Healthcare and Medical Devices, Vol. 130, 2024, 88–92
https://doi.org/10.54941/ahfe1004841

Stepped Wedge Design Analyses Under Pandemic's Period Effect: Alternative Approach Using Non-Specific Neck Pain Intervention Data

Matthew Kerry[1], Andrea Aegerter[1,2], Achim Elfering[2], and Markus Melloh[1]

[1]Zürich University of Applied Sciences (ZHAW), Institute of Public Health, Winterthur, Switzerland
[2]University of Bern, Faculty of Human Sciences, Institute of Psychology, Bern, Switzerland

ABSTRACT

Use of stepped wedge design (SWD) trials have increased exponentially over the past decade (Hooper & Eldridge, 2020). Concomitantly, due to the increasing prevalence of neck pain in the workforce, interventions are necessary and have to be evaluated. Stepped-wedge designs are adaptive, so they can and should adjust to externalities. For example, the COVID-19 pandemic introduces a period effect that could perturb the design. To understand SWD's potential vulnerability to the secular trend of a pandemic (or other period effect in future conductance), we compare a classical SWD analysis to an item-response theory (IRT) approach that only utilizes before-after segments of the data collected. A national "closed-cohort" SWD was examined (n = 120 with 4 measurements, 480 measurements). The outcome measure is the European Quality of Life instrument (EQoL-5D-5L). Our "gold standard" SWD analysis yielded a significant effect of Cohen's d = .29, SE = .009. In comparison, our newly proposed IRT model yielded a similarly significant effect of Cohen's d = .31, but with power loss as indicated by a higher SE = .19. Finally, our "crude", classical paired t-test yielded a greater effect size of Cohen's d = .36, SE = .007. For IRT, the average-relative parameter bias was 7% and considered below the ignorable 10–15% threshold (Rodriguez, Reis, & Haviland, 2016). For paired t-test, the average-relative parameter bias was an unacceptable 24%. An IRT-alternative to SWD designs with before-after data yields unbiased effects, but loses power. The IRT approach may be replicated in another SWD design outside of the period of the COVID-19 pandemic to understand it's potential under "normal" study conditions.

Keywords: Stepped wedge design (SWD), Item response theory (IRT), Bifactor, Neck pain

INTRODUCTION

Use of stepped wedge design (SWD) trials have increased exponentially over the past decade (Hooper & Eldridge, 2020). An SWD is defined as a baseline collection of observations where no clusters are exposed to an intervention, followed by random and sequential crossover of clusters from control to intervention until all clusters have been exposed. Concomitantly, due to the

© 2024. Published by AHFE Open Access. All rights reserved.

increasing prevalence of neck pain in the workforce, interventions are necessary and have to be evaluated. Stepped-wedge designs are adaptive, so they can and should adjust to externalities. For example, the COVID-19 pandemic introduces a period effect that could perturb the design. To understand SWD's potential vulnerability to the secular trend of a pandemic (or other period effect in future conductance), we compare a classical SWD analysis to an item-response theory (IRT) approach that only utilizes before-after segments of the data collected. Understanding the relative-advantages of classical SWD analyses may hold import for recommendations regarding continuous data collections (i.e., response burden) under "known" externalities in future research.

Several complex analytic approaches to SWDs have previously been reviewed, but a complementary approach may be a more advanced measurement theory (Li & Wang, 2022). For example, Li and Wang note that SWD analyses can be classified as either conditional (cluster-specific) or marginal (population-averaged) regression models (2022). Both of these models, however, are rooted in classical test theory. In complement, a model-based measurement approach, IRT may be adopted for analysing SWD data (Embretson, 1999). IRT assumptions enable accounting for secular trends in the SWD design, so long as within-cluster equality constraints are appropriately applied. For example, regardless of regression-based model, Li & Wang note that high-quality SWDs should report ICCs, as well as modeling assumptions for secular trends and random-effects (2022). IRT's distribution-free parameter estimates enables meaningful assessment of change regardless of baseline values (Embretson & Poggio, 2012).

Research Questions

Should all participants continue to be measured after intervention (burdensome measurement)? Can we impose a different analysis (two-time point, uncontrolled before-after) on the SWD to obtain period/event-robust effect estimates? Specifically, a before-after extension of IRT's bifactor model for assessing change is applied to the current dataset (Cai, 2010). The logic-flow illustrating our planned "horizontal-block comparisons" and the IRT model used for estimation is displayed in Figure 1 below.

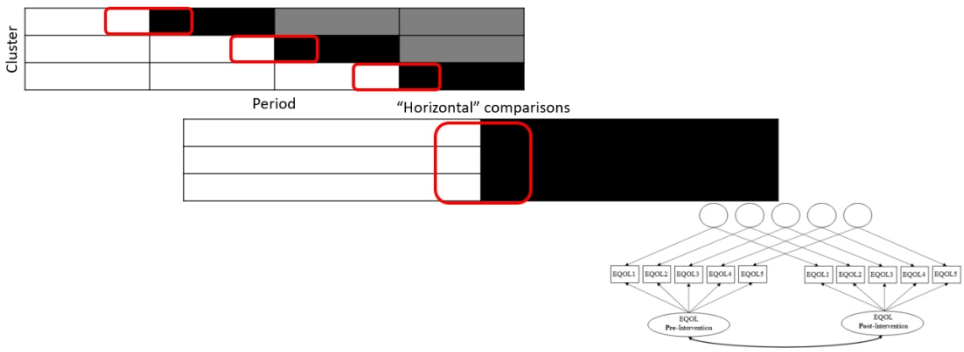

Figure 1: Illustration of SWD horizontal comparisons to before-after redesign approach using IRT before-after bifactor model.

SAMPLE & METHODS

The dataset comes from a nationally funded project entitled NEXPRO (Aegerter et al., 2022). This may be classified as a "closed-cohort" variant of SWDs (Li & Wang, 2022). In a closed-cohort design, a suitable population is identified at the beginning of the study with repeated follow-up measurement after cross-over, but no adjustments are made in terms of participant attritions and consequent additions. Approximately half of our sample (n = 120 with 4 measurements, 480 measurements) comprises employees working in the health-system education context with common neck problems.

The outcome measure is the European Quality of Life instrument (EQoL-5D-5L).

Three analytic estimates are reported: 1) We compare classic SWD analyses (n = 296) with 2) IRT estimates from before-after segments (n = 194) of the dataset, and 3) A classical paired t-test. Effect size estimates and standard errors are reported to allow interpretations of estimate bias and power, respectively. The classic SWD analysis served as "gold standard" comparator for accurate estimates that takes full power-advantage of all measurements. Specifically, a generalized linear mixed-effects model with robust estimates was used, entering random-intercepts for repeated-measurement and fixed-effects for cluster, time, and intervention to estimate changes in EQoL.

Specifically, we compare effect sizes and standard errors across analytic approaches. By comparing effect size estimates across analyses, we should gain understanding as-to potential bias from including additional measures during pandemic perturbation. Complementary, comparing standard errors should give insight into potential power loss from "ignoring" additional measurements. Applying parameter inputs from Hooper and Eldridge's (2021) illustration of SWD variants, for example, we obtained an anticipated power estimate of $1-\beta = .70$ for our planned "horizontal-block comparisons".

RESULTS

Our "gold standard" SWD analysis yielded a significant effect of Cohen's d = .29, SE = .009. In comparison, our newly proposed IRT model yielded a similarly significant effect of Cohen's d = .31, but with power loss as indicated by a higher SE = .19. Finally, our "crude" classical paired t-test yielded a greater effect size of Cohen's d = .36, but with "gold standard"-similar SE = .007 (perhaps due to ignoring pandemic-secular trend perturbation. These results are summarized in Table 1 below.

Table 1. Summary across models of effect size (potential bias) and standard error (proxy power) estimates.

Model-Analytic Approach	Effect Size	Standard Error
SWD "Gold Standard"	.29	.009
IRT "Before-After" Bifactor	.31	.19
CTT "Paired T-test"	.36	.007

For IRT vs. "gold standard", the average-relative parameter bias (d = .29 vs. d=.31) was 7% and considered below the ignorable 10–15% threshold (Rodriguez, Reis, & Haviland, 2016). For paired t-test, the average-relative parameter bias was an unacceptable 24% (d = .29 vs. d=.36).

DISCUSSION

IRT reduces bias but loses power due to measurement specification. Thus, if researchers are interested in obtaining accurate (unbiased) intervention-effect estimates and wish to reduce response burden by ignoring follow-up (or pre-advanced) measurements (perhaps due to pandemic externalities, lengthy measurements, or vulnerable populations), then the IRT approach would be appropriate, if sacrificing power and potential statistical significance.

We provided an analytic alternative to SWDs in cases of unforeseeable externalities, such as the period effects of a global pandemic. Such "secular trends" are consistently called on for accountability in SWD analyses (Li & Wang, 2022). In the current study, imposition of "equality constraints" in an IRT-based modeling approach is equivalent to the "period-additivity" assumption of common linear-mixed effects models, although IRT is inherently non-linear (Kennedy-Shaffer, Gruttola, & Lipsitch, 2019). Indeed, our robust approach may most closely related to the non-parametric, between-period ("horizontal") comparisons elaborated by Thompson and colleagues (2018).

In addition to our period-robust analytic approach, several design alternatives may also mitigate bias from "secular trends" in SWDs. For example, another practical consideration to minimize potential disruption from period effects may be to simply contract the timescale of the rollout. Obviously, there may be meaningful limitations in terms of saturation or detection of effects. Still, from a statistical power standpoint, Hooper & Eldridge (2021) illustrated how contracting the timescale or other "clustering tweaks" may still succeed in attaining a nominally acceptable power rate of 80%. The shorter the trial overall, the shorter the potential contamination from a pandemic or other externality.

LIMITATIONS

There are several substantive limitations to the current study that are noteworthy. We restrict our analysis to a simple, 5-item EURO-QoL measure, although much more complex measures with several more items will obviously increase modeling complexity. Also, depending on the motivation of the SWD, the currently proposed "before-after" restructuring may or may not be advisable. For example, if the primary motivation for the SWD is to ensure every participant / cluster receives treatment, then perhaps the current "before-after" restructuration is inadvisable, as effect "duration" may be of more interest than simple "effect presence". On the other hand, if the SWD is motivated primarily for logistical purposes, or is being administered in a program that is already prepared for full-rollout, then perhaps the "before-after"

restructuration and effect estimate is desirable. Finally, in cases where contamination between control and intervention is more likely (e.g., proximal clusters), the "before-after" approach may be less advisable.

CONCLUSION

An IRT-alternative to SWD designs with before-after data yields unbiased effects, but loses power. The IRT approach may be replicated in another SWD design outside of the period of the COVID-19 pandemic to understand it's potential under "normal" study conditions.

REFERENCES

Aegerter, A. M., Deforth, M., Volken, T., Johnston, V., Luomajoki, H., Dressel, H., Dratva, J., Ernst, M. J., Distler, O., Brunner, B. and Sjøgaard, G., 2023. A Multi-component Intervention (NEXpro) Reduces Neck Pain-Related Work Productivity Loss: A Randomized Controlled Trial Among Swiss Office Workers. *Journal of occupational rehabilitation*, *33*(2), pp. 288–300.

Cai, L., 2010. A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), pp. 581–612.

Cook, K. F., O'Malley, K. J. and Roddey, T. S., 2005. Dynamic assessment of health outcomes: time to let the CAT out of the bag?. *Health services research*, *40*(5p2), pp. 1694–1711.

Embretson, S. E. and Hershberger, S. L. eds., 1999. *The new rules of measurement: What every psychologist and educator should know*. Psychology Press.

Embretson, S. E. and Poggio, J., 2012. The impact of scaling and measurement methods on individual differences in growth.

Hooper, R. and Eldridge, S. M., 2020. Cutting edge or blunt instrument: how to decide if a stepped wedge design is right for you. *BMJ quality & safety*.

Kennedy-Shaffer, L., De Gruttola, V. and Lipsitch, M., 2020. Novel methods for the analysis of stepped wedge cluster randomized trials. *Statistics in Medicine*, *39*(7), pp. 815–844.

Li, F. and Wang, R., 2022. Stepped wedge cluster randomized trials: a methodological overview. *World Neurosurgery*, *161*, pp. 323–330.

Li, F., Kasza, J., Turner, E. L., Rathouz, P. J., Forbes, A. B. and Preisser, J. S., 2023. Generalizing the information content for stepped wedge designs: A marginal modeling approach. *Scandinavian Journal of Statistics*, *50*(3), pp. 1048–1067.

Maleyeff, L., 2023. *Treatment effect heterogeneity in cluster randomized trials* (Doctoral dissertation, Harvard University).

Thompson, J. A., Davey, C., Fielding, K., Hargreaves, J. R. and Hayes, R. J., 2018. Robust analysis of stepped wedge trials using cluster-level summaries within periods. *Statistics in medicine*, *37*(16), pp. 2487–2500.