Fact sheet: Quantitative methods

This fact sheet will cover some of the core factors within quantitative methodology, including sample sizes, measuring outcomes quantitatively and surveys. For information on different statistical analyses with quantitative data, please refer to the fact sheet on an overview of statistical analyses.

Top TSIP evaluation tips

- The larger the number of participants in your programme evaluation, the more likely it is that you will actually detect the true impact of your programme, and the stronger your claim will be that your impact would also occur in the wider population
- In some circumstances, it may be sensible to make your treatment group larger than your comparison group (or vice versa), but you need to be careful that this does not weaken your evaluation design
- Outcome measures need to be reliable and valid, and validated measures are usually tested thoroughly for these two aspects

Sample Sizes[1]

The statistical analysis of quantitative data (with or without a comparison group) aims to show how confident we can be that any difference in change in outcome indicators in the treatment group (compared to a comparison group or compared to itself) is due to the programme itself, rather than i) chance or ii) other factors. Whilst the latter is governed by the make-up of the comparison and treatment groups, the former is governed by their size – the sample size.

In fact, the level of confidence about your results is governed by the interaction between three factors:

- Sample size, i.e. how many participants are there in the evaluation?
- Standard deviation of your outcome indicator, i.e. how much do the outcome indicators typically vary by in each participant?
- Effect size[2], i.e. what is the *standardised* difference in outcome indicator(s) at time 1 versus time 2 (or at time 2 in the treatment group versus time 2 in the comparison group)? Standardising the difference means taking into account the units and context of the measure (e.g. a change of 4 inches is different from a change in 4 centimetres). Once standardised, you can compare effect sizes across measures and even studies. As a rough guideline, in social intervention research, an effect size of 0.2 is "small", of

[1] Learn more about sample sizes on pp. 127-138 in Togerson 2008, and on pp. 142-156 Rea 1992

[2] Learn more about effect sizes on pp. 174-178 in Gorard 2013

0.5 is “moderate” and of 0.8 and above is “large”[3], although these are somewhat arbitrary). The effect size is thus a useful and comparable measure for the magnitude of a programme’s impact.

Example: If you find an effect size of 0.3 for a tutoring programme designed to improve school grades, this means that on average, the children only achieved slightly better grades (e.g. from a low C to a high C), while another programme that achieved an effect size of 0.8 caused children to improve by more than one grade (e.g. from a D to a B).

The bigger the effect size, the smaller the sample size can be, and vice versa. It is possible to estimate what sample size you need, or what effect size you need to achieve, using an online calculator like this one.

Potential problems

Sample size too small: Having too small a sample size is perhaps the most common flaw in evaluations of social programmes. Unless the effect size is very large, sample sizes of less than 100 are unlikely to show a statistically significant impact. However, with large sample sizes (e.g. in excess of 1000) you are much more able to detect even a very small impact (i.e. it hits statistical significance). Evaluations should only be carried out when you have a sufficiently large sample size to show statistical significance for the expected effect size (otherwise it is a waste of resources since even if there was an effect, your analysis would not be able to detect it).

Attrition: It is important to take attrition into account when choosing your sample size – in most evaluations the number of individuals receiving the intervention will fall over time, due to dropouts for various reasons. To cater for this, you should aim to have a sample size that is higher than the minimum that you think you will need to find statistical significance.

Unbalanced treatment and comparison group ratio[4]

The size of the treatment and comparison groups do not have to operate on a 1:1 ratio – it is possible to have a larger or smaller treatment group compared to the comparison group, albeit with an impact on the effect size that can be detected. The following table includes some examples of how the ratio in sample size changes with sample size and what this means for what effect size it is possible to detect:

[3] Please note that these numbers are dependent on the type of effect size. These numbers apply for Cohen’s *d,* which is perhaps the most commonly used type of effect size. Cohen’s *d* involves dividing the difference in outcome indicators by the pooled standard deviation of those indicators – in other words, it frames the impact of the intervention in relation to how much the outcome indicators typically vary by in each participant.

[4] Learn more about unequal randomisation on pp. 108-113 in Togerson 2008

	N = 150		N = 210		N = 270		N = 330	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
1:1	75	75	105	105	135	135	165	165
Effect size	0.46		0.39		0.34		0.31	
2:1	100	50	140	70	180	90	220	110
Effect size	0.49		0.41		0.36		0.33	
3:1	113	38	158	53	203	68	248	83
Effect size	0.53		0.45		0.40		0.36	

Interpretation of the table: If you have a sample size of 270 with 135 in both the treatment and the control group, the smallest effect size that you can statistically detect is 0.34 (i.e. you can confidently detect any effect sizes bigger than 0.34). However, if you have the same sample size but 203 in your treatment and 68 in your comparison group, you can only detect effect sizes of 0.4 and larger. So, if your programme can bring about an effect size of 0.37, you will find this effect if you use a 1:1 ratio but won't find it if you use a 3:1 ratio.

When deciding what ratio is most appropriate for your trail, you should take the following factors into consideration:

- Ethics[5]: If it is genuinely unclear (from an evidence-based perspective) whether the programme is effective or not (also known as the principle of equipoise), it is ethically justifiable to "deny" half of your sample the intervention. In fact, one could argue that it is unethical to provide more than half of your sample with your intervention since you cannot be certain that you are actually benefiting them (and not harming them). On the other hand, interventions that are virtually proven to be effective (e.g. the Incredible Years parenting intervention for child disruptive behaviours) can justify making the treatment group twice as large as the comparison group (and you may want to give the comparison group the intervention after the conclusion of the trial due to the extensive evidence showing its effectiveness).
- Cost: The larger the difference in sample size between treatment and comparison groups, the larger your overall sample size needs to be. The larger the sample size, the costlier your trial will be (since data collection typically is the most expensive part of an evaluation). However, changing the ratio may save money in some scenarios. For example, if the treatment is extremely expensive, you can decrease the treatment group size and increase the comparison group size. Or if considerable resources have to be invested to set up the delivery of the treatment, you may want to increase your treatment group size to use the treatment to full capacity and reduce the comparison group size.
- What is commonly done: Generally, the ratio does not exceed 3:1 and is typically 2:1 or 3:2 (if it is not 1:1).

[5] Learn more about ethical and data protection considerations on pp. 79-80 in the Magenta book 2011

Measuring outcomes

Outcomes will generally be measured quantitatively in three different ways:

- By survey (e.g. Strengths & Difficulties Questionnaire)
- By existing administrative data (e.g. re-offending rates, GCSE grades)
- By direct measurement (e.g. testing blood pressure)
- By observation (e.g. teachers rating a pupil's behaviour)

Here are some general rules / guidelines when it comes to measuring outcomes[6]:

Reliability: Outcome measures should be reliable, which means they should return the same value when participants are re-measured in the same circumstances (test-retest reliability). Standardised questionnaires have generally undergone thorough testing for their reliability, which is why they are generally preferable to bespoke questionnaires that may have been poorly designed and worded. However, if there are no appropriate questionnaires, it is possible to test a newly created questionnaire for test-retest reliability. Ideally the newly created questionnaire should also have at least some level of input from an evaluator with experience of designing questionnaires.

Validity: Outcome measures should also have high levels of validity, which means that they should be measuring what they say they are measuring. Standardised questionnaires are often tested for validity and thus should generally be used where possible. However, if it is inappropriate for your particular target population or if you are measuring only one aspect of a concept, its validity can be drastically reduced when you use it in your trial and the questionnaire may need to be adapted to improve validity (though this will have implications for test-retest reliability).

Levels of measurement: In order to measure and analyse the data for your outcome indicator accurately, you need to identify the variable type of your data:

- **Nominal (categorical data):** Separate categories that have no order or scale to them (e.g. ice cream flavours)
- **Ordinal (categorical data):** Finite rankings of some sort (e.g. GCSE grades)
- **Interval (continuous data):** Variables that can be measured to any level of precision (e.g. latitude) but without a meaningful zero point
- **Ratio (continuous data):** Variables that can be measured to any level of precision (e.g. age) with a meaningful zero point

[6] Learn more about reliability and validity on pp. 10-13 in Boslaugh 2012

Surveys[7]

Surveys can generally be administered:

- In person
- On the phone
- By mail
- Online

Guidelines for survey questions:

- **Closed:** Should generally be closed so that the responses can be quantitatively coded
- **Clear:** Should avoid colloquial language, jargon or ambiguity
- **Not double-barrelled:** For example, should not relate to the respondents' 'skills and knowledge' – both skills and knowledge should have their own question
- **Not manipulative:** Should not be preceded with information that might sway the response in some way (e.g. by insinuating a moral or ethical judgement)
- **Not emphasised:** There should not be some words that are bolded or italicised
- **No emotional words or phrases:** For example, a word like 'subversives' has a negative connotation that might influence the response

Books referenced in this fact sheet

Boslaugh, S. (2012), *Statistics in a nutshell*, O'Reilly Sebastopol

Gorard, S. (2013), *Research design: Creating Robust Approaches for the Social Science*, SAGE Publications Ltd. London

HM Treasury (2011), *The Magenta Book – Guidance for Evaluation*, HM Treasury

Rea, L. M., Parker, R. A. (1992), *Designing and conducting survey research: a comprehensive guide*, Jossey-Bass, San Francisco

Togerson, D. J., Togerson, C. J. (2008), *Designing Randomised Trials in Health, Education and the Social Sciences*, Palgrave Macmillan Basingstoke

Glossary

Attrition—when participants in either treatment or comparison groups leave the evaluation (e.g. they disappear or refuse to continue).

Categorical data—involving variables made up of categories of objects/entities (e.g. gender).

Continuous data—involving variables that can be measured to any level of precision (e.g. temperature).

Effect size—the difference between two groups described in standard deviation units (i.e., difference divided by the standard deviation).

[7] Learn more about survey design on pp. 8-72 in Rea 1992

Equipoise—a state of existing knowledge about an intervention where there is there is genuine uncertainty over whether the intervention will be beneficial.

Impact—the long-term, cumulative effect of programmes/interventions over time on what they ultimately aim to change.

Indicator—a quantitative or qualitative variable that provides a valid and reliable way to measure achievement, assess performance, or reflect changes connected to an intervention.

Outcome—short-term and medium-term effect of an intervention's outputs, such as change in knowledge, attitudes, beliefs, behaviours.

Standard deviation—an estimate of the average variability (spread) of a set of data measured in the same units of measurement as the original data (square root of the variance).

Variance—an estimate of average variability (spread) of a set of data.