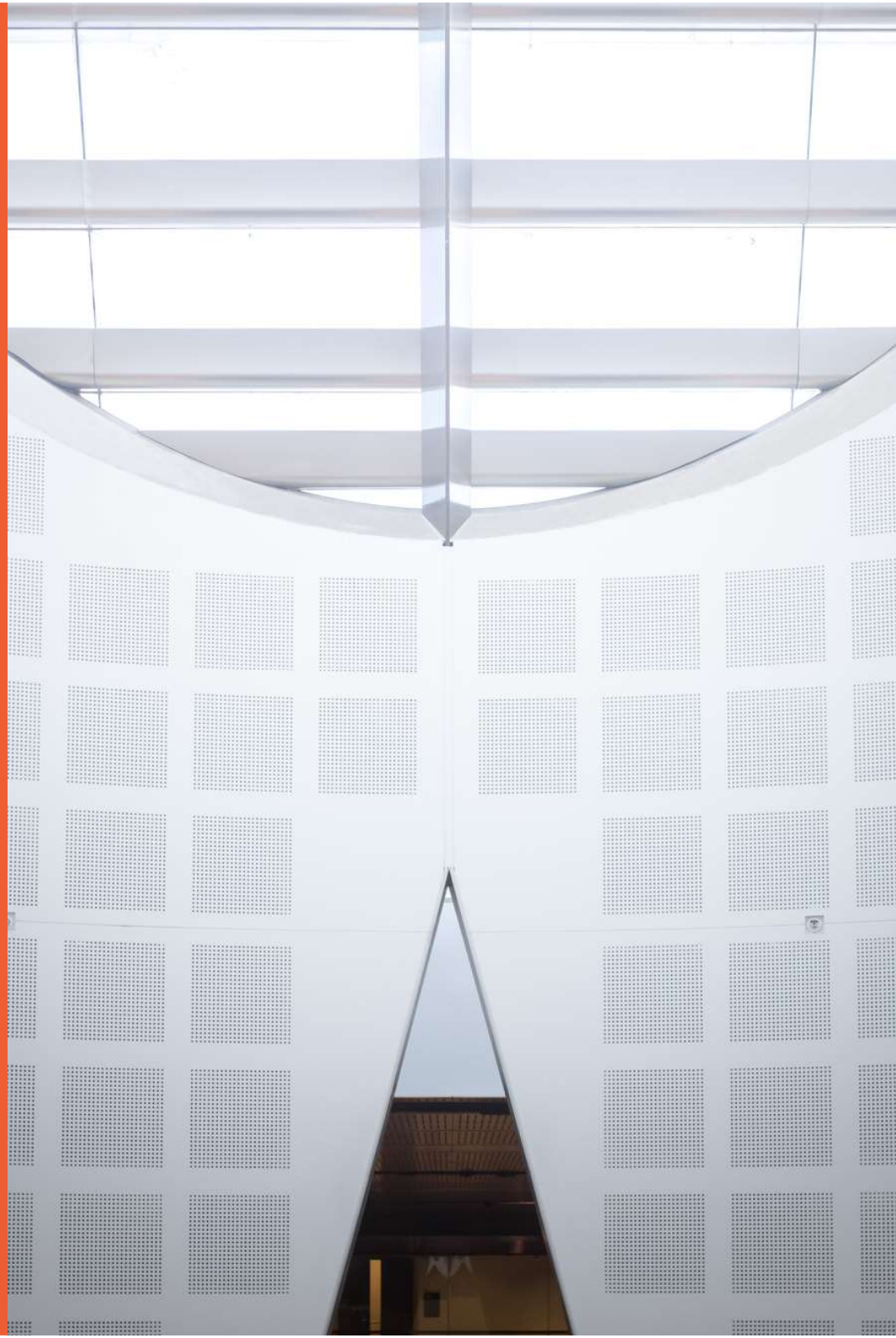
Surveys 2: Advanced Topics

Presented by
Chris Howden
Sydney Informatics Hub
Core Research Facilities
The University of Sydney

THE UNIVERSITY OF SYDNEY



Acknowledging SIH

All University of Sydney resources are available to Sydney researchers **free of charge.** The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

"The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

For further information about acknowledging the Sydney Informatics Hub, please contact us at **sih.info@sydney.edu.au**.

We value your feedback

- We aim to help HDR students and researchers in a wide range of fields across different faculties
- We want to hear about **you** and whether this workshop has helped you in your research.

- Later in this workshop there will be a link to a survey
- It only takes a few minutes to complete (*really!*)
- Completing this survey will help us create workshops that best meet the needs of researchers like you

During the workshop

- Ask short questions or clarifications during the workshop. There will be breaks during the workshop for longer questions.
- Slides with this blackboard icon are mainly for your reference, and the material will not be discussed during the workshop.

Challenge Question

- A wild boar is coming towards you at 200mph. Do you:?
  - A. Ask it directions
  - B. Wave a red flag
  - C. Wave a white flag
  - D. Begin preparing a trap

After the workshop

These slides should be used after the workshop as **Workflows** and reference material.

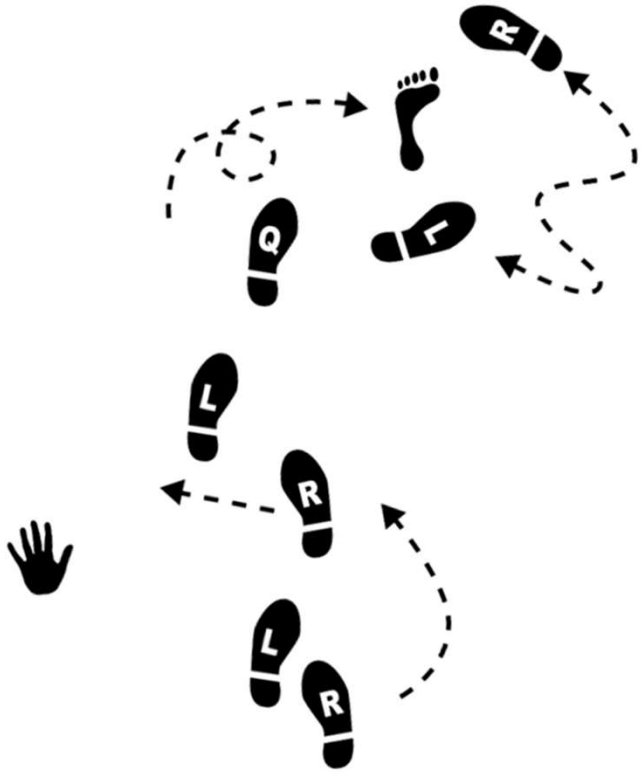
- Todays workshop gives you the **statistical workflow**, which is software agnostic in that they can be applied in any software.
- There are also accompanying **software workflows** that show you how to do it. We won't be going through these in detail. But if you have problems we have a monthly hacky hour where people can help you.

1on1 assistance

- You can email us about the material in these workshops at any time
- Or request a consultation for more in-depth discussion of the material as it relates to your specific project. Consults can be requested via our Webpage (link is at the end of this presentation)

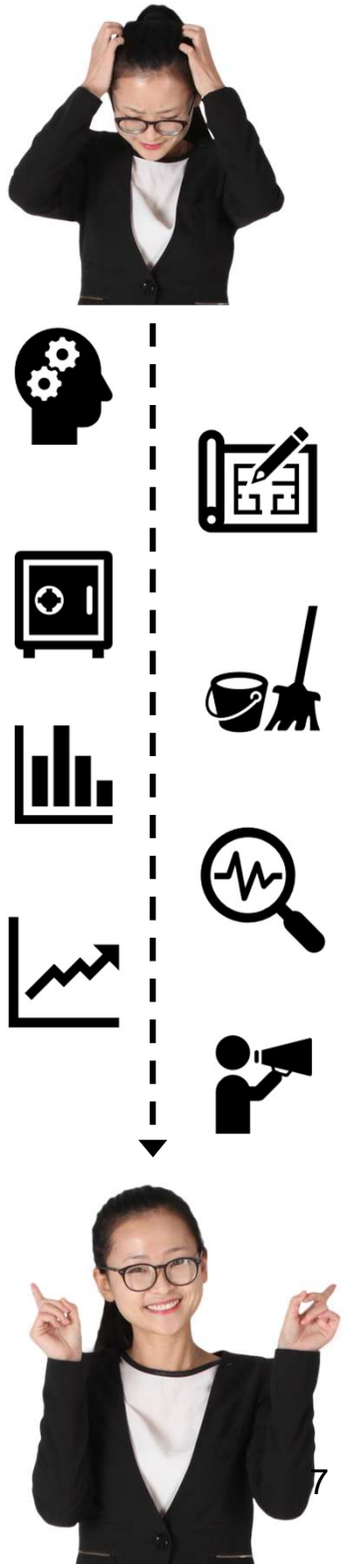
Research Workflow

- Why do we use a research workflow?
  - As researchers we are motivated to find answers *quickly*
  - This drive can cause problems if we don't think systematically
  - … and we need to in order to:
    - Find the right method
    - Use it correctly
    - Interpret and report our results accurately
  - The payoff is huge, we can avoid mistakes that would affect the quality of our work *and* get to the answers sooner



- So… what is a workflow?
  - The process of doing a statistical analysis follows the same general "shape".
  - We provide a general research workflow, and a specific workflow for each major step in your research
    (currently **experimental design, power calculation, analysis using linear models/survival/multivariate/survey methods**)
  - You will need to tweak them to your needs

General Research Workflow



1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**

Content

Survey Design and Validation

- Initial Design
- Pre Testing
- Post Testing
- Dimension Testing
  - Exploratory Factor Analysis using Factor Analysis
  - Confirmatory Factor Analysis using Structural Equation Modelling

Index Creation

Conjoint/Choice Models and Best Worst (Max Diff)

A Conversation is better than a Presentation

So please speak up and ask questions!

People think differently. So I may need to explain things in 2 or 3 different ways!

Survey Design and Validation

THE UNIVERSITY OF SYDNEY

Survey Modes: 4 common methods

1. Online
2. Paper
3. Face to Face (F2F)
4. Computer Assisted Telephone Interviews (CATI)

Survey Modes: Online and paper surveys

Online is the dominant survey mode so we focus on them. Paper is less commonly used.

Online's dominance began in about 2005 in developed countries and has spread with the ubiquity of the internet and smartphones. As this is relatively recently older references may be of little use as they will not consider online.

The main reason was their **cheap cost,** enabling substantially more sample for less money due to fewer labour costs (no interviewers required, no data entry).

A frequently **overlooked problem is their reliance on online panels, which calls into question their representativeness and hence if they can be generalised to the wider population.** Refer to our Experimental Design workshop for more info on samples that can be generalised to the wider population. A common way around this is to recruit one's own sample and not rely on established panels.

Paper and online share many similarities in terms of how respondents answer them. However online:

- Can have fancier and interactive questions
- Complex adaptive filters and piping
- Is cheaper

Survey Modes: Interviewer Surveys such as Face to Face (F2F) and Computer Assisted Telephone Interview (CATI)

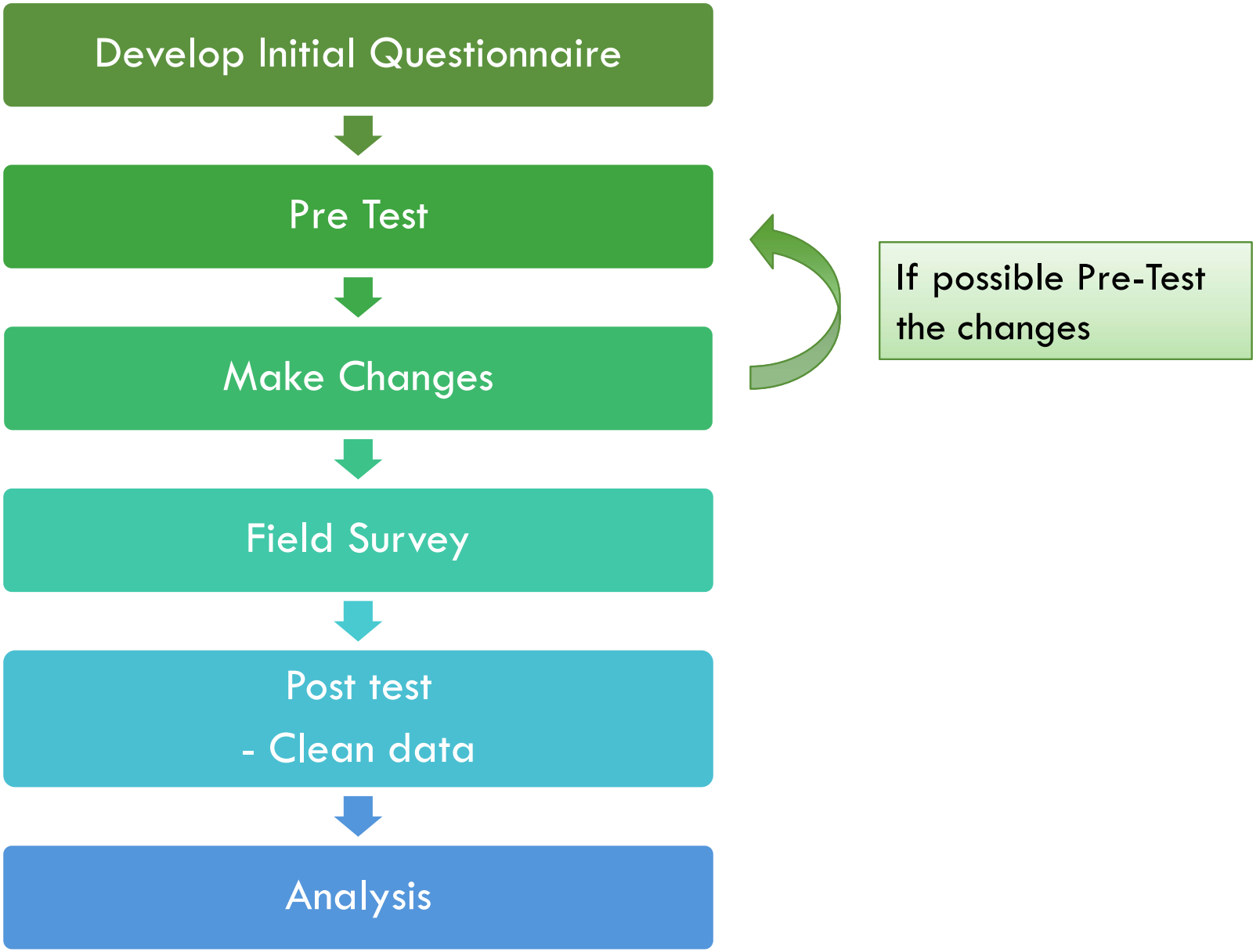
Prior to online when people often asked the questions there was the **added complexity of factoring in the interviewer.** For example

- if it was a **sensitive topic** people may not answer truthfully if giving their answer to another person
- **different interviewers may bias** the responses differently based on how the ask the question.

There are established methods for including the interviewers in the validation and pretesting which is not needed with online/paper surveys.

Interviewer surveys are not covered in detail in this workshop due to lack of time and their uncommon use. For a detailed discussion of topics relevant to interviewers refer to Presser et al (2004) Methods for Testing and Evaluating Survey Questionnaires.

Validation Workflow



How to make a great Survey: The basics from Surveys I

1. Write a draft in word based on your needs, desktop research and qualitative work.
2. Leave it at least a few days, ideally a week, then review. Keep doing until no new edits.
3. Feedback from friends and colleagues.
4. Set up in Survey Tool e.g. REDCap or Qualtrics.
5. Enter some pilot data: You can make it up by just thinking of some likely but very different respondents and enter as they would.
6. Export data and ensure you have set it up so the data exports in an easy to analyse format.
7. Leave it at least a few days, ideally a week, then review. Keep doing until no new edits.
8. Send link to friends and colleagues and get some feedback.
9. Leave a final week, review and send off to Ethics for approval.
    1. Even small changes can be problematic as new ethics approval is often required, this is one reason it is so important to get it right before submitting to ethics.
10. Go Live!!!
11. Review the first 12-50 respondents for any problems
    - Look for missing categories you should consider adding by reviewing the open ender linked to 'other' to see if it has a lot of responses representing the same thing. (Particularly worth keeping an eye on as the survey progresses to avoid time consuming back coding later.)

This method assumes a straight forward survey with established questions and scales. So little testing required.

Let's look at how to do more detailed testing if required.

Testing and Design Research: How much is enough?

Less, if the survey is about:

- Easily understood concepts e.g. How satisfied are you?
- Established scales e.g. LIKERT scales.
- Validated or commonly used questions or formats.

More, if the survey is about:

- Hard to understand concepts e.g. do you have joint custody (as opposed to physical custody)
- Sensitive topics, to ensure accurate answers and low non response e.g. drug use.
- Translated into different languages.
- Children respondents.
- New scales.
- New or novel questions or formats.

Pre Testing: Mandatory

Pilot Survey

- Purpose is to identify:
  - obvious problems
  - missing
    - Categories in categorical variables
    - Dimensions of interest
- Recommended Sample Size: 12-25 cases (Sheatsley, 1983, p226). 20-50 (Sudman, 1983, p181)
- Perform the standard Post Test Questionnaire Metric Tests on them e.g. flatliners, outliers, response time, etc.
- Add these Free Text questions
  - Split the questionnaire into sections and at the end of each ask "Was there anything difficult to answer, any improvements you might suggest or anything else you want to tell us about the last section?"
  - Include a final open ender at the very end asking the same thing.
  - It's often useful to retain a 'is there anything else you want to tell us?'
- These respondents can be colleagues, friends, etc. However at least a few should be unknown to the researchers.

Debrief pilot respondents, or at least some of them.

Pre Testing

Optional

- Qualitative e.g. Focus Groups, Cognitive testing.
- Behaviour testing i.e. observing how respondents react to the interviewers questions, and how the interviewers behave. For example, questions that needed frequent repeating maybe problematic.
- Experimental Testing i.e. evaluate different questions with different respondents.

Declared vs undeclared

Declared pretesting means people are aware of their involvement. This can lead to different responses. And is ***particularly a problem when pretesting interviewers*** since they may be on their 'best behaviour' which is different to the their normal interviewing style.

Post Testing

THE UNIVERSITY OF SYDNEY

3 types of Post Testing

Questionnaire metrics

- ***Mandatory:*** no excuse not to do
- Involves looking at relevant metrics for each question to look for problems e.g. the distribution to see if there are with a lot of missing data or with poor differentiation, response time, flatliners, etc.
- If the sample is large and being asked over a long period of time it's worth monitoring these metrics while in field and tweak the survey if necessary. However:
  - Any such changes may make it harder to combine the data and analyse it. So ensure the cure isn't worse than the disease!
  - This may be unpractical to do in an Academic setting due to the need to resubmit for Ethics approval.

Reliability metrics (aka equivalence)

- ***Optional:*** usually requires respondents/interviewers/coders to answer the same question more than once so not always possible. And not necessary for well established questions and scales.
- Tests the 'reliability' of the answers e.g. if someone answers the same question again how often do they give a different answer?

Evaluating and fine tuning statements used to quantify dimensions

- ***Optional*:** not always relevant, nor necessary for well established questions and scales.
- Methods for evaluating and finetuning the statements used to quantify the dimensions e.g. Confirmatory Factor Analysis via Structural Equation Modelling (CFA via SEM).

Questionnaire metrics and EDA (Exploratory Data Analysis)

THE UNIVERSITY OF SYDNEY

Non response: #/% who didn't answer i.e. missing values.

Response time: note that longer times may indicate a harder question and not necessarily a problem with it. What is a problem are *Racers* i.e. people finishing too quickly, they could even be bots.

Response distributions: are they behaving as expected? Look for:

- Outliers
- Poor differentiation i.e. only "Agree" being used in a 5 point LIKERT scale.
- Flatliners (refer to Surveys 1).

Categorical variables: are they behaving as expected? Look for:

- missing categories you should consider adding by reviewing the open ender linked to 'other'. If it has a lot of responses representing the same thing consider adding them as a hard coded option.
  - Particularly worth keeping an eye on as the survey progresses to avoid time consuming back coding later.
  - But can be hard to do in an Academic setting as it often requires going back through Ethics. Which is why it is so important to do some pilot questionnaires and qualitive work before submitting to Ethics.

Reliability Metrics

THE UNIVERSITY OF
SYDNEY

Reliability Metrics

There are too many different metrics and scenarios to cover in the time available. If in doubt it is often easiest to use those accepted in your domain.

Rather than try to cover all the different metrics we will give you the information you need to do the analysis, even if you need to slot in a metric not covered here. So we will cover:

- Introduction
- Analysis Workflow
  - Suggested Metrics
- References
  - Has a lot of information on validation of interviewers, not so much on online surveys. Presser. S, Rothgeb J.M., Couper M.P, Lessler J.T, Martin. E, Martin. J, Singer. E (2004) *Methods for Testing and Evaluating Survey Questionnaires*. Wiley-Interscience
  - A short online course in R. https://www.datanovia.com/en/courses/inter-rater-reliability-measures-in-r/

Reliability Metrics: 2 main types

Intra Rater Reliability

- The **error/variance within** a rater: caused by respondents not answering the question identically each time.
- A common source of error, albeit often a rather small one.
- Often evaluated with test-retest data.

Inter Rater Reliability

- The **variance between** raters: caused by different raters scoring the same thing differently.
- This is more usually a problem when we are using raters to quantify something of interest e.g. trained panellists evaluating the same food products, psychological research when behaviour is being coded.
  - And is also a problem when we have different interviewers administrating the survey Face to Face (F2F) or over the phone (CATI-computer assisted telephone Interview).

Sources of Error/Variance

Measurement Error

- Also known as Scale Error
- The error when a respondent is trying to give the same answer, but the scale or collection method used prevents that.
- Examples:
  - Line scales i.e. try marking 50 on a 100cm line scale inevitably results in scores of 49, 51 etc.
  - Unanchored LIKERT scales

Respondent Error

- Is the respondent themselves not being consistent e.g. given exactly the same piece of cake a well trained panellist will mark sweetness approximately the same, a poorly trained one will give very different scores.

Interviewer Variance

- The variance associated with different interviewers e.g. a surly interviewer might elicit different answers to a happy one.

Question/Instrument Error

- The Error caused by an ambiguous or confusingly worded question.

Fixing Error/Variance

Measurement Error

- Use a better scale! So avoid line scales, add text anchoring, etc.

Respondent Error

- Train the respondents better e.g. better lead in's to questions, pre survey briefing, for coders more examples and benchmarks, etc.

Interviewer Variance

- Train the interviewers better e.g. train them to be more consistent using benchmarks.

Question/Instrument Error

- Use qualitative work to improve it e.g. cognitive testing.

How to collect the data for Intra Rater Reliability

Test-Retest: Requires the respondent to answer the same question twice there are at least 3 ways to do this.

Method	PRO	CON
Ask the entire survey at a later time	• Complete data for each respondent. • Respondents less likely to remember previous answer i.e. independence.	• If done at a later date their response can legitimately change. • Usually annoys respondent which can affect results. • Can rarely get all respondents to participate.

Ask a subset of the survey at a later time	• Easier for respondent then redoing entire survey, so
	  • possible better results.
	  • more respondents will participate
	• Respondents less likely to remember previous answer i.e. independence.
		• If done at a later date their response can legitimately change.

Sneak in the same question in the same survey	• Easy to get all respondents to participate. • Easy for respondents to do, likely better results.	Respondents more likely to remember answer i.e. independence is less likely.

Analysis Workflow

1) **Determine what method and metrics to use** by deciding if the data is:
    1) A Continuous, Nominal or Ordinal scale
    2) From 2 raters or more

2) EDA

- Used to understand where and how the raters are agreeing vs disagreeing. Helps diagnose and fix the problem.
- Hard to use for more than 3 sets of raters/ratings
- If lots of questions are evaluated it's usually easier to first look at the Agreement Metric to identify which have poor agreement and then use these tables to diagnose where the problem is.

3) Calculate Reliability Metric

- Are a simple 1 score metric representing Agreement making them easy to compare between studies and if there are lots of questions one wishes to test.
- These are often interpreted in a similar way to Pearsons correlation co-efficient i.e. 0 means no Agreement, 1 = Strong Agreement, -1 = Strong Disagreement.

Continuous Scale Workflow

EDA

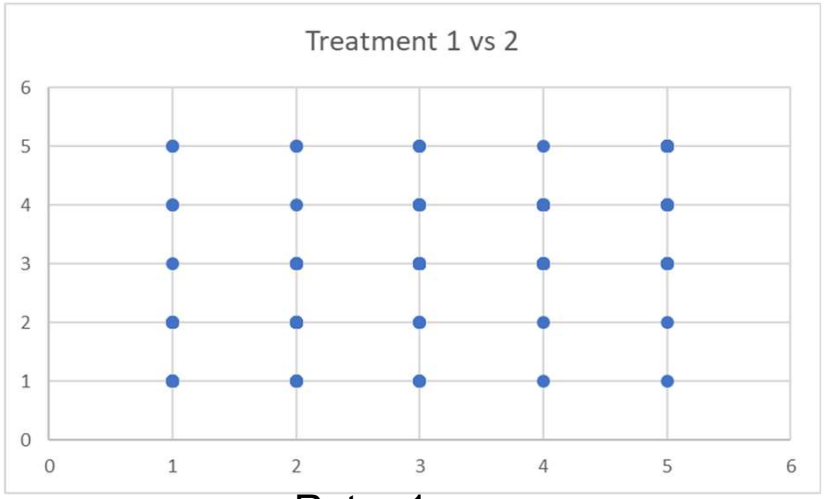
- Plot the data
- If there are only few possible answers on the scale you will need to either jitter them or use a bubble plot e.g. Likert scales.

Calculate Reliability Metric

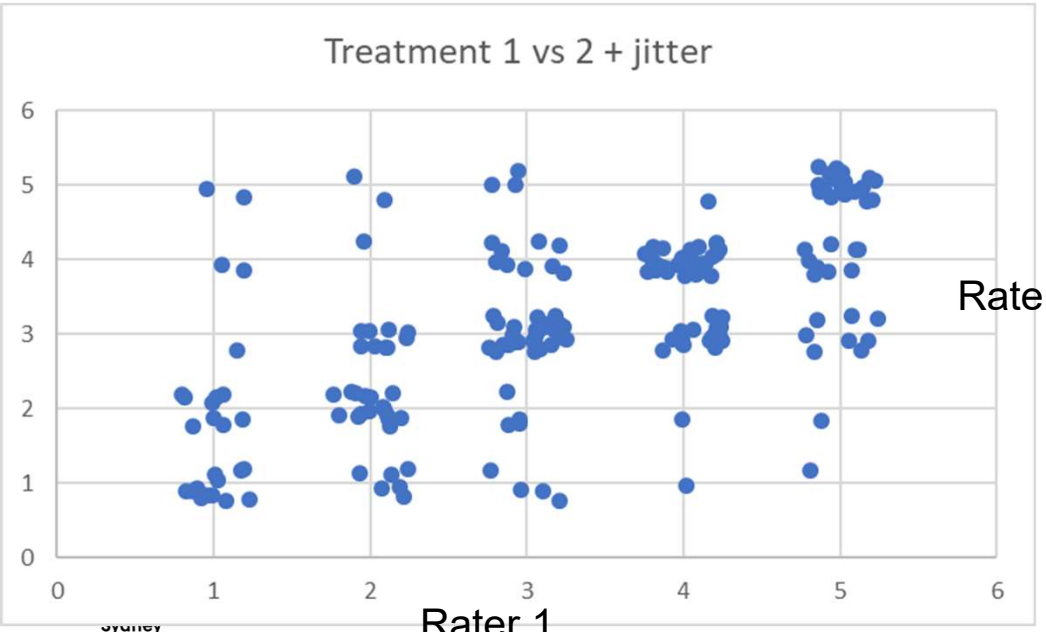
- Intra Class Correlation Coefficient (ICC)
  - Good for Intra and Inter rater reliability, depending on how we define groups.
  - Will work for more than 1 rater.
  - A high ICC (close to 1) indicates high similarity between values from the same group/rater.
  - A low ICC (ICC close to 0) means that values from the same group/rater are not similar.
- Concordance Correlation Coefficient
  - Good for Inter rater reliability.
  - Gives a correlation metric that includes a bias factor which is the difference from the 1:1 line. As opposed to just using the Pearson linear correlation which only tells us if they are correlated, not if they are in agreement.
  - Only works for 2 raters, although one might repeat it for all combinations of raters if there aren't too many.

EDA: Why you need a jitter if there are only a few points e.g. LIKERT Scales

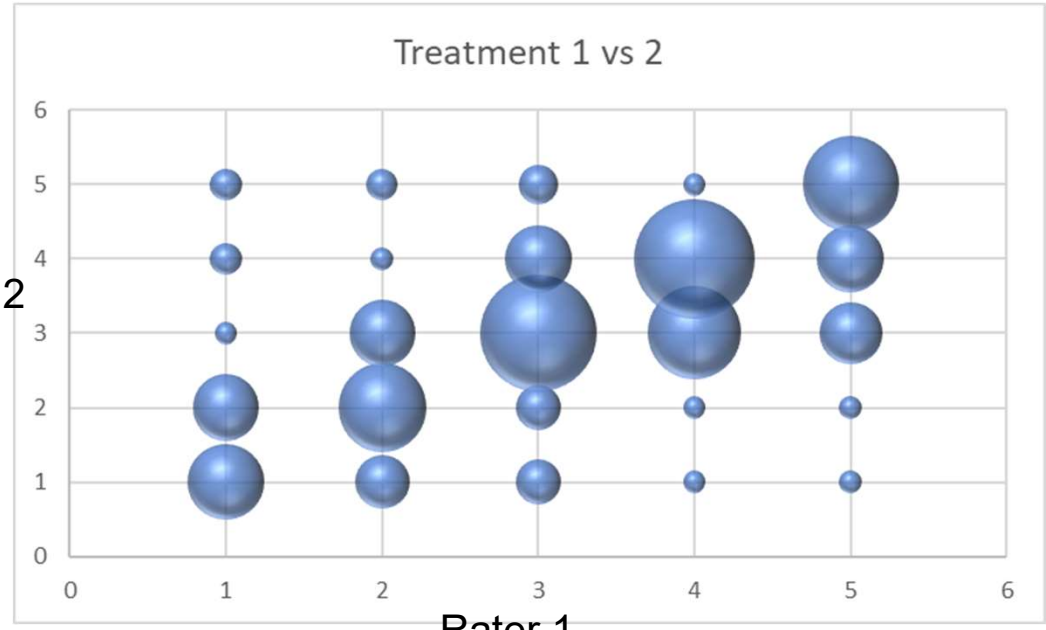
As the number of scores are limited it often comes out as a grid!!! Which doesn't help us much since we don't know how many times each combination actually occurs!



Fix by adding some jitter



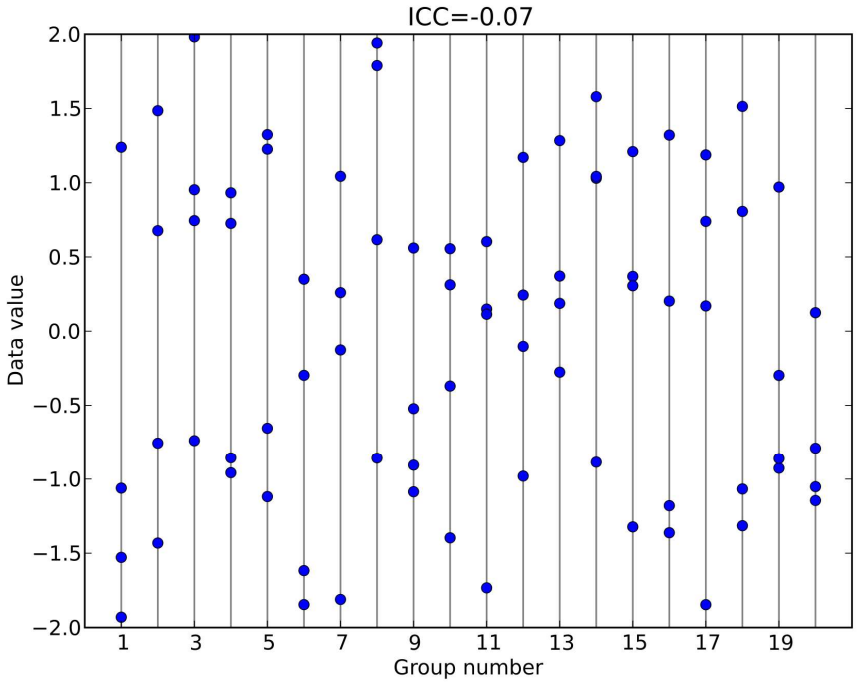
Or a bubble plot

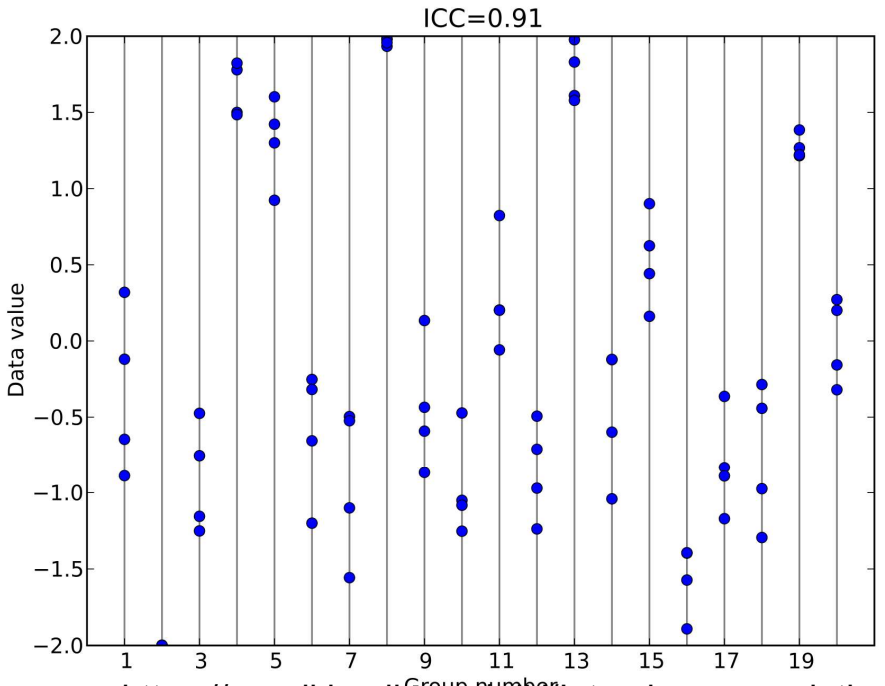


Intra Class Correlation (ICC)

Intra Rater: If each group/line is a different doctor telling us how sick they think a patient is over multiple days. Then the dot plot on left shows low intra rater reliability, while the dot plot on right shows high intra rater reliability. (Assuming patient is stable).

Inter Rater: If each group is a different patient and the dots different doctors telling us how sick they think they are. The dot plot on left shows low inter rater reliability, while the dot plot on right shows high inter rater reliability.





https://en.wikipedia.org/wiki/Intraclass_correlation

Concordance Correlation Coefficient

The solid line is the 1:1 equivalence line.

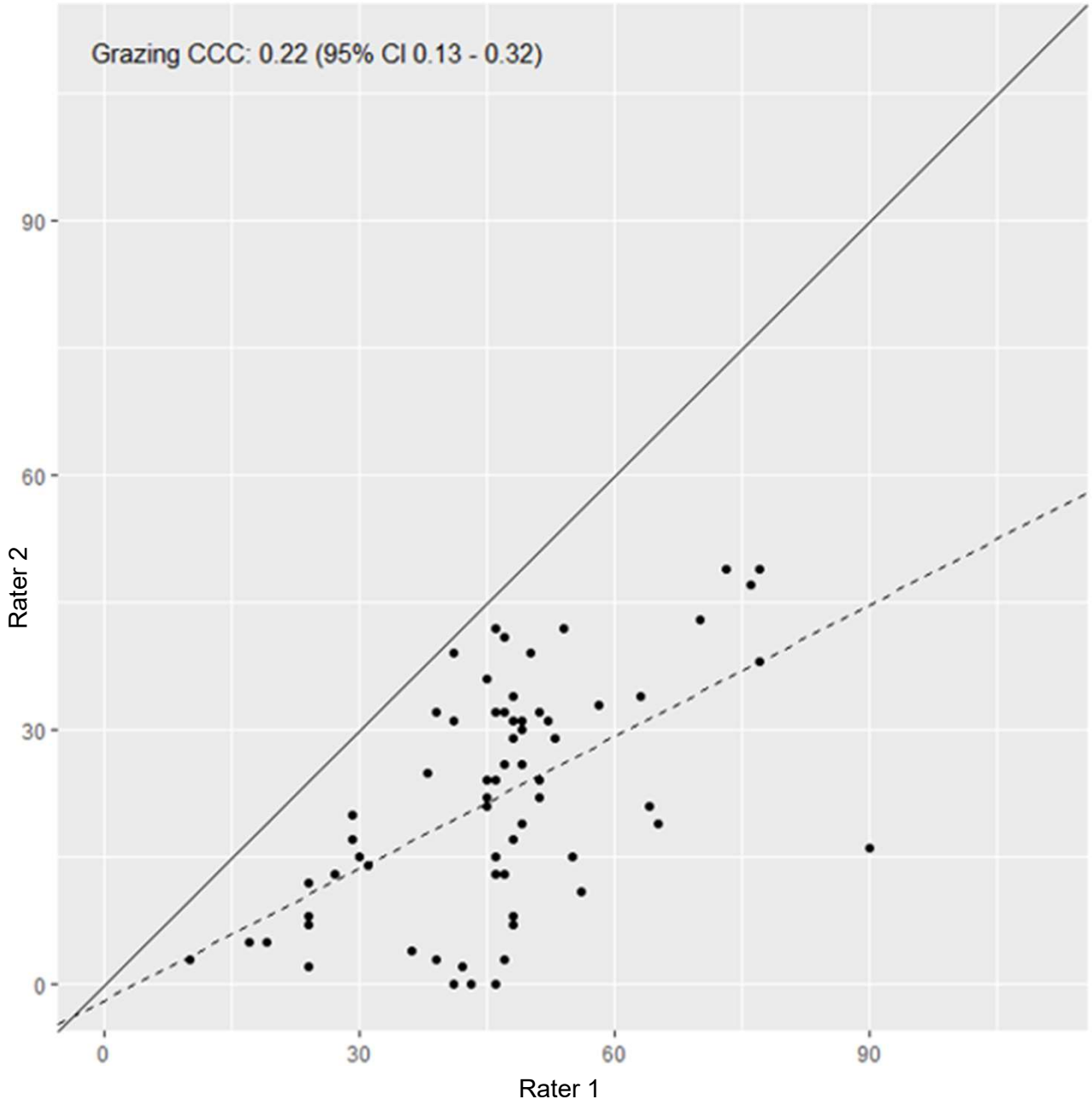
The dotted line is the actual correlation.

The metric shown here is the Concordance Correlation Coefficient.

This example show us there is fairly good correlation.

But rater 2 is consistently under rater 1.

So we might be able to bring them into alignment using a linear correction.



Categorical Scale Workflow: EDA

- Uses 2x2 contingency tables to understand where the raters agree and disagree. Usually uses Prediction Accuracy and % which shouldn't be used as the overall metric as they don't address the problems raised below. But they are easy to understand and do help us diagnose where the problems are.
  - diagonal is #/% they agree with
  - off diagonal is #/% they disagree with
- Note that 2 way tables only works for 2 raters since it only shows the 2 way possibilities. One can use a '3 way' table if there are 3 raters, once we have more than that it gets difficult to keep track of though!
- Example below shows that the problem is that Rater 1 is Strongly agreeing with things that Rater 2 is only Agreeing with.

				Rater 1		
	Count	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
	Strongly Disagree	20	2	1	0	0
Rater 2	Disagree	3	30	3	1	0
	Neither	1	2	40	5	1
	Agree	0	1	4	80	**25**
	Strongly Agree	0	0	2	4	20

				Rater 1		
	%	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
	Strongly Disagree	8	1	0	0	0
Rater 2	Disagree	1	12	1	0	0
	Neither	0	1	16	2	0
	Agree	0	0	2	33	**10**
	Strongly Agree	0	0	1	2	8

EDA: 3 rater table example

Dillon and Mulani (1984, p. 449)

A cognitive response study for which three raters classified 164 subjects in $K = 3$ categories positive (**1**), neutral (**2**) and negative (**3**).

Rater 3		**1**			**2**			**3**	
Rater 2		**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Rater 1	**1**	56	1	0	5	3	0	0	0	1
	2	12	2	1	14	20	4	0	4	2
	3	1	1	0	2	1	7	2	1	24

Cross-classification of 164 subjects by three raters

Prediction Accuracy (counts and %)

	PRO	CON
Both	• Easy to understand. • Can be broken down in agreement tables.	Doesn't factor in: • # we expect to get right by chance. • Sample size of each score
Counts	• Shows the actual number so we know if we have a few or a lot. e.g. a % of 2% might mean only 2 people or 1000 depending on the sample size.	• A bit harder to interpret than the % since we also need to factor in sample size. • Hard to compare between studies with unequal n.
%	• Easy to compare between studies with different sample size.	

Categorical Scale Workflow: Analysis Common metrics

There are a lot of different metrics. These are only a few. If in doubt it is often easiest to use those accepted in your domain.

Metric	# of raters	Scale	Missing data
Cohens Kappa	2	Ordinal or nominal. Best for nominal (some say only for nominal)	No
Fleiss Kappa	2+	Ordinal or nominal.	No
Krippendorfs alpha	2+	Ordinal or nominal.	Yes

Cohens: What's a good enough score?

Interpretation: -1 = Perfect Disagreement, 0 = No pattern, 1 = Perfect Agreement

Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics*. **33** (1): 159–174.

- I'm told that they supplied no evidence to support it, basing it instead on personal opinion.
- They characterized values:
  - < 0 as indicating no agreement
  - 0–0.20 as slight
  - 0.21–0.40 as fair
  - 0.41–0.60 as moderate
  - 0.61–0.80 as substantial
  - 0.81–1 as almost perfect agreement

Fleiss, J.L.; Cohen, J.; Everitt, B.S. (1969). "Large sample standard errors of kappa and weighted kappa". *Psychological Bulletin*. **72** (5): 323–327

- I'm told that they supplied no evidence to support it, basing it instead on personal opinion.
- They characterized values:
  - 0.40 as poor
  - 0.40 to 0.75 as fair to good
  - 0.75 as excellent

Tricks

If only a few raters we can do the analysis for 2 raters at a time and repeat it.

To diagnose where the problems are run the metric for all of them, and then each pair of raters.

They grouped statements in to 3 categories. Positive statements (I love swimming), neutral statements (swimming is OK fun), negative statements (I hate swimming). This is represented in the Categories (i) column.

Challenge Question: One of their research questions was that Raters would have a harder time with neutral questions. What does this table say about that?

Raters did indeed have problems with Neutral statements (0.2462), and were consistent for Positive and Negative questions. Raters 1 and 2 had poor agreement (0.160), while rater 3 seemed to have the opposite opinion to the other 2 rates (-0.474, -0.570).

Evaluation of raters:

Categories	Consistencies			
(i)	Raters 1, 2 and 3	Raters 1 and 2	Raters 1 and 3	Raters 2 and 3
1 *Pos.*	0.7040	0.706	0.788	0.810
2 *Neu.*	0.2462	0.160	− 0.474	− 0.570
3 *Neg.*	0.6306	0.764	0.714	0.709
Overall	0.550	0.567	0.329	0.413

References

Sheatsley, P. (1983) Questionnaire construction and item writing, in P. Rossi, J. Wright, and A. Anderson (eds.), *Handbook of Survey Research*. San Diego, CA: Academic Press

Sudman, S. (1983) Applied Sampling, in in P. Rossi, J. Wright, and A. Anderson (eds.), *Handbook of Survey Research*. San Diego, CA: Academic Press

Presser. S, Rothgeb J.M., Couper M.P, Lessler J.T, Martin. E, Martin. J, Singer. E (2004) *Methods for Testing and Evaluating Survey Questionnaires*. Wiley-Interscience

Evaluating and Fine tuning Statements used to Quantify Dimensions

THE UNIVERSITY OF SYDNEY

What are Dimensions/Factors and why use them?

Prior research and qualitative work often identifies dimensions of interest the researcher wants to understand. Indeed this is often the primary reason for the survey.

Even if not the focus it's still a good idea to identify possible dimensions that might impact the research prior to developing the survey. For example:

- Business: Price, Quality, Animal Welfare
- Vaccines: Education, previous bad experiences

These dimensions are then included in the survey, which is used to Quantify their impact.

If one is using **rating scales it is common to assign 2-5 statements per dimension and use these to quantify each one.** The simplest way is to simply average them, a more complex way is to create an index from them.

Creating and Validating the Dimensions: Steps

Step 1) Initial Design

- Assign statements to each dimension and field the survey to get data. If in doubt on which to include simply include them all (within reason) since further steps can be used to select the best.

Step 2) Select and Refine the statements used to define each Dimension

- Remove unreliable statements
- Remove redundant statements
- Refine statements used in each dimension e.g. some statements might be more appropriate for a different dimension.
- Refine how they are being asked.
- Ensure they are 'loading' onto the dimensions as expected i.e. validation.
- We want at least 2 statements in each dimension to ensure it is robust and stable i.e. if they incorrectly answer or there is poor DE for 1 statement it is 'corrected' by the other's. If we only had 1 statement then this incorrect data has great impact on the analysis.
  - 3 is usually the minimum to avoid Heywood cases during CFA/SEM.

Step 3) Model and confirm the statements used to define each Dimension

- Create a formal Path Model for each dimension based on the input statements.
- Fit the Path Model.
- Confirm statements are 'loading' onto the dimensions as expected i.e. validation.

Creating and Validating the Dimensions: Analyses

Select and refine the statements used to define each Dimension

1. Reliability analysis: to ensure reliable statements are used.
2. Pairwise Scatterplot and Correlations: Grouped by statements in each dimension.
3. Exploratory Factor Analysis (EFA)
    1. EFA is actually a concept, usually done with factor analysis. Which is why it is often confounded with that method.
    2. Helps us understand the correlation amongst the statements to determine what dimensions might exist.

Model and confirm the statements used to define each Dimension

1. Confirmatory Factor Analysis (CFA)
    1. CFA is actually a concept, usually done with Structural Equation Modelling SEM). Which is why it is often confounded with that method.
        - A rough hack is to use Factor Analysis on each dimension.
    2. Confirms if the statements are organising themselves into pre-defined dimensions as expected.

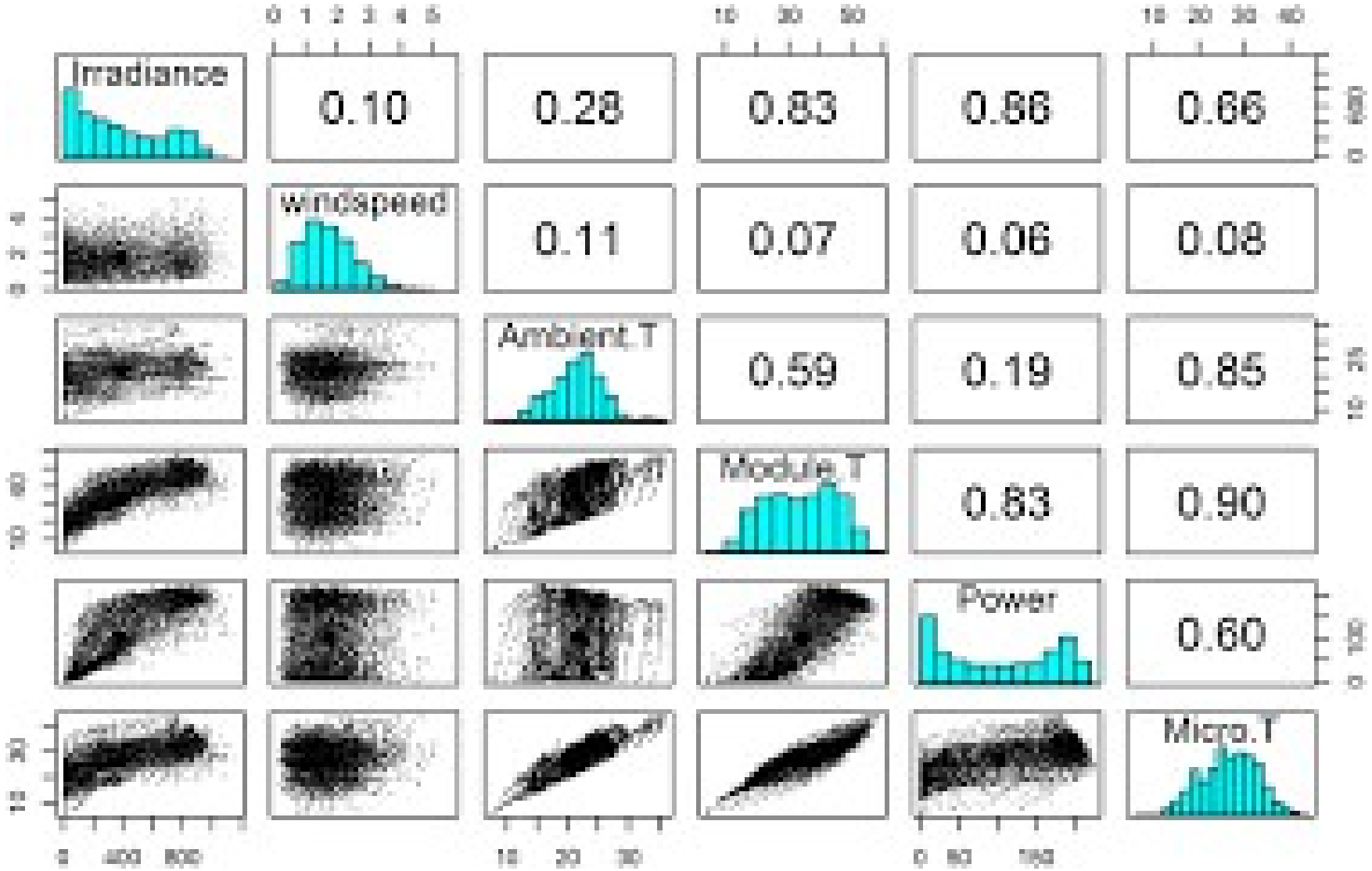
Pairwise Scatterplot and Correlations

We are looking for:

1. **Large Positive Correlations Statements,** means they are redundant so we can remove some of them (obviously keeping at least one) e.g. height(inches), height (cm), weight. We'd only keep one of height. Keeping in mind that we tend to want 3-4 items in each dimension.
2. **Correlations close to 0.** Means either they are:
    1. Not related and at least 1 should be removed.
    2. Different aspects of the dimension and can be kept e.g. "I always buy products on special if they are good value" and "I always buy products that have a consistently low price" are very different and might have very low correlation, but both could be included in a Price dimension.
        - In this case one might decide to split them into 2 sub-dimensions. However this may over complicate things and not be required.
3. **Negative correlations.** Have a close look at these ones to decide if the negative correlation makes sense and can be kept or is a problem. Common issues are that:
    1. The statement is phrased as negative one when the rest are positive (or vice versa), it is often best to change it so they are all in alignment. This can be kept.
    2. It makes no sense and needs to be dropped or changed.
4. **Linearity** (since the underlying correlation metrics used in EFA/FA and CFA/SEM assume this. If not linear then a non standard factor analysis that doesn't use linear correlation can be used).

Some papers and books suggest that we should only keep statements that have large positive correlations, and correlations close to 0 should be removed. I disagree with this since large positive correlations can mean redundant statements while correlations close to 0 can mean we are capturing different aspects of the dimension.

Scatterplot Matrix



https://www.researchgate.net/figure/Pairwise-scatter-plot-matrix-histogram-and-correlation-coefficients-of-all-related_fig3_280031491

Exploratory Factor Analysis

We are looking for:

1. **The same correlation issues** we looked for during the previous pairwise correlation.
2. **Factors with unexpected statements or dimensions.** This indicates something isn't working and maybe this statement should be assigned to a different dimension, or this dimension split in 2.
3. **How highly correlated statements correlate (load out) differently for the part they aren't correlated on.** This would usually be a factor that explains a small amount of variance.
4. Look out for the 1st factor just being a high/low rater dimension, this will be an average of most (if not all) of the statement. Depending on the survey you may, or may not, be interested in these people. For example if it's a satisfaction survey you want to distinguish between low vs highly satisfied people. But if it was an attitudinal survey you may not be interested in people who are grumpy (low Likert scores) vs optimistic (high Likert scores).

One potential problem with EFA is that it rarely combines 'cleanly' into the expected dimensions. We often need to force it to, which is what CFA does. We can get around this problem during EFA by also doing a separate factor analysis for each dimension to better understand that specific correlation structure.

- This isn't entirely unexpected. One reason for this is that our dimensions may be at different scales within the data e.g. we might have 3 price and 1 quality dimension. The 3 price dimensions might then split into 2 price dimensions not the 3 we want.
- Note that as we go up in scale this actually has to happen, especially if we have fewer factors in the EF than dimensions we are looking for! Eventually even non correlated sub-dimensions might be combined into the same factor.

In general we are removing those with:

- Poor Reliability (test/retest metrics)
- Redundant
- Theoretically silly
- Loading onto multiple domains (unless we are OK with non-orthogonal factors)
- Factor loading < some cut-off (say 0.6). Be careful here, the items included can influence them so do this last and remove statements one at a time then rerun the factor analysis to see which loadings are <cut-off.

Model and confirm the statements used to define each Dimension

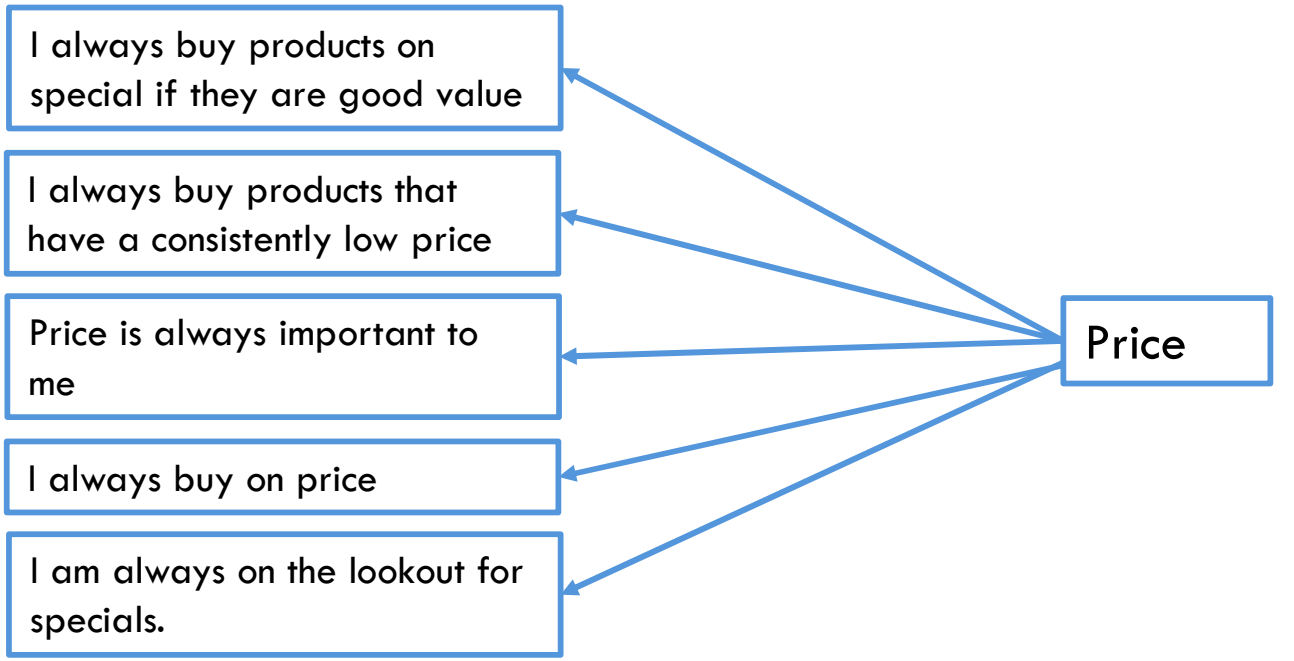
Confirmatory Factor Analysis (CFA)

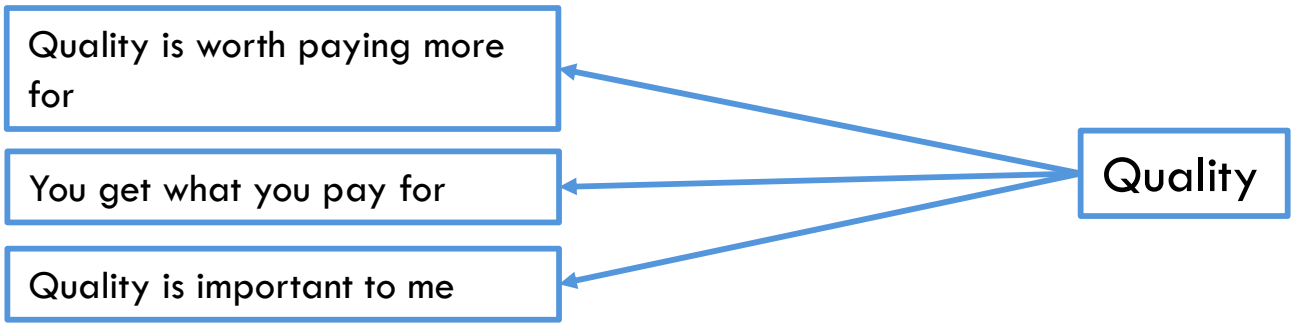
1. CFA is actually a concept, usually done with Structural Equation Modelling SEM). Which is why it is often confounded with that method.
   - A rough hack is to use Factor Analysis on each dimension.
2. Confirms if the statements are organising themselves into pre-defined dimensions as expected.

We do this by:

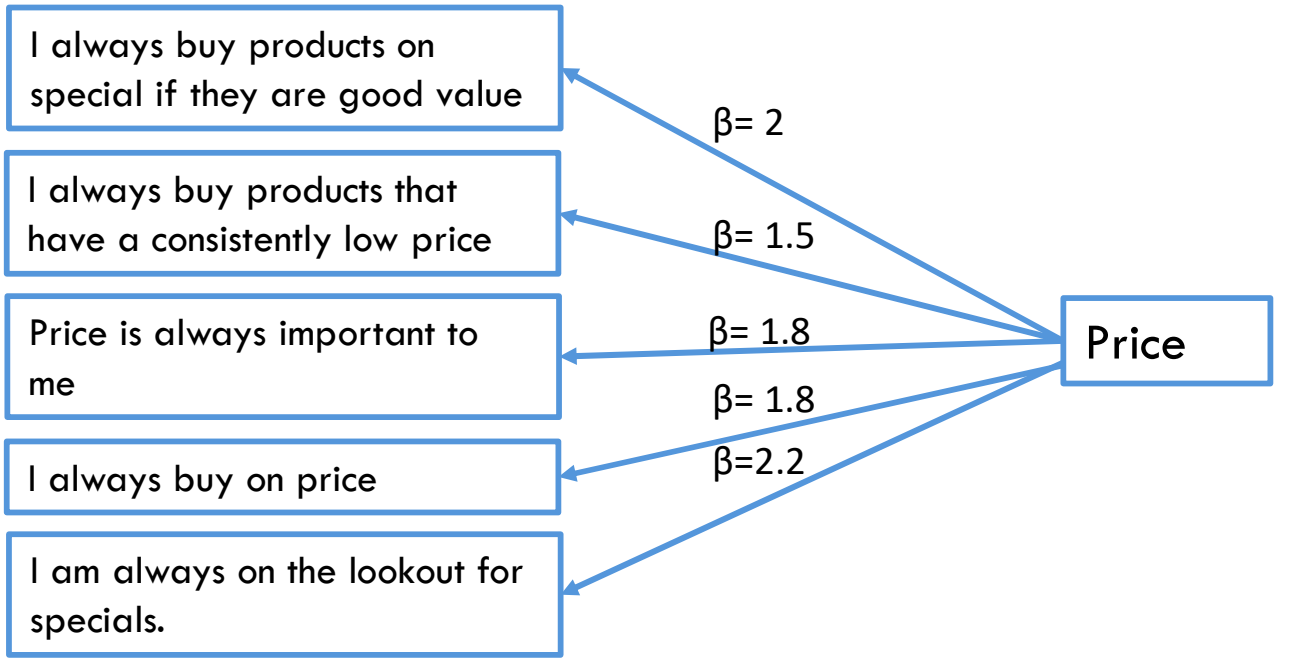
1) Defining the model structure usually using a Path Model
2) Fitting the model
3) Testing the model's assumptions and Goodness of Fit
4) Interpreting the model to confirm statements are 'loading' onto the dimensions as expected i.e. validation.

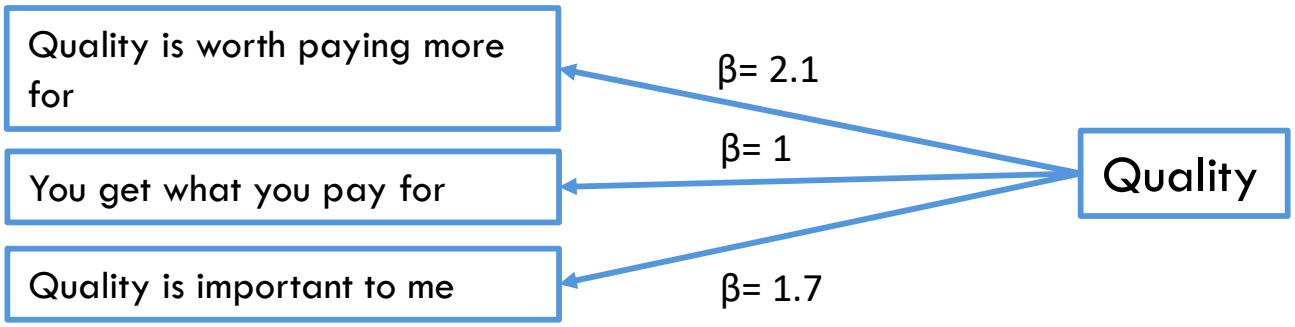
Define the Path Model





Fitting the Path Model





Testing the model's assumptions and Goodness of Fit

There are numerous ways to test the model assumptions and Goodness of Fit. Both should always be done and to some extent the preferred metrics are domain specific. We will cover only the most important and commonly done.

Testing Model Assumptions

Appropriate correlation metric used Linear correlation is the standard metric and should have already been tested with the pairwise scatterplots.

Testing the model's assumptions and Goodness of Fit

Model Fit Statistics

Parameters/loadings

- Direction, magnitude and statistical significance of parameter estimates
- Check loadings are all high (>0.6 assuming standardised coefficients), if not then maybe we have more than 1 factor here.

Testing the model's assumptions and Goodness of Fit

Model Fit Statistics

Exact fit indices - Matsunaga (2010)

Chi-Squared

- How close the model fits the data's covariance matrix.
- Significance means there is a significant difference, which ***could*** mean poor fit (see con below). Meaning we are actually looking for non significance since this indicated no evidence of a poor fit, which isn't ideal since we don't want to 'accept' the null. All we are really saying is that there is no evidence of a poor fit.
- CON: Highly susceptible to sample size. Larger sample size can lead to a statistically significant result, but the fit is still acceptable.

Testing the model's assumptions and Goodness of Fit

Model Fit Statistics

Plus at least 2 others from (Hu and Bentler as per Matsunaga 2010).

- **Approximate Fit Indices:** how close the model fits the data.
  - RMSEA (Root Mean Square Error of Approximation). Estimates the amount of Error or Approximation per DF, and as such accounts for sample size.
    - Acceptable benchmarks:
      - <0.06 (Hu and Bentler)
      - <0.08 (Marsh; Thompson)
    - Unacceptable benchmarks:
      - >0.1
- **Incremental Fit Indices:** model fit over a "null" model (no structural path, factor loadings or inter-factor correlations).
  - Some common indices are CFI (Comparative Fit Index), TLI (Tucker-Lewis/Non-Normed Index), RNI (Relative Noncentrality Index). Acceptable benchmarks (lower is OK):
    - <0.95 (Hu and Bentler)
    - <0.9 (Russel)
- **Residual Indices:** Covariance residuals between data and model.
  - SRMR (Standardised Root Mean Square Residual). Average standardised residuals. Acceptable benchmarks (lower is OK):
    - <0.10 (Hu and Bentler; Kline)

Always do EFA with CFA since

CFA might show they fit the underlying model you want to test. With all statements loading onto dimensions as expected.

BUT EFA might show they load in slightly unexpected ways that might help improve the overall model e.g. 2 statements in different dimensions might be very highly correlated.

Heywood cases

Heywood cases are a common problem.

They occur when we see negative variances and R-squared values greater than 1. Neither are theoretically possible, meaning we have a problem and the rest of the estimates are not reliable.

It occurs when the model lacks enough information to estimate the dimension and is fixed by adding more statements to it.

This is 1 reason why we say 3 is the bare minimum # of statements per dimension.

References

Matsunaga M (2010) How to Factor Analyse your data right. Do's Don'ts and How To's. *International Journal of Psychological Research*.

Analysis

THE UNIVERSITY OF SYDNEY

Analysis using CFA

Research Objective Example

Some argue that cheap private label brands have a negative impact on local farmers and suppliers. For example:

- Woolies/Coles selling milk for less than they were buying it put a downward pressure on prices at the farmyard gate.
- Woolies/Coles don't invest in new products. So if you want a delicious new Tim Tam this has to come from Arnotts profits being invested in new product development.

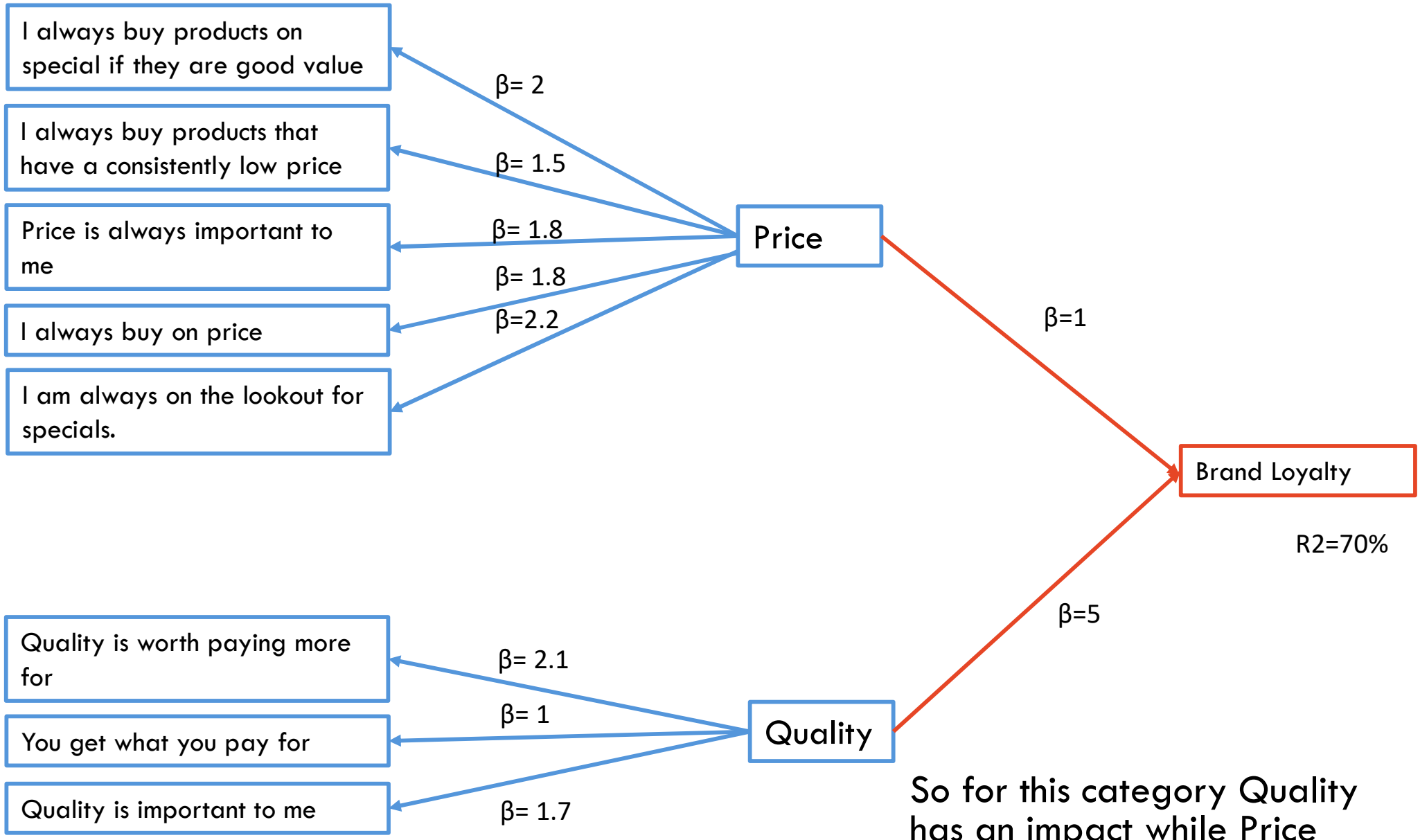
Say we wanted to identify categories which may be at risk.

One way is to test if Price and Quality effect Brand Loyalty.

- Categories where quality is a strong driver of brand loyalty and price is not should be more able to withstand competition from cheap private label brands.
- On the other hand categories where people are open to buying cheaper products and quality is not an issue (as opposed to low quality is acceptable) are more at risk e.g. commodities such as sugar and flour.

To do this we could add Brand Loyalty as a 'response' variable to our Price and Quality CFA. And look at the Beta coefficients for them to understand how they impact Brand Loyalty.

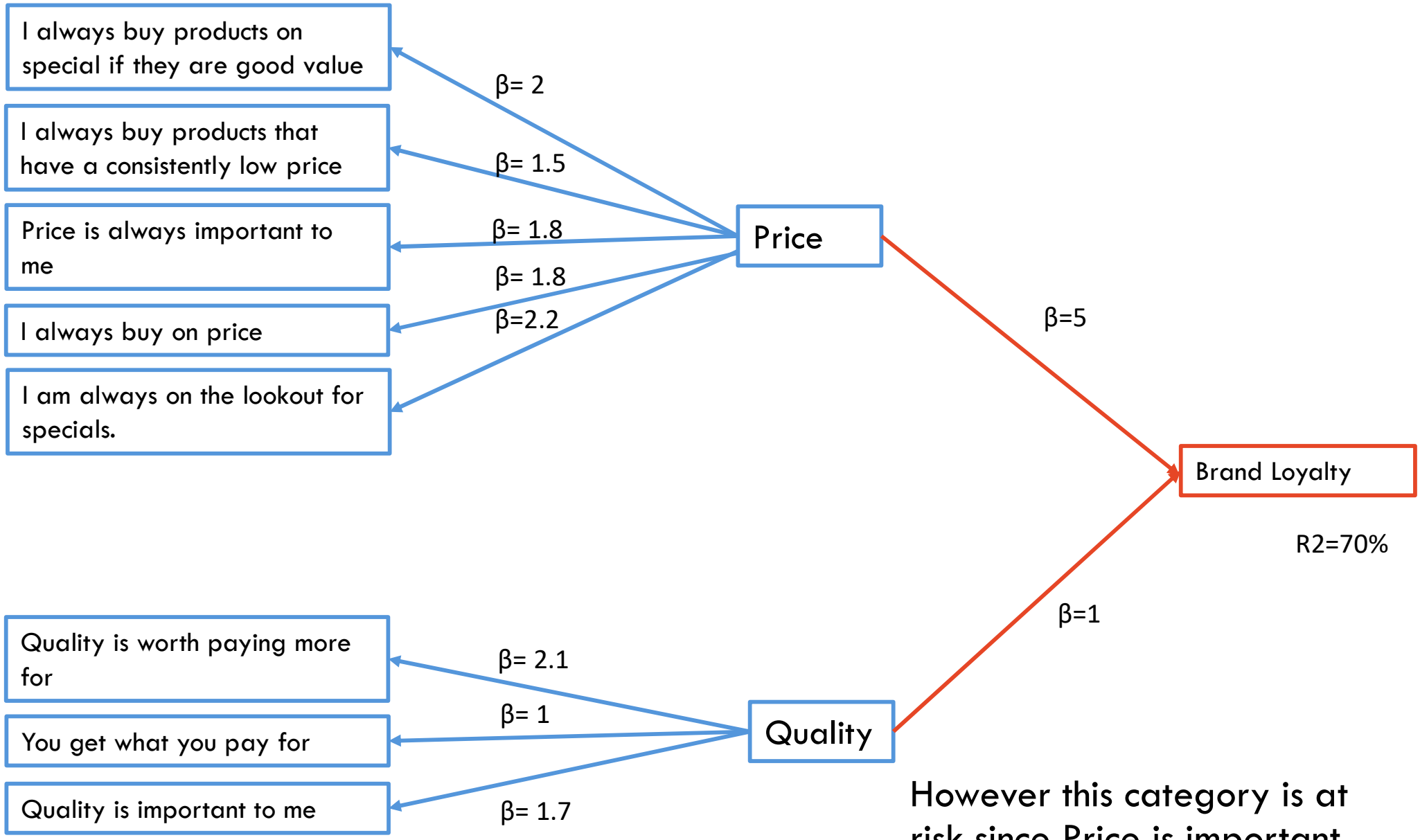
Adding a response to the Path Model



So for this category Quality has an impact while Price had a much smaller one. So it is less at risk

NB: Due to time constraints standard model diagnostics and GoF tests not shown. In practise they are necessary.

Adding a response to the Path Model



However this category is at risk since Price is important but Quality is not.

NB: Due to time constraints standard model diagnostics and GoF tests not shown. In practise they are necessary.

Index Creation

Also known as metrics,

THE UNIVERSITY OF SYDNEY

What is an Index?

Some **function of input metrics i.e. an equation.**

The most **simplest being the average of all input's** e.g. Price/Quality index = average of 8 statements. More complex indices weight the inputs by something.

2 broad categories

1. **Created from *apriori* dimensions measured using Likert style scales**
   - In scope for this workshop
   - Why we do it
     - **Remove survey bias towards dimensions with more statements.** EG: in our example there are 5 price statements and 3 quality statements. If we took a simple average of them then the index is skewed towards Price. But if we first calculate Price and Quality indices which are then averaged this bias is removed.
     - **Reduces the effect of the survey instrument** (also a type of bias). Different surveys having different statements e.g. if 2 different surveys have slightly different statements but both first calculate Price and Quality indices their difference is reduced.
     - **Ensure the index explains as much of the input statements variance as possible.** One of the properties of Factor analysis and PCA is to find a weighted average that explains maximum variance, which is usually more than that explained by a straight unweighted average.

2. **Created from different types of metrics, often on different scales**
   - EG: Consumer Price Index, Body Mass Index, etc
   - Too broad a subject and out of scope of this workshop

Different methods

Simple Averages

Simply take the averages of inputs as defined in the path model e.g. Price is the average of its 5 inputs, Quality the average of its 3 inputs and the Price/Quality index the average of these 2 dimensions.

Weight by Importance

A common improvement is to weight the averages by the **importance** of the statement or dimension. Importance might be stated in the questionnaire, calculated through a driver analysis, or some other method such as $ Market Share.

- Some weight the dimensions by the variance explained. Be careful though as this may just reintroduce the bias caused by # of statements so is rarely useful (since a factor with more statements will often explain more of the overall variance).

Different methods

Confirmatory Factor Analysis (CFA/SEM) and Factor Analysis (FA)

Creates **weighted averages,** with the weights being the **Beta Coefficients (CFA/SEM) or Factor Loadings (EFA/FA).** Has the benefit that these weights are designed to ensure the dimension explains as much of the input variance as possible. One might interpret these weights as a type of importance.

If using CFA follow the steps in the ***Evaluating and Fine tuning Statements used to Quantify Dimensions*** section to create the model.

Using FA instead of CFA is a bit of a hack. It usually involves 1 of 2 methods:

1. **Doing a different FA** on each dimensions statements and (usually/hopefully) using the first factor to represent the dimension. **Advantage** of this method is that you **force the dimensions** you want to be created.
2. **Doing a single FA** on all statements with each factor representing a different dimension. In this case statements with low loadings that are not shown in the factor table are often set to 0, this gives factor scores more aligned to the loadings shown in the factor table. **Disadvantages** of this method are that i) the **required dimensions rarely fall out** from a single FA, and ii) simply setting scores to 0 means all the other factor metrics, loadings and scores aren't quite right. This method is not recommended.

Applying CFA and FA models indices to new data

CFA and FA will often create the factor scores for each respondent as part of the model fitting process. These are the index scores. However how do we calculate the index on new data?

To do this first recognise that the **index is just like a regression equation** where we multiply respondents statement scores by the statements beta coefficients (CFA/SEM) or the loadings (FA) and then sum them.

Some things to look out for:

1. Ensure **appropriate model diagnostics** have been conducted on the models i.e. assumption and Goodness of Fit tests.
2. **Does the input data need to be standardised?**
    1. Sometimes the input data will have been standardised prior to analysis. A common standardisation is to subtract by the mean and divide by the SD i.e. the normal standardisation. If you need to do this for new data remember that you should almost always use the original data's mean and SD, not that from the new data set. If you didn't the same raw scores from different data sets would have different standardised scores and thus factor scores/indices, which rarely makes sense.
3. **Check your index formula** by applying it to the original modelled data and ensure it matches the factors scores that the software generates.

Conjoint and Choice Models

THE UNIVERSITY OF SYDNEY

Example 1: Health Research

A Nutrition researcher wants to understand to understand what types of nutritional claims motivate peoples eating habits.

Rather than just ask people to rank how important various nutritional claims are the researchers decide to create various scenarios and ask them to rate their preference for each one.

Nutritional Factors Researcher Wants to Evaluate

Factors	Levels		
Fat Content	Less than 3 grams	Less than 10% Fat	97% Fat Free
	Less than 5 grams	Less than 5% Fat	
Health Claim	National Heart Foundation Approval	Both	
	School Canteen Approval	Neither	
Source of fibre	Source of fibre		
	No claim		

Full Profile Conjoint Question 1

Q1) Please read the benefits in the below box and tell us how interested you would be in purchasing a product with these claims for yourself or your family, assuming that it was sold where you normally shop, at a reasonable price?

Less than 5% Fat
National Heart Foundation Approved
Source of Fibre

☐	☐	☐	☐	☐
Definitely do not want to buy	Probably do not want to buy	Not sure	Probably want to buy	Definitely want to buy

Full Profile Conjoint Question 2

Q2) Please read the benefits in the below box and tell us how interested you would be in purchasing a product with these claims for yourself or your family, assuming that it was sold where you normally shop, at a reasonable price?

97% Fat Free
Approved for School Canteens
Source of Fibre

☐	☐	☐	☐	☐
Definitely do not want to buy	Probably do not want to buy	Not sure	Probably want to buy	Definitely want to buy

Full Profile Conjoint Question 3

Q3) Please read the benefits in the below box and tell us how interested you would be in purchasing a product with these claims for yourself or your family, assuming that it was sold where you normally shop, at a reasonable price?

97% Fat Free
Approved for School Canteens

☐	☐	☐	☐	☐
Definitely do not want to buy	Probably do not want to buy	Not sure	Probably want to buy	Definitely want to buy

Choice Model Question 1

Q1) Please pick the product you prefer

Less than 5% Fat National Heart Foundation Approved Source of Fibre	97% Fat Free Approved for School Canteens Source of Fibre	97% Fat Free Approved for School Canteens
☐	☐	☐

Choice Model Question 2

Q2) Please pick the product you prefer

Less than 5% Fat Approved for School Canteens Source of Fibre	Less than 5% Fat National Heart Foundation Approved	97% Fat Free Approved for School Canteens Source of Fibre
☐	☐	☐

Full Profile Conjoint vs Choice Models

Full Profile Conjoint

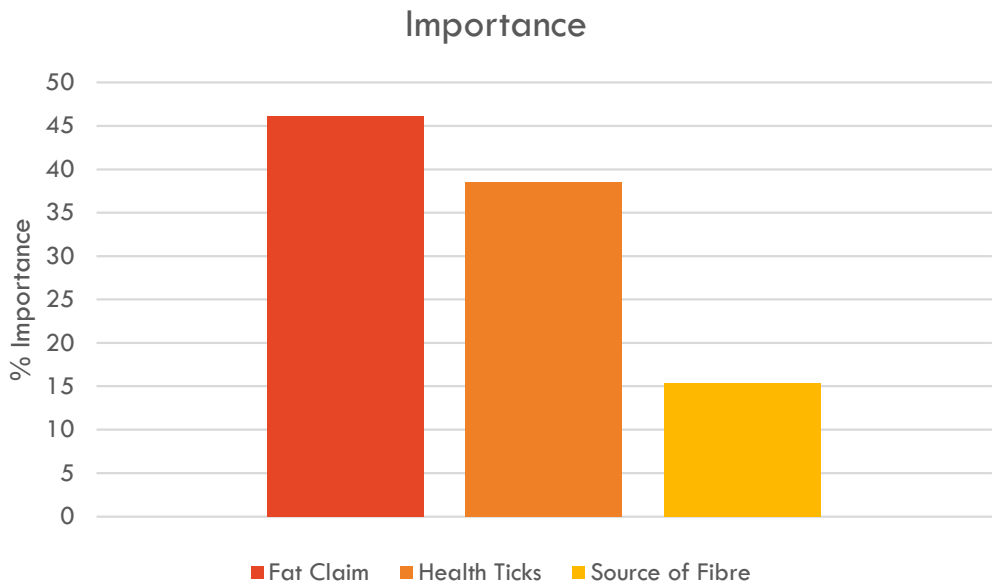
- Ask respondents to ***rate*** a single scenario.
- We get information on each scenario, so more efficient in that respect.
- Usually analysed using some type of regression on a LIKERT scale. Linear Regression is commonly used, ordinal or multinomial are other options.

Choice Models

- Ask respondents to ***select*** which scenario they prefer.
- More realistic since asking people to pick which scenario they prefer, so often preferred.
- Usually analysed using Multinomial regression.

Results: Factor Importance's

Fat claims and health ticks are more important to these respondents than source of fibre claims.



Results: Factor Level impacts on Conjoint PI

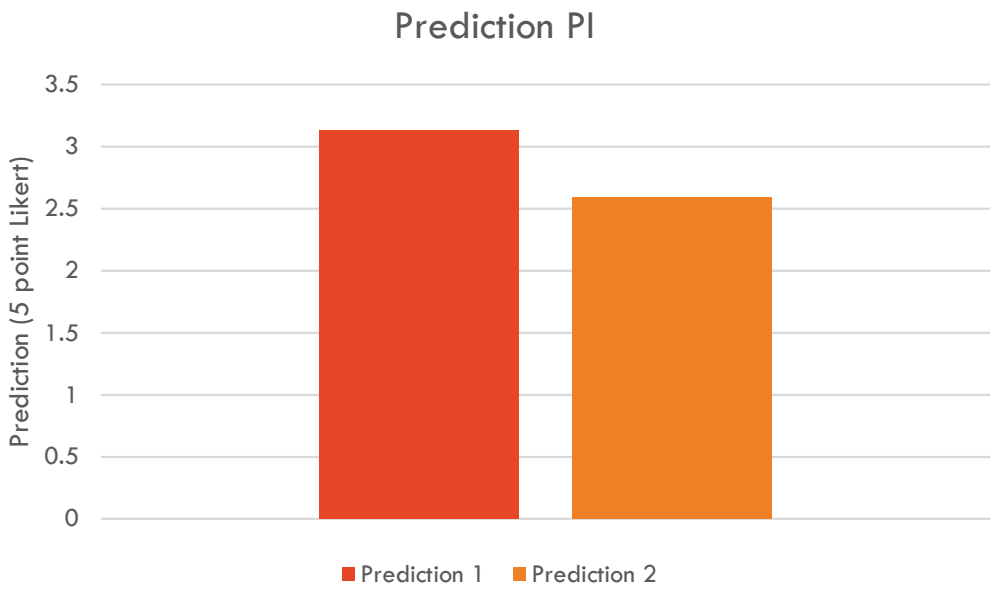
	Total	Heavy Users	Light Users
N	200	75	125
%	100%	38%	63%
Importance (%)			
Fat Claim	46	78	82
Health Ticks	38	16	9
Source of Fibre	15	6	9
Total	100	100	100
Driver Coefficients			
Fat Claim			
absent	-0.01	0.04	-0.05
Less than 3 grams of fat per bar	-0.02	0.09	0.01
97% Fat Free	0.3	0.39	0.30
Less than 10% Fat	-0.3	-0.34	-0.40
Less than 5% Fat	0.01	0.07	0.02
Less than 5 grams of fat per bar	0.02	0.03	0.01
Health Ticks			
absent	-0.3	0.18	0.08
National Heart Foundation Approved	0.05	0.03	0.01
Approved for School Canteens	0.05	0.16	0.07
Both	0.2	0.04	0.03
Source of Fibre			
Present	0.1	0.05	0.02
Absent	-0.1	0.10	0.10
Intercept	3	2.90	3.10

Fitted using linear regression on a 5 point LIKERT scale (scored on 1-5). Meaning the total sample can be interpreted as:

- ***97% fat free*** moves the LIKERT score up 0.3, so is about 3 times more important than ***Source of Fibre*** which only moves it up 0.1.
- A type of ***sensitivity*** analysis

Note how we can redo the analysis for different splits of the data such as Heavy vs Light users.

Results: Predict various Scenarios



Prediction 1	
Intercept	3
Less than 3 grams of fat per bar	-0.02
National Heart Foundation Approved	0.05
Source of Fibre Claim Present	0.1
Sum	3.13

Prediction 2	
Intercept	3
Less than 3 grams of fat per bar	-0.01
National Heart Foundation Approved	-0.3
Source of Fibre Claim Present	-0.1
Sum	2.59

Final Conclusions

- 97% Fat Free tended to be the most motivating of the fat claims and all of the claims presented.

- The best combination of possible claims is 97% Fat free, both Heart Foundation and School Canteen approval and Source of fibre. (Since these have the highest values in the Table)

Synopsis: Conjoint and Choice models

Conjoint and Choice models are from the same family of analysis. Sometimes people use the term conjoint instead of choice model.

Tells us what **drives human behaviour by showing people real life scenarios** with **all factors of interest shown at the same time,** rather than asking about each one in isolation which is what standard ratings questionnaires do.

Meaning they model the underlying **behaviour/*purchase heuristic*** people use to sift through information and decide what to do.

Experimental and Questionnaire Design

Asks respondents to respond to the actual product/task/option in a variety of ***scenarios*** constructed from different combinations of the factors we want to test.

And **then models their answers to tease apart the affect of each factor.**

- As opposed to conventional surveys which would simply ask how important each factor is independently.

The s**cenarios are constructed from different combinations** of the factors the researcher wants to test.

- All the combinations of the different factors usually results in far too many scenarios then we can ask respondents. So a strict Statistical Design is used to:
  - enable us to predict all combinations, even those not rated by respondents
  - efficiently estimate the impact of each factor in an unbiased way with minimum variance.

Benefits over Rating Scales

Gives us a more **realistic assessment** since the factors of interest are presented in a more realistic manor.

The results also tend to show **better differentiation** than metrics commonly used to report ratings. Making the right decision easier to identify as the winners and losers are much clearer.

Evaluates the **combined effect of factors,** rather than the traditional line scale approach which looks at each one in isolation.

We can also model **interactions** between factors.

We can create a **simulators that predict different scenarios** (often done in EXCEL).

Other Benefits

Tests options, scenarios and factors that don't (yet) exist Meaning it is an ***experimental study***, not an observational one. With the corresponding benefits on causal inference.

Closer to reality than other methods

- Since it makes people perform a ***forced trade off*** *between the* different scenarios and their factors. This is closer to what happens in the real world and can give a more accurate understanding of how people ***trade off between the different factors,*** *resulting in* ***better predictions.***
  - A good example of this is when people say 2 factors are both equally important on a 5 point LIKERT scale, but when forced to trade off they consistently ***pick one over the other.***
- Measures **psychological trade-offs** that consumers make when evaluating several attributes together.
- Can uncover **hidden drivers** which may not be apparent to the respondent themselves.
- Measures in a **less 'rational'** way then asking people to rate the factors independently. Which some researchers feel results in more accurate results.

Other uses

Legal Cases: Quantifying Damages

- Apple vs Samsung 2012. Apple used choice models to quantify their damages claim against Samsung for patent infringement. Original case awarded them ***US$1 Billion in damages***! Choice models were used to understand the effect consumers willingness to pay (WTP) computation based on simulating shares of preference for Samsung's devices with and without the alleged patent-infringing technology. A second expert for the plaintiff combined Hauser's WTP estimates with supply-side analysis to arrive at the final claim of damages.

Designing an advertising strategy to reduce road fatalities

For example: could be used to calculate chance of Drink Driving predicted by factors such as 'distance from home', 'back road route available', etc.

Expert Opinion Analysis: Use various experts to find the best Water Sharing Plans in the Murray Darling

In order to better understand what factors are important in Water Sharing Plans one might develop various factors based on the science and community concerns which are then used to create Scenario's. Experts are then used to evaluate it and the results used to tell us which factors are the most important.

Other stockholders can also answer it so we can understand what is important to them.

Business / Market Research Uses

New Product and Concept Ideas

- Maximise sales while minimising cannibalisation
- Maximise profit

New Pack and Claim Ideas

- The above plus:
  - Maximise Return On
  - Investment (ROI)

Prices

That optimise share and revenue

PLUS:

- Measure Brand Equity, in $$$$ and % preference, against other brands (such as private label).

Business / Market Research Uses

Packaging Optimisation

Driver analysis

Source of Share

Volumetrics

ROI

Pricing Optimisation

Product Optimisation

Range Optimisation

NB: Not all studies can do all of these things. Very specific designs are required to achieve some of these outcomes. These simply represent all the possibilities available given the right sample, design, model and simulator is used.

Optimising Price

Price Elasticity Chart

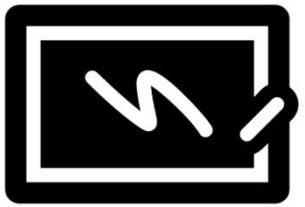
Price is optimised by using Choice models to create price elasticity curves, one for each product.

These are created by setting all products to their RRP and then changing the target products price points. The % preference at each price point is then charted.

The key outcome of price elasticity curves are the shape of the graph, not the % preferences. As the graph is flat between the RRP of $3.90 and $4.29 it means we can increase our RRP by 8%.

It is not an unreasonable assumption that no change in preference means no change in volume. So although choice models don't estimate market share they can be used to make pricing decisions that also consider the impact on volume.

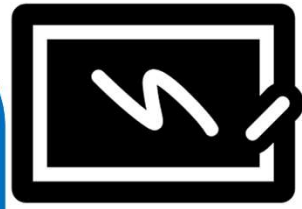
What can choice models do?



Choice models accurately estimate the % of people who prefer a product. This % preference can then be used for:

1. **Price optimisation, via price elasticity curves**
2. **To determine source of share and cannibalisation**
3. **To determine ROI for product improvements, claims and advertising**

Using Choice models for Market Share aka Volume Estimation



Standard choice models tell us what people would prefer to buy in a 'level playing field', not what they actually buy in the real world. We assume a 'level playing field' for these things:

Distribution
Awareness
Marketing
Instore promotions

Most standard market research and even some volume estimation methods also makes these assumptions.

This means that standard Choice models can't be used to predict absolute Volume.

Although there are more complicated volume forecasting choice models that include these market factors and are used to predict volume. However they need to be treated with caution since the market calibration methods often remove much of underlying preference structure.

Using Choice models for Market Share aka Volume Estimation

BUT, just because they don't estimate absolute volume doesn't mean they don't do other things very well.

The majority of market research does not directly predict volume, yet we still use its insights . E.g. Qual, Purchase Intent, Overall Liking.

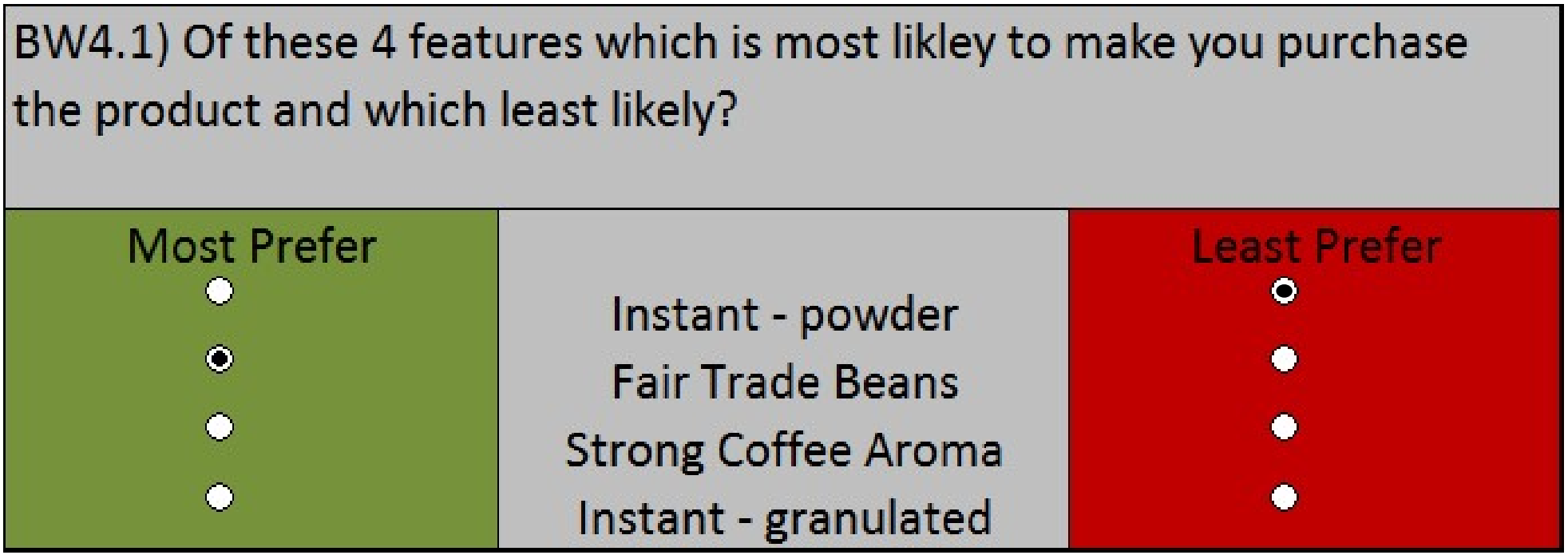
Choice models predict preference, which is a good relative indication of volume. So if a choice model suggests an increase in preference this a good indication that volume will also increase.

As this measure of preference is taken in a competitor context it is the best relative measure of volume used in standard market research. Certainly better than monodic PI.

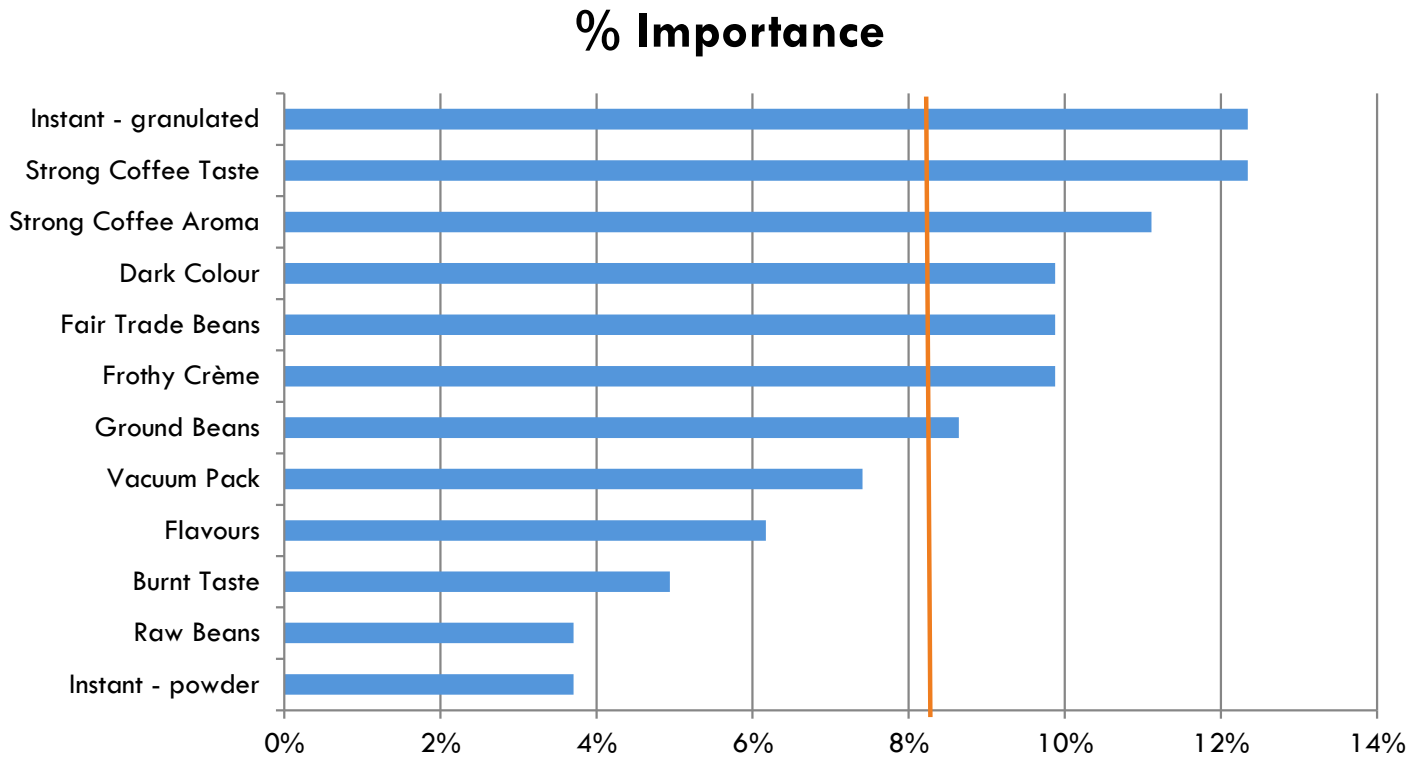
Best Worst (Max-Diff) Models

THE UNIVERSITY OF SYDNEY

Example Question



Example Results



Orange line is the “line of equal importance”, which is a useful way to highlight items over indexing. It is 100%/# items.

Best Worst (Max-Diff)

- Best Worst (also known as Max-Diff) is **closely related to conjoint models** and can be viewed as a simple version of.

- Used to evaluate the **preference of a long list of options** from a single factor.

- As a forced trade off method they share many of the same benefits as Conjoint models over other methods such as rating scales. The key benefit being **increased discrimination** over rating scales.

- **Similar to ranking the options.**
  - However people are good at ranking the first 1-3 and last 1-3 options, with the middle options ranked with a lot of error. Best Worst allows these middle options to be estimated with much higher accuracy.

- They are a **more advanced method of paired comparisons** that give us more information per question by comparing more than 2 options.

When you should use it

If you want the **clear cut, easy to tell story you get through better discrimination** when using rankings or Forced Trade off methods like Choice models.

AND you want the **more detailed information that comes with preferences** (which rating scales give you).

Problem

Gives relative, not absolute preference. This means that although we know the relative preference between the different options we don't know which are actually wanted!

The conventional way to deal with this is to use "Anchored Best Worst". Which anchors the relative measures to a question that explicitly asks if they want it. There are various ways to do this e.g. Dual Response.

Benefits over Ratings

Simple and intuitive meaning (%).

- It's a ***distance rank***.
- Say we have some items with the scores: 40%, 20% & 10%. This tells us:
  - Their rank
  - The item that scored 40% is twice as important as the one that scored 20%., which is in turn twice as important as the one that scored 10%.
  - But that increase in preference is greater (and hence more important) for 40% vs 20% (20% increase) compared to 20% vs 10% (only a 10% increase).

Better discrimination (clearer winners) then ratings due to Forced Tradeoff.

- Makes it much easier to build a story compared to ratings which often have very similar averages. Such small differences between them mean:
  - conclusions aren't convincing and a 'story' is hard to create
  - don't give a researcher a lot of confidence to really focus on certain attributes.

Benefits over Ratings

No scale effects

- Since it uses forced trade off of ***choices instead of scales***. Scale effects prevent direct comparison between groups where such effects exist such as:
  - Nationality e.g. Western vs Eastern
  - Different Collection modes use the scale differently CATI vs online
- Should be used in Global studies where we expect a cultural difference in scale use prevents direct comparison between countries ratings.

Exceptional Segmentations: Usually better than those created from ratings due to more discrimination.

Easier to prevent respondents gaming the answers or using survey bots since goodness of fit metrics will detect those with poor fit. Note that people can also game choice models by using simple algorithms e.g. always pick Brand A.

Benefits over Ranking: Accurately estimates middle ranks

- When ranking more than 5 things the **top and bottom are reliably picked while the middle ones aren't.** However as a Best Worst only asks 3-5 at a time we get around this problem.

- So even if all you want is reliable rankings Best Worst is the best way to get them if you are ranking more than 5 things.

- This means that some people say that if there are less than 5 statements we could just rank them all. And they'd be right, if all that is required is ranking. However Best Worst also has the other benefits as explained here so there may still be a benefit in using Best Worst even if there are 5 or less options.

Benefits over Ranking: Best 1 score summary

- One of the biggest problems with Ranks is that there is **no good single score summary**
  - We can't simply report the % of 1st ranks since the other ranks are vitally important.
    - For example, consider Product Testing. Would you prefer to come 1st for say 20% of people and last the rest of the time, or 2nd all the time? If we only report 1st rankings than we would never know that overall most people place us 2nd.
    - Something that is picked first 25% of the time but last 75% of the time can come out ahead of something that everyone ranks in the top 3.
  - One can use 'average' ranking, which although easily interpretable has its problems too.

- Because Best Worst has design 'connectivity' overlap and uses both the Bests and the Worsts it factors in all the ranks and is considered by some as the best 1 score summary of rank style data.

Benefits over "Pick all that apply" type questions

Tells us what people *prefer*, not just *what* they want.

- Selection methods on the other hand don't tell us the preference between the things selected, they just tell us how many people picked them (this is made worse when they can select more than one thing since it puts them all on the same footing). This means they can't differentiate between things that are very important and determine behavior vs 'nice to have'.
- For example: selection methods can make an option that a lot of people want but with very low preference appear more important than options that options with that fewer people want but rank 1st.

Best Worst also captures peoples 'Worsts'. So say one had 2 things that were equally picked, but one of them was the 'least important' for lots of people. Best Worst would factor this in and give it a lower score, while selection methods would mark them the same.

NB: "Pick all that apply" is a common question type where we list all the options and ask the respondents to pick all that they like i.e. multiple response. The % each are picked is then reported as a form of pseudo ranked results. A variation of these are "pick the most important", "pick the 2 most important" etc.

Benefits over "Pick all that apply" type questions

Allows us to include and benchmark to "Cost of Entry", while still having fine grain reporting

- Since they will ***dominate we often exclude "Cost of Entry" things,*** particularly when selecting from a list.

- However ***it's important to understand how new features compare to "Cost of Entry"*** things since this gives us a benchmark to compare against to see just how important new features are overall. However once we have them benchmarked to Cost of Entry we then want to drill down to the other features to get a better feel for them since this really helps in allocating resources.

- The Magic of Best Worst is that we can **remove statements that are dominating and then recalculate** the remaining statements to understand the relative preference of those that remain. Other methods like "pick all that apply" don't give us that. In other words it lets us look at any combination of statements from those we asked and see their relative preference, as if the others were never even asked!!!

How it works

- We get info from the Bests, the Worsts AND those that are never picked. It's because we use a special design with **connectivity**. This means all statements are connected, even if not shown together. This allows us to infer rank and importance for each statement from how people score all the *other* statements. It's one of the key reasons people use it for large lists, it gives better information than ranking but doesn't require everything to be explicitly ranked.

- An example of connectivity is that we can know something is 2nd even if it is never get picked as either Best or Worst. Consider the simple example where we have 3 items in 1 table. The one not picked as either Best or Worst is ranked 2nd. We know this because the 3 items are *connected*. The design actually extends this connectivity to ALL the tables, so even if something is never picked as Best or Worst we can still give it an accurate rank.

- If we have two choices that are chosen as Best as often as each other then their Worsts would be the tie breaker.

The End! Any Questions?

Survey Platforms

The University provides access to **REDCap, Qualtrics and MS Forms.** These are the preferred platforms, using others may cause researchers to not meet their legal obligations on criteria such as data security and respondent confidentiality. For more info on suitable survey research platforms please review: https://sydneyuni.service-now.com/sm?id=kb_article_view&sysparm_article=KB0019511.

Please contact **Research Data Consulting** for help with **RedCap.** https://redcap.sydney.edu.au/surveys/?s=3W48H9833H

Further Assistance: Sydney University

SIH

- **1on1 Consults** can be requested on our website: www.sydney.edu.au/research/facilities/sydney-informatics-hub.html OR Google "Sydney Informatics Hub" with the "I'm feeling lucky" button
- **Training** Sign up to our mailing list to be notified of upcoming training: https://signup.e2ma.net/signup/1945889/1928048/
  - Research Essentials
  - Experimental Design
  - Power Analysis
- **Online library.** Useful links and the most recent version of all our workshops.
  - https://sydney-informatics-hub.github.io/stats-resources/
- **Hacky Hour** www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/hacky-hour.html OR Google "Sydney Hacky Hour"

OTHER

- **Open Learning Environment (OLE) courses**
- **LinkedIn Learning:** https://linkedin.com/learning/
  - **SPSS** https://www.linkedin.com/learning/machine-learning-ai-foundations-linear-regression/welcome?u=2196204

Further Assistance

WEBSITES

- For conjoint info, Sawtooth Software has a lot of technical papers https://sawtoothsoftware.com/

BOOKS AND PAPERS

- Getting Started with Conjoint Analysis, by Bryan Orme.
- Becoming an Expert in Conjoint Analysis, by Bryan Orme and Keith Chrzan.
- Applied MaxDiff, by Keith Chrzan and Bryan K. Orme.

REDCAP & QAULTRICS TEMPLATES

- Have a lot of validated survey instrument templates on file for your use. Worth looking there before you set up your own!

Acknowledging SIH

All University of Sydney resources are available to Sydney researchers **free of charge.** The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

"The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

For further information about acknowledging the Sydney Informatics Hub, please contact us at **sih.info@sydney.edu.au**.