the market for performing interim analyses. On one side, there exist freely available open-source solutions such as several *R* packages. For the

application of group sequential designs, first and foremost *gsDesign* by Anderson (2016) with a wide-spanning functionality, but also the *R* packages *GroupSeq* (Pahl, 2018) and ldbounds (Casper & Perez, 2018), both with fewer *R* functions, can be recommended for two-sided testing. For adaptive testing (which is beyond the scope of the present article) and two-sided testing, *adaptTest* by Vandemeulebroecke (2009) as well as the newly available *R* package *rpact* (Wassmer & Pahlke, 2019) can be recommended. Moreover, *rpact* can also be applied for group sequential designs such as *gsDesign*. On the other side, ADDPLAN® and EAST® 6 are currently the two most popular commercially available software solutions for performing interim analyses. First, ADDPLAN® is a fully validated statistics software platform, which allows for in-stream data cleaning and testing of high-value sophisticated clinical designs with a variety of interim analyses. Second, EAST® 6 *The Comprehensive Trial Design Solution* provides a user-friendly interface for the quick construction of trial designs and enables numerous applications of interim analyses. Furthermore, the commercially available statistics software SAS® 9.4 provides the two procedures SEQDESIGN and SEQTEST for group sequential testing. We applied the statistics software *R* (R Core Team, 2019) and created our own and richly annotated *R* code primarily using the *R* package *gsDesign* for the application of group sequential designs and *pwr* for sample size estimation.

Appendix B: Workflow Chart



Methodology
2020, Vol.16(1), 75–91
https://doi.org/10.5964/meth.2811