A primer on validity and design terminology in comparative designs

Paul White,

Applied Statistics Group,

Faculty of Environment and Technology,

Univeristy of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul.white@uwe.ac.uk

Paul Redford,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

paul2.redford@uwe.ac.uk

James Macdonald,

Department of Health and Social Sciences

Faculty of Health and Applied Sciences,

University of the West of England, Brsitol,

Bristol BS16 1QY, UK

james.macdonald@uwe.ac.uk

***Abstract*— An overview of terminology (independent and dependent designs, dependent and independent variables) and validity (internal, external, and measurement) is given using six different scenarios. For simplicity of exposition, all examples have factors at two levels.**

***Keywords*— design terminology, validity**

I. Introduction

Initially, some of the terminology used in the quantitative research sciences can appear to be a little confusing. For instance, a researcher might talk about an independent variable or an independent design, or a dependent variable and a dependent design, and then proceed to talk about different types of validity (e.g. internal, external, or measurement validity).

It is quite important to have a good grasp of the terminology used; the terminology used helps describe a design and the resulting statistical analysis should mimic and capture the design.

In the following, a number of plausible scenarios are outlined briefly. In each case we will consider and discuss

(a) whether we have independent or dependent samples
(b) the independent variable
(c) the dependent variable

In addition, where appropriate, for each research situation we will consider threats to (a) *internal validity* and (b) *external validity* and (c) *measurement validity*

Before proceeding on to the example scenarios, it may be instructive to note that dependent designs are associated with “paired data”, “matched data”, “repeated measures”, “within-subjects”, or “blocked designs”. Similarly, independent designs are often referred to as between-subjects designs.

Also, an independent variable (IV) is a “factor”, or “explanatory variable” and the dependent variable (DV) is the “outcome” or “response” or what is being “measured”.

Likewise, for the two-sample case, commonly used techniques for comparing two independent samples on a *scale* dependent variable, include the independent samples t-test, Welch’s t-test, or the Mann Whitney Wilcoxon (aka the Mann Whitney test, or aka the Wilcoxon Rank Sum test). For the two-sample case, commonly used techniques for comparing two dependent samples with a scale dependent variable, include the paired samples t-test and the Wilcoxon Signed Rank test (aka the Wilcoxon test). Accordingly, knowing whether a design comprises independent samples or dependent samples will go a long way to helping to select the most appropriate analytical technique.

II. Worked examples with discussion

A. *Example 1*

To test the effect of background music on productivity, a group of workers is observed. For one month, they had no music. For another month, they had background music.

B. *Discussion (Example 1)*

The *dependent variable (DV)* in this example is productivity. We would need to know a bit more about productivity to ascertain how it is precisely quantified and hence consider measurement validity.

The factor of interest, i.e., the *independent variable (IV)*, is Background Music. The IV is a factor with two levels:- (a) background music is present or (b) background music is absent. Note this is one factor (one IV) with two levels.

Do we have an independent design or a dependent design? In this case we have repeated the same measure (productivity) on each participant under two different states of nature (music present, music absent) and as such we have a repeated measures design or a *dependent design*.

Suppose we analyse the resulting data and discover that, in the sample, productivity in the Music Condition is significantly higher than the No Music Condition. Could we really say, that Background music *caused* the increase in productivity? It is doubtful. There may be extraneous factors at play (e.g. the Hawthorne effect [1]). What if the no music condition was in July and the music condition was in August? If this was the case then Month (July, August) is completely *confounded* with Music (absent, present) and the conclusion could be in terms of month rather than music. Based on what is known about this example it would be hard to argue that background music was the cause and hence the internal validity is in doubt. Without good internal validity it is impossible to generalise to a wider population (i.e. the external validity must be in doubt too).

C. Example 2

To test the effect of background music on productivity, a group of workers (n = 20) are observed without any background music and a second different group (n = 20) are observed with background music playing.

D. Discussion (Example 2)

In this design we have two separate or independent groups. The design is therefore an independent design aka a between-subjects design.

In this design the outcome (dependent variable) is productivity. Of course, we would need more information on what is meant by productivity and how it might be quantified. Measuring productivity could be difficult in, say office workers, and hence we would need to know more to comment on *measurement validity* (i.e. are we really measuring productivity).

Suppose in a statistical analysis of the resulting data we find a statistically significant increase in productivity in the background music condition. Could we really say, that Background music *caused* the increase in productivity? It is doubtful. For instance, those in the background music condition might not be naïve to the purpose of the study, there may be the Hawthorne effect [1] or similar. This casts doubt on the internal validity.

Further, suppose that we did not have concerns about the internal validity and we wonder whether our conclusions from the sample could be generalised (external validity). In this case we only have n = 20 per condition and it would be nigh on impossible to argue that n = 20 could give representative data of a much wider population and accordingly the external validity is in question.

Example 3

A new weight reducing diet was tried on n = 60 women. The weight of each woman was measured before the diet and again after being on the diet for ten weeks.

E. Discussion (Example 3)

In this example each woman is measured before and after the diet. We have therefore repeated “weight” on each participant under two different states of nature (before diet, after diet) in a longitudinal design. This, therefore, is a repeated measures design (a dependent design) with weight as the dependent variable. The factor of interest is DIET which has two levels (before, after).

Measurement validity is not in doubt in this case. Measuring weight (mass) is not difficult.

Suppose, in a statistical analysis we discover that mean weight after diet is significantly lower than before the diet. Would we be happy to conclude, for the women who took part, that the average decrease in weight loss was *because* of the diet? The answer to this is “probably not”! The women taking part may have altered their behaviour e.g. increase in exercise or similar. It could be suggested that a better design might have incorporated a control group.

Of course, n = 60 is large in some respects but not sufficiently large to be representative of a much larger population, so even if we did have good internal validity, there would be doubts over the external validity.

F. Example 4

To compare the average weight gain of pigs fed on two different rations, twenty pairs of pigs were used. The pigs in each pair were littermates. One pig in each pair was given ration A, the other ration B.

G. Discussion (Example 4)

In this case the factor of interest, or independent variable, is Ration. Ration is a two-level variable (Ration A, Ration B). That is the easy bit!

Pigs in a litter are (usually) very similar to one another but can be quite distinctive from pigs in a different litter. It could be argued that a pair of pigs from a litter are essentially identical in all key aspects (almost as if a pig had been cloned). We therefore have a *matched pairs* design (but not a repeated measures design).

For each pair of pigs we might want to see if the per-pig change in weight differed between the pig on Ration A and the pig on Ration B. This point of view aligns with viewing the design as a dependent design. An alternative point of view is to acknowledge that putting one pig from each pair into Ration A and the other into Ration B is a good idea, but to then argue, that because they are actually different pigs, to analyse them as if they were two independent samples. So, which is the correct way of viewing the situation?

If the data (dependent variable) in a matched pairs design is positively correlated then there is a statistical advantage of analysing the data as a dependent design using, say, the paired samples t-test or the nonparametric Wilcoxon test. This statistical advantage is derived from the power of the test (see [2]). As a broad rule, as the degree of correlation in matched pairs design increases then the power of the test increases if a paired analysis is undertaken. It is for this reason that computer packages such as SPSS will conduct a correlation analysis prior to a paired samples t-test.

If the data in a matched pairs design is not positively correlated then there is a statistical advantage in analysing as an independent design.

Suppose the result of a statistical test indicated a significant increase in weight gain for Ration A compared to Ration B. For the sample of pigs, could we attribute this effect to the rations? The use of randomisation and a hard outcome measure (i.e. not a self-report) of weight, pigs being naïve to the motivation of the study, all add to the internal validity. However, the sample size is small and therefore difficult to generalise to all breeds of pig. This prima facie evidence of Ration A being superior to Ration B on weight gain is subject to future replication.

H. Example 5

To investigate potential institutional gender bias, a sample of university lecturers was taken. The purpose behind the sampling was to compare salaries of male and female lecturers.

I. Discussion (Example 5)

The factor of interest is gender, which, in this case, is a two-level (male, female) variable. There is no logical mechanism to match or pair any one male lecturer with any one female lecturer. As such, we have an independent design aka a between-subjects design.

The dependent variable is salary.

Salary is relatively easy to measure.

Suppose in the sample we find that average salary for males significantly exceeded average salary for females; would this then translate into "institutional bias" or event "institutional bias on salary"? There are far too many variables unaccounted for to draw any such conclusion (e.g. length of service, number of career breaks, and so on). We might need a better design to answer the question as posed.

J. Example 6

To investigate a claim of potential institutional gender bias, a HR department matched male and female lecturers on their start date, qualifications on entry, initial position on pay spine, and age. The purpose behind the sampling was to compare salaries of male and female lecturers.

K. Discussion (Example 6)

The factor of interest is gender, which, in this case, is a two-level (male, female) variable. The design is a matched pairs design with pairs matched on their start date, qualifications on entry, initial position on pay spine, and age.

The dependent variable is salary.

Salary is relatively easy to measure.

The question is whether to analyse the data as a dependent design or an independent design.

One school of thought is, similar to the pig example (Example 3), to consider that there will be a positively correlation on salary for the matched pairs and to therefore exploit this by conducting a paired analysis.

A second school of thought is to acknowledge that design ensures the male and female samples to be balanced on the matching criteria but to argue that there are possibly other important factors not used in the matching process and to the proceed to analyse the data as if they were two independent samples.

Note that different statistical conclusions could be drawn whether the data is analysed as dependent samples or independent samples.

III. Summary

This brief note has largely considered some terminology associated with two-level comparative designs. In general, the terminology used helps describe a design and the resulting statistical analysis should mimic and capture the design. Example scenarios have been used to draw out and consolidate what is meant by an independent variable, a dependent variable, and by independent and dependent designs.

Note that some two-level designs include both independent and dependent samples. These latter situations are referred to as partially overlapping samples and parametric [3, 4] and non-parametric [5] methods for analysing these data are available.

It must be said that statistical terminology can at times be quite poor and deceptive. For instance, statistical significance is quite different from something being "significant" (as in important or substantive). In ANOVA we may talk about a "main effect" but this does not mean it is the most important effect, and ANOVA (analysis of variance) might seem strange when looking for differences between means!, and regression equations might not contain any regression! The list goes on. Please, mind your statistical language

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Learning and Teaching Initiative in the Faculty of Health and Applied Sciences, University of the West of England, Bristol in supporting the wider Qualitative and Quantitative Methods teaching programme.

SERIES BIBLIOGRAPHY

White P, Redford PC, and Macdonald J (2019) A primer on validity and design terminology in comparative designs, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the chi-squared test of association (2 by 2), *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the independent samples t-test and the Welch test, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –6

White P, Redford PC, and Macdonald J (2019) An example motivated discourse of the paired samples t-test, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –5

White P, Redford PC, and Macdonald J (2019) A primer on statistical hypotheses and statistical errors, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –4

White P, Redford PC, and Macdonald J (2019) A reverse look at p-values, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –5.

White P, Redford PC, and Macdonald J (2019) That assumption of normality, *Qualitative and Quantitative Research Methods Project*, University of the West of England, 1 –11

White P, Redford PC, and Macdonald J (2019) Cohen's *d* for two independent samples, *Qualitative and Quantitative Research Methods Project,* University of the West of England, 1 –4

REFERENCES

[1] McCambridge J, Witton J and Elbourne DR (2014) Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects, *Journal of Clinical Epidemiology*, Vol 67, 267 - 277

[2] Van Voorhis CRW and Morgan BL (2007), Understanding Power and Rules of Thumb for Determining Sample Sizes, *Tutorials in Quantitative Methods for Psychology*, Vol. 3, No 2, p. 43-50.

[3] Derrick, B., Toher, D. and White, P. (2017) How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods in Psychology,* 13 (2). pp. 120-126.

[4] Derrick, B., Russ, B., Toher, D. and White, P. (2017) Test statistics for the comparison of means for two samples which include both paired observations and independent observations. *Journal of Modern Applied Statistical Methods,* 16 (1). pp. 137-157.

[5] Derrick, B., White, P. and Toher, D. (2018) Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations. *Jounal of Modern Applied Statistical Methods.*