0.1

Scales of measurement

As the previous section indicates, the outcome of a psychological measurement is called a variable. But not all variables are of the same qualitative type and so it's useful to understand what types there are. A very useful concept for distinguishing between different types of variables is what's known as **scales of measurement**.

0.1.1 Nominal scale

A **nominal scale** variable (also referred to as a **categorical** variable) is one in which there is no particular relationship between the different possibilities. For these kinds of variables it doesn't make any sense to say that one of them is "bigger' or "better" than any other one, and it absolutely doesn't make any sense to average them. The classic example for this is "eye colour". Eyes can be blue, green or brown, amongst other possibilities, but none of them is any "bigger" than any other one. As a result, it would feel really weird to talk about an "average eye colour". Similarly, gender is nominal too: male isn't better or worse than female. Neither does it make sense to try to talk about an "average gender". In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That's it.

Let's take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This "transport type" variable could have quite a few possible values, including: "train", "bus", "car", "bicycle". For now, let's suppose that these four are the only possibilities. Then imagine that I ask 100 people how they got to work today, with this result:

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

So, what's the average transportation type? Obviously, the answer here is that there isn't one. It's a silly question to ask. You can say that travel by car is the most popular method, and travel by train is the least popular method, but that's about all. Similarly, notice that the order in which I

list the options isn't very interesting. I could have chosen to display the data like this…

Transportation	Number of people
(3) Car	48
(1) Train	12
(4) Bicycle	10
(2) Bus	30

…and nothing really changes.

0.1.2 Ordinal scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can't do anything else. The usual example given of an ordinal variable is "finishing position in a race". You *can* say that the person who finished first was faster than the person who finished second, but you *don't* know how much faster. As a consequence we know that 1st > 2nd, and we know that 2nd > 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

Here's a more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change. I then go and ask some people to pick the statement (from four listed statements) that most closely matches their beliefs:

(1) Temperatures are rising because of human activity

(2) Temperatures are rising but we don't know why

(3) Temperatures are rising but not because of humans

(4) Temperatures are not rising

Notice that these four statements actually do have a natural ordering, in terms of "the extent to which they agree with the current science". Statement 1 is a close match, statement 2 is a reasonable match, statement 3 isn't a very good match, and statement 4 is in strong opposition to current science. So, in terms of the thing I'm interested in (the extent to which people endorse the science), I can order the items as $1 > 2 > 3 > 4$. Since this ordering exists, it would be very weird to list the options like this…

(3) Temperatures are rising but not because of humans

(1) Temperatures are rising because of human activity

(4) Temperatures are not rising

(2) Temperatures are rising but we don't know why

…because it seems to violate the natural "structure" to the question.

So, let's suppose I asked 100 people these questions, and got the following answers:

Response	Number
(1) Temperatures are rising because of human activity	51
(2) Temperatures are rising but we don't know why	20
(3) Temperatures are rising but not because of humans	10
(4) Temperatures are not rising	19

When analysing these data it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 out of 100 people were willing to *at least partially* endorse the science. And it's *also* quite reasonable to group (2), (3) and (4) together and say that 49 out of 100 people registered *at least some disagreement* with the dominant scientific view. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 out of 100 people said…what? There's nothing sensible that allows you to group those responses together at all.

That said, notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we *can't* do is average them. For instance, in my simple example here, the "average" response to the question is 1.97. If you can tell me what that means I'd love to know, because it seems like gibberish to me!

0.1.3 Interval scale

In contrast to nominal and ordinal scale variables, **interval scale** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. A good example of an interval scale variable is measuring temperature in degrees celsius. For instance, if it was $15°$ yesterday and $18°$ today, then the $3°$ difference between the two is genuinely meaningful. Moreover, that $3°$ difference is *exactly the same* as the $3°$ difference between $7°$ and $10°$. In short, addition and subtraction are meaningful for interval scale variables.[*1]

[*1]Actually, I've been informed by readers with greater physics knowledge than I that temperature isn't strictly an interval scale, in the sense that the amount of energy required to heat something up by $3°$ depends on it's current temperature. So in the sense that physicists care about, temperature isn't actually an interval scale. But it still makes a cute example so I'm going to ignore this little inconvenient truth.

However, notice that the $0°$ does not mean "no temperature at all". It actually means "the temperature at which water freezes", which is pretty arbitrary. As a consequence it becomes pointless to try to multiply and divide temperatures. It is wrong to say that $20°$ is *twice as hot* as $10°$, just as it is weird and meaningless to try to claim that $20°$ is negative two times as hot as $-10°$.

Again, lets look at a more psychological example. Suppose I'm interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I'm going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely daft for me to divide 2008 by 2003 and say that the second student started "1.0024 times later" than the first one. That doesn't make any sense at all.

0.1.4 Ratio scale

The fourth and final type of variable to consider is a **ratio scale** variable, in which zero really means zero, and it's okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take $3.1 - 2.3 = 0.8$ seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took $3.1/2.3 = 1.35$ times as long as Alan did to answer the question. And the reason why you can do this is that for a ratio scale variable such as RT "zero seconds" really does mean "no time at all".

0.1.5 Continuous versus discrete variables

There's a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables. The difference between these is as follows:

- A **continuous variable** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A **discrete variable** is, in effect, a variable that isn't continuous. For a discrete variable it's sometimes the case that there's nothing in the middle.

These definitions probably seem a bit abstract, but they're pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then Cameron's response time will lie in between if he took

Table1   The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

. . . .

3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond, meaning that his RT would lie in between Cameron's and Alan's. And while in practice it might be impossible to measure RT that precisely, it's certainly possible in principle. Because we can always find a new value for RT in between any two other ones we regard RT as a continuous measure.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete. There isn't a type of transportation that falls "in between" trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete. Although "2nd place" does fall between "1st place" and "3rd place", there's nothing that can logically fall in between "1st place" and "2nd place". Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete. Since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10. Table 1.1 summarises the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I'm trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like "discrete variable" when they mean "nominal scale variable". It's very unfortunate.

0.1.6   Some complexities

Okay, I know you're going to be shocked to hear this, but the real world is much messier than this little classification scheme suggests. Very few variables in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were

hard and fast rules. It doesn't work like that. They're guidelines, intended to help you think about the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the **Likert scale**. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands, of them and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

> Which of the following best describes your opinion of the statement that "all pirates are freaking awesome"?

and then the options presented to the participant are these:

(1) Strongly disagree
(2) Disagree
(3) Neither agree nor disagree
(4) Agree
(5) Strongly agree

This set of items is an example of a 5-point Likert scale, in which people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items are explicitly described. This is a perfectly good example of a 5-point Likert scale too:

(1) Strongly disagree
(2)
(3)
(4)
(5) Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On

the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as interval scale.[*2] It's not interval scale, but in practice it's close enough that we usually think of it as being **quasi-interval scale**.

[*2]Ah, psychology …never an easy answer to anything!

1. 研究デザインについての短い導入

> 実験が終わった後に統計家にアドバイスを求められることは，ほとんどの場合，死後の
> 検査を求めるようなものである。統計家は実験がなぜ死んだかをいうだけだろう。
> – Sir Ronald Fisher[1]

この章では，研究をデザインし，データを集め，そのデータが機能するかどうかチェックする，といったことについての基本的なアイデアについて考えることから始めよう。あなた自身が研究をデザインするのに十分な情報生えられないかもしれないが，ほかの人がやる研究を査定するのに必要な基本的ツールの多くは手に入れられると思う。しかし，この本の狙いはデータの収集よりもデータの分析にあるので，ごくあっさりと全体像を見るにとどめる。この章は，ふたつの意味で「特別」である。まず，ほかの章よりも心理学に特有の状況を扱っている。次に，研究の方法論についての科学的問題にかなりの重点を置いていて，データ分析の統計的問題にはそれほど重点をおいていない。とはいえ，このふたつの問題はお互いに関係しているから，この問題をちょっとばかり詳しく扱うのが，統計のテキストの伝統的なやり方になっている。この章は研究デザインについては **Campbell1963** に，測定の尺度に関する議論については **Stevens1946** に依っている。

1.1

心理学的測定への導入

データ収集を理解するにはまず，測定について考えなければならない。つまり，ここで我々がやろうとしていることは，人の行動や心についての何かを測るということだ。では「測る」とは何だろう？

1.1.1 心理学的測定についてのいくつかの考え方

測定そのものは微妙な概念だが，基本的には「なにか」に数字やラベル，あるいはほかのよく定義された記述を割り当てる方法を見つけること，と言える。であるから，次に挙げるようなものも心理学的測定の中に含まれる。

- 私の **年齢** は *33* 歳です。
- 私は**アンチョビが** *好きではない* 。
- 私の **染色体的な性別** は *男性である。*
- 私の **性自認** は *男性である*[*2]

上の短いリストには**太字になっているところ**が "測定されようとしている対象" であり，*イタリックになっているところ* が "測定されたもの" である。じつは，それぞれのケースで生じうる測定全体について考えることで，少し拡張することができる。

- 私の**年齢**は (年単位で数えるから)*0, 1, 2, 3 …,* となる。上限がどうなるかはちょっと曖昧だが，現実的には最大でも *150* ぐらいだと思っていれば良い。それ以上生きた人はいないのだから。
- *アンチョビが好きかどうかは*，答え方として*好き*あるいは*嫌い*，はたまたどちらでもないとか，ときどきね，と答えるだろう。
- 私の**染色体的な性別**は，ほぼ間違いなく*男性 (XY)* か，*女性 (XX)* だろうが，ごくわずかな例外もある。それはクラインフェルター症候群 *(XXY)* というやつで，ほとんど男性の染色体とおなじである。そのほかにもこうした可能性はあるかもしれない。
- 私の**性自認**は*男性*か*女性*のいずれかだが，私の染色体的な性別とは一致していない。私は自分をどちらでもないとするかもしれないし，はっきりと*トランスジェンダーだ*，と答えるかもしれない。

このように，ある事象 (年齢とか) は取りうる値が明らかのものもあれば，ちょっと特殊なものも

[*2]うむ... ちょっと引っかかるだろ？ このセクションはこの本の最も古いパートだから，ちょっと時代遅れで小っ恥ずかしいものになっている。これが書かれているのは 2010 年で，ここに書いてある事実は*確かに正しい*。これを今の 2018 年にも取って考えると，私はもう 33 歳じゃないし，驚くことじゃない。私の染色体が変わったとは思えないし，私の遺伝子は今も XY 型だとおもう。一方，性自認は... ううん。タイトルページには，今私のことを Daniel じゃなくて Danielle と書いているからわかってもらえるかと思うけど (訳註；Daniel は男性名的表記，Danielle は女性名的表記)，最近は性別に関するアンケートでは男性と答えないようにしていて，むしろ "she/her" で呼ばれる方が嬉しくなっている。話すと長くなるけど！実は，この本ではこのことをどう扱うか少し考えた。この本では作者の声が直接含まれているので，この本での表記を全部Danielle にすると随分違ったものになるんじゃないかと思う。でもそれはかなりの作業量になるから，この本での私の名前は"Dan" で通そうと思う。とにかく Danielle のニックネームとして Dan はバッチリだと思わない？ ちっちゃいことは気にしない。私のことをどう呼んだら良いかわからない読者のために，少しでも気楽になってもらえればと思ってこれを書いてる。あ，アンチョビはまだ好きじゃないけどね (笑)。

ある。しかし，誰かの年齢と言ってもこれよりちょっと微妙なものであることは指摘しておきたい。たとえば，上の例では年齢を一年単位で測定することを想定していた。しかしもしあなたが発達心理学者であったら，これはちょっと雑すぎてて，年齢は*年と月 (月齢)* で表現するかもしれない (もし子供が 2 歳 11 ヶ月であれば，これを "2.11" と書いたりする)。新生児に興味があれば，年齢は生まれてから*何日め*かで表現するだろうし，もしかすると*生後何時間経った*かで表現するかもしれない。言い換えると，あなたが測定値に何を許容するかというのが重要なのだ。

　これをもう少し詳しく見てみると，"年齢" という概念は実際まったく正確ではないことに気づかれるかもしれない。一般的に言って，我々が "年齢" というときは，暗に "生まれてからの時間の長さ" を意味している。しかし，それが常に正しいわけではない。たとえば，新生児の目の動きをどのようにコントロールしているかに興味があったとしよう。もしそれぐらい若い子に興味があれば，あなたは "生まれたとき" から始めることが意味のある点だとは思わなくなるかもしれない。アリスが生後 3 週間で，ビアンカが 1 週間後に生まれてきたとして，彼女らに "生まれてから 2 時間後" に出会った時彼女らが "同じ年齢" だったといえるだろうか？ 社会的には，日常生活で年齢について語る時に，生まれた時を起点とするが，それは世界に生まれたって独立した存在として自らを扱ってきた人としての時間の総量，として定義しているからだ。科学的な観点からは，興味があるのはそこだけではない。生物としての人間を考えるなら，概念的には意識を持った生き物として成長してきた過程を考える方が便利だし，その観点から行くとアリスとビアンカは同じ年齢だとは言えない。だからあなたが "年齢" という概念を定義しようとするなら，ふたつのやり方があることになる。意識の時間的長さ，生まれてからの時間的長さ，である。大人を対象にしているときは，その差はほとんど意味をなさないが，新生児の場合はそうではないのだ。

　この問題を超えた，方法論的な問いがある。どの "測定法" が，誰かの年齢を見出すのに用いられるべきだろうか。上で述べたように，そこにはいくつかの可能性がある。

- 単に "あなたは何歳ですか" と尋ねる。この事故報告式は早くて，やすくて，簡単だ。しかしその人が質問を十分理解している場合に限るし，人によっては歳を誤魔化すことがある。
- 権威者 (例えば親御さん) に "あなたのお子さんは何歳ですか？" と尋ねる。この方法は早くて，子供を扱うときには簡単な方法だ。親はいつも周りにいるのだから。でも "意識を持ち始めてから" を考えるときはうまくいかない。というのも，親はその子がいつ意識を持ち始めたかを知る由がないからだ。そのときは，別の権威者 (例えば産科医) に頼る必要があるだろう。
- 公的な記録を調べる。例えば出生証明書とか，死亡証明書である。これは時間はかかるし，手間もかかるが，それなりの利点がある (例えば故人を対象にする場合は。)

1.1.2 操作可能にする: あなたの測定を定義する

上の章で論じだれていたことはすべて，**操作化**という概念に関係している。もう少しこの概念を正確にいうと，操作化とは大きな意味があるが，何かしらぼんやりした概念を，正確な測定に落とし込む手続きのことを指す。この操作化のプロセスは，以下のいくつかの異なる要素を含んでいる。

- あなたが測定しようとしているものを明確にすること。例えば "年齢" も "生まれてからの時間" なのか，"意識が芽生えてからの時間" なのか，どちらがあなたの研究の文脈に沿うものなのか？
- それを測定するのにどの手法にするのか決定すること。事故報告式の測定年齢を使うのか，親に聞くのか，公的な記録を取るのか，どれだろう？ もし事故報告式を使うのなら，どういうフレーズでその質問をするだろう？
- 測定が取りうる値のセットを決定すること。この値は常に数量的なものとは限らないことに注意しよう。測定する年齢は数量的かもしれないが，どんな値が許されているのかをよく考えて見てほしい。年単位の年齢なのか，年月なのか，日，時間，どれだろう？ ほかの測定の種類によっては (例えばジェンダー)，値は数量的なものにならない。しかしすでに述べたように，私たちはどんな値が許されているかを考える必要があるのだ。もし事故報告式でジェンダーを尋ねたら，私たちは相手にどんな選択肢を許すだろうか。単に "男性" か "女性" だけで十分なのか？ "その他" のオプションは必要ないのか？ あるいは，特別な選択肢を考えて，回答者自身の言葉を使うことを許すのか？ そしてもしすべての言語報告を可能な値のセットに含めるのであれば，その回答をどのように解釈するのか？

操作化はトリッキーなビジネスで，"唯一の正解" があるものではない。"年齢" や "性別" のインフォーマルな概念を操作化する方法を選択し，それをフォーマルな測定にすることを選んでも，それはあなたが何を測定したいかに依存するのだ。大抵の場合，あなたの領域で活躍している人たちの科学的コミュニティーが，どうしたらいいかについて十分に確立されたアイデアをすでに持っていることと思う。いいかえると，操作化はケースバイケースで考える必要があると言うことだ。とはいえ，個々人の研究プロジェクトにはそれぞれ特別な問題が他にもいっぱいあるが，一般的な側面もあるものだ。

次に進む前に，用語を整理しておきたい。相互に関連する次の四つのことがある。

- **理論的構成概念**。これはあなたが測定しようとしているもの，例えば "年齢"，"性別"，"ジェンダー" や "意見" といったものだ。理論的な構成概念は直接観測されないし，実際にはちょっと曖昧なものであることが多い。
- **測定**。測定はあなたが客観的にしようとする際につかうツールや手法のこと。調査における質問，行動観察や脳のスキャンなど，すべて測定に数えられる。
- **操作化**。"操作化" という用語は，測定と構成概念の論理的なつながりを指す。あるいは，理

論的構成概念から測定を導き出そうとするプロセスのことである。

- **変数**。最後に，新しい用語を。変数とは，世界に存在する何かに対して，私たちが測定を適用した時の結果である。つまり，変数は実際の “データ” で，最終的に我々のデータセットになるものである。

実際には，科学者でさえこれらの擁護間の区別を曖昧にしようとする傾向があるが，その違いを理解しようとすることはとても有益である。

1.2
測定の尺度水準

前のセクションで示したように，心理学的な測定の結果は変数と呼ばれる。しかしすべての変数が同じ質的なタイプではないので，どういうタイプがあるのかを知っておく必要がある。変数のタイプの違いを区別する大変便利な概念として**測定の尺度水準**というものがある。

1.2.1　名義尺度水準

名義尺度水準の変数は (**カテゴリカルな** (質的な) 変数と言われることもあるが)，異なる取りうる可能性との間に特定の関係がないもののことである。この種の変数は，他の値 “より大きい” とか “より良い” といったいかなる意味も持たないし，その平均を出しても全くなんの意味もない。典型的な例として，“目の色” を考えてみよう。目はブルー，グリーン，ブラウンだったり他の色であったりするが，それのどれかが他のものよりも “大きい” とは言わない。だから，“平均的な目の色” について語るのも憚られる。同様に，ジェンダーもそうだろう。男性が女性よりも良いとか悪いとかいうものではない。同じく “平均的な性別” というのも意味がない。つまり，名義尺度水準の変数は，異なる変数の値は異なるということしか言えない。それだけだ。

もう少し注意深く見てみよう。私が人の通勤・通学手段について研究しているとしよう。私が測定しようとしている変数は，仕事に行くためにどういう種類の移動手段を使うかである。“移動のタイプ” 変数は，それほど多い値を取るものではない。せいぜい “電車”，“バス”，“自家用車”，“自転車” ぐらいだろう。ここでは，この四つの可能性しかないことにしよう。その上で，100 人の人に対して

移動手段	人数
(1) 電車	12
(2) バス	30
(3) 自家用車	48
(4) 自転車	10

どうやって仕事場に行ったかを調べて，次のような結果になったとする。

　さて，移動手段の平均はなんだろう？ 明らかにその答えはひとつではない。愚問としか言いようがない。車で移動することが最もポピュラーな方法で，電車で移動するのがもっともポピュラーでない，それだけである。同時に，私が選択肢を列挙した順番が面白くないことにも注意してほしい。データは次のように表示することだってできるのだ。

移動手段	人数
(3) 自家用車	48
(1) 電車	12
(4) 自転車	10
(2) バス	30

…実質的には何も変わってない。

1.2.2　順序尺度水準

　順序尺度水準にある変数は，名義尺度水準の変数より少し構造化されているが，それほどで大きな変化ではない。順序尺度水準の変数は，異なる状態に自然な意味のある順序を与えたものであるが，それ以上のことはない。よく使われる例としては，"レースの最終順位" である。最初にゴールした人に，2 位の人よりも早かったということはできるが，どれぐらい速かったかということはできない。結果的に，1 位 >2 位ということはわかるし，1 位 >2 位ということもわかるが，1 位と 2 位の差が 2 位と 3 位の差よりも大きいかもしれない。心理学的に面白い例もある。私が気候変動に対する人々の考え方に興味をもっているとしよう。私は，その人の信念にもっとも当てはまる文章がどれかを，(以下にリストした 4 つの文章から) ピックアップして人に尋ねる。

> (1) 気温は人間の活動のせいで上昇している
> (2) 気温は上昇しているがなぜかはわからない
> (3) 気温は上昇しているが人間のせいではない
> (4) 気温は上昇していない

　この 4 つの文章は，確かに自然な序列がついている。"現在の科学に対してどれほど同意するか" という意味でだ。文章 1 は明らかに科学的営みにマッチしているし，2 はかなり適合している。3 はそれほどでもないし，4 は現在の科学と逆の立場だ。だから，私の興味のある用語 (人は科学を支持

する度合い) でいうと，項目を 1 > 2 > 3 > 4 の順に並べることができる。この順序づけがある以上，次のように選択肢を並べると奇妙なことになる。

(3) 気温は上昇しているが人間のせいではない
(1) 気温は人間の活動のせいで上昇している
(4) 気温は上昇していない (2) 気温は上昇しているがなぜかはわからない

...これは項目の自然な "構造" に違反しているように見えるからだ。では，100 人にこの質問をして，以下のような答えを得たとしよう。

反応	度数
(1) 気温は人間の活動のせいで上昇している	51
(2) 気温は上昇しているがなぜかはわからない	20
(3) 気温は上昇しているが人間のせいではない	10
(4) 気温は上昇していない	19

このデータを分析するときは，グループ (1),(2),(3) を一緒にするのは合理的だと思われるし，100 人のうち 81 人が少なくとも一部は現在の科学を支持していると言えそうだ。また，グループ (2),(3),(4) をまとめることもあり得る話で，100 人のうち 49 人が現在の科学的観点に，少なくともいくらか同意していないといえる。しかし，(1),(2) と (4) を一緒にすることはおかしなことで，100 人のうち 90 人がどうこうしてるとは言えない。これらの反応を一緒くたにする合理的な根拠がないからだ。

　つまり，私たちがこれらの項目が，意味のあるグルーピングで自然に序列づけできるということはできるが，その平均を求めるということはできない。例えば，ここでの例でいうと，"平均的な" 反応は 1.97 である。もしその意味がわかるというのなら，ぜひ知りたいのでどうぞ教えてください！

1.2.3　間隔尺度水準

　名義や順序の尺度水準に比べると，間隔尺度水準の変数や比率尺度水準の変数は，かなり意味を持った数値である。感覚尺度水準の変数は，数字の間の差分が解釈可能だが，"自然な" ゼロの値をもつものではない。間隔尺度水準の良い例は，セ氏で測られる気温である。例えば，昨日は 15° で今日は 18° だというとき，この 3° の違いに意味がある。もっというと，この 3° の違いは，7° と 10° の間の 3° と全く同じである。つまり，間隔尺度水準では足し算引き算に意味があるのだ[*3]。

[*3]実は，かなり物理学の知識がある読者に教わったのだが，この温度は厳密な間隔尺度水準ではない。というのも，3° 温度を上げようとするのに必要なエネルギーの量は，その時の気温に依存するらしい。だ k ら，物理学的に注意していうなら，温度は間隔尺度とはいえないのだ。しかし例としていいものなので，このちょっとした不都合な真実は無視しようと思う。

しかし注意してほしいのは，0° が "温度がない" ことを意味するものではない点だ。これが意味するのは "水が凍る温度" であって，これはちょっと恣意的なものだ。だから，温度を掛け算したり割り算したりしようとするのはポイントがずれていることになる。20° は 10° の二倍熱いとは言えないし，20° が −10° のマイナス 2 倍というのも意味がないことだ。

あらためて心理学的な例を見てみよう。大学一年生の態度が時間とともにどう変わっていくかに興味があったとしよう。もちろん，各学生が入学した年を記録するだろう。これは間隔尺度水準の変数だ。2003 年に学生になった人は，2008 に始めた学生よりも 5 年早い。しかし，2008 を 2003 で割って，後者の学生を前者の学生に比べて "1.0024 時間後の人" とは言わないだろう。そんなの全く意味がないじゃないか。

1.2.4 Ratio scale

The fourth and final type of variable to consider is a **ratio scale** variable, in which zero really means zero, and it's okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take $3.1 - 2.3 = 0.8$ seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took $3.1/2.3 = 1.35$ times as long as Alan did to answer the question. And the reason why you can do this is that for a ratio scale variable such as RT "zero seconds" really does mean "no time at all".

1.2.5 Continuous versus discrete variables

There's a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables. The difference between these is as follows:

- A **continuous variable** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A **discrete variable** is, in effect, a variable that isn't continuous. For a discrete variable it's sometimes the case that there's nothing in the middle.

These definitions probably seem a bit abstract, but they're pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then Cameron's response time will lie in between if he took 3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond,

Table1.1   The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

. . . .

meaning that his RT would lie in between Cameron's and Alan's. And while in practice it might be impossible to measure RT that precisely, it's certainly possible in principle. Because we can always find a new value for RT in between any two other ones we regard RT as a continuous measure.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete. There isn't a type of transportation that falls "in between" trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete. Although "2nd place" does fall between "1st place" and "3rd place", there's nothing that can logically fall in between "1st place" and "2nd place". Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete. Since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10. Table 1.1 summarises the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I'm trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like "discrete variable" when they mean "nominal scale variable". It's very unfortunate.

1.2.6   Some complexities

Okay, I know you're going to be shocked to hear this, but the real world is much messier than this little classification scheme suggests. Very few variables in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were hard and fast rules. It doesn't work like that. They're guidelines, intended to help you think about

the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the **Likert scale**. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands, of them and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

> Which of the following best describes your opinion of the statement that "all pirates are freaking awesome"?

and then the options presented to the participant are these:

(1) Strongly disagree
(2) Disagree
(3) Neither agree nor disagree
(4) Agree
(5) Strongly agree

This set of items is an example of a 5-point Likert scale, in which people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items are explicitly described. This is a perfectly good example of a 5-point Likert scale too:

(1) Strongly disagree
(2)
(3)
(4)
(5) Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5"

part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as interval scale.[*4] It's not interval scale, but in practice it's close enough that we usually think of it as being **quasi-interval scale**.

1.3

Assessing the reliability of a measurement

At this point we've thought a little bit about how to operationalise a theoretical construct and thereby create a psychological measure. And we've seen that by applying psychological measures we end up with variables, which can come in many different types. At this point, we should start discussing the obvious question: is the measurement any good? We'll do this in terms of two related ideas: *reliability* and *validity*. Put simply, the **reliability** of a measure tells you how *precisely* you are measuring something, whereas the validity of a measure tells you how *accurate* the measure is. In this section I'll talk about reliability; we'll talk about validity in section 1.6.

Reliability is actually a very simple concept. It refers to the repeatability or consistency of your measurement. The measurement of my weight by means of a "bathroom scale" is very reliable. If I step on and off the scales over and over again, it'll keep giving me the same answer. Measuring my intelligence by means of "asking my mum" is very unreliable. Some days she tells me I'm a bit thick, and other days she tells me I'm a complete idiot. Notice that this concept of reliability is different to the question of whether the measurements are correct (the correctness of a measurement relates to it's validity). If I'm holding a sack of potatos when I step on and off the bathroom scales the measurement will still be reliable: it will always give me the same answer. However, this highly reliable answer doesn't match up to my true weight at all, therefore it's wrong. In technical terms, this is a *reliable but invalid* measurement. Similarly, whilst my mum's estimate of my intelligence is a bit unreliable, she might be right. Maybe I'm just not too bright, and so while her estimate of my intelligence fluctuates pretty wildly from day to day, it's basically right. That would be an *unreliable but valid* measure. Of course, if my mum's estimates are too unreliable it's going to be very hard to figure out which one of her many claims about my intelligence is actually the right one. To some extent, then, a very unreliable measure tends to end up being invalid for practical purposes; so much so that many people would say that reliability is necessary (but not sufficient) to ensure validity.

Okay, now that we're clear on the distinction between reliability and validity, let's have a think

---

[*4]Ah, psychology …never an easy answer to anything!

about the different ways in which we might measure reliability:

- **Test-retest reliability**. This relates to consistency over time. If we repeat the measurement at a later date do we get a the same answer?
- **Inter-rater reliability**. This relates to consistency across people. If someone else repeats the measurement (e.g., someone else rates my intelligence) will they produce the same answer?
- **Parallel forms reliability**. This relates to consistency across theoretically-equivalent measurements. If I use a different set of bathroom scales to measure my weight does it give the same answer?
- **Internal consistency reliability**. If a measurement is constructed from lots of different parts that perform similar functions (e.g., a personality questionnaire result is added up across several questions) do the individual parts tend to give similar answers.

Not all measurements need to possess all forms of reliability. For instance, educational assessment can be thought of as a form of measurement. One of the subjects that I teach, *Computational Cognitive Science*, has an assessment structure that has a research component and an exam component (plus other things). The exam component is *intended* to measure something different from the research component, so the assessment as a whole has low internal consistency. However, within the exam there are several questions that are intended to (approximately) measure the same things, and those tend to produce similar outcomes. So the exam on its own has a fairly high internal consistency. Which is as it should be. You should only demand reliability in those situations where you want to be measuring the same thing!

1.4 ———

The "role" of variables: predictors and outcomes

I've got one last piece of terminology that I need to explain to you before moving away from variables. Normally, when we do some research we end up with lots of different variables. Then, when we analyse our data, we usually try to explain some of the variables in terms of some of the other variables. It's important to keep the two roles "thing doing the explaining" and "thing being explained" distinct. So let's be clear about this now. First, we might as well get used to the idea of using mathematical symbols to describe variables, since it's going to happen over and over again. Let's denote the "to be explained" variable Y, and denote the variables "doing the explaining" as X_1, X_2, etc.

Table1.2   The terminology used to distinguish between different roles that a variable can play when analysing a data set. Note that this book will tend to avoid the classical terminology in favour of the newer names.

role of the variable	classical name	modern name
"to be explained"	dependent variable (DV)	outcome
"to do the explaining"	independent variable (IV)	predictor

. . . .

When we are doing an analysis we have different names for X and Y, since they play different roles in the analysis. The classical names for these roles are **independent variable** (IV) and **dependent variable** (DV). The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). The logic behind these names goes like this: if there really is a relationship between X and Y then we can say that Y depends on X, and if we have designed our study "properly" then X isn't dependent on anything else. However, I personally find those names horrible. They're hard to remember and they're highly misleading because (a) the IV is never actually "independent of everything else", and (b) if there's no relationship then the DV doesn't actually depend on the IV. And in fact, because I'm not the only person who thinks that IV and DV are just awful names, there are a number of alternatives that I find more appealing. The terms that I'll use in this book are **predictors** and **outcomes**. The idea here is that what you're trying to do is use X (the predictors) to make guesses about Y (the outcomes).[*5] This is summarised in Table 1.2.

1.5

Experimental and non-experimental research

One of the big distinctions that you should be aware of is the distinction between "experimental research" and "non-experimental research". When we make this distinction, what we're really talking about is the degree of control that the researcher exercises over the people and events in the study.

[*5]Annoyingly though, there's a lot of different names used out there. I won't list all of them — there would be no point in doing that — other than to note that "response variable" is sometimes used where I've used "outcome". Sigh. This sort of terminological confusion is very common, I'm afraid.

1.5.1 Experimental research

The key feature of **experimental research** is that the researcher controls all aspects of the study, especially what participants experience during the study. In particular, the researcher manipulates or varies the predictor variables (IVs) but allows the outcome variable (DV) to vary naturally. The idea here is to deliberately vary the predictors (IVs) to see if they have any causal effects on the outcomes. Moreover, in order to ensure that there's no possibility that something other than the predictor variables is causing the outcomes, everything else is kept constant or is in some other way "balanced", to ensure that they have no effect on the results. In practice, it's almost impossible to *think* of everything else that might have an influence on the outcome of an experiment, much less keep it constant. The standard solution to this is **randomisation**. That is, we randomly assign people to different groups, and then give each group a different treatment (i.e., assign them different values of the predictor variables). We'll talk more about randomisation later, but for now it's enough to say that what randomisation does is minimise (but not eliminate) the possibility that there are any systematic difference between groups.

Let's consider a very simple, completely unrealistic and grossly unethical example. Suppose you wanted to find out if smoking causes lung cancer. One way to do this would be to find people who smoke and people who don't smoke and look to see if smokers have a higher rate of lung cancer. This is *not* a proper experiment, since the researcher doesn't have a lot of control over who is and isn't a smoker. And this really matters. For instance, it might be that people who choose to smoke cigarettes also tend to have poor diets, or maybe they tend to work in asbestos mines, or whatever. The point here is that the groups (smokers and non-smokers) actually differ on lots of things, not *just* smoking. So it might be that the higher incidence of lung cancer among smokers is caused by something else, and not by smoking per se. In technical terms these other things (e.g. diet) are called "confounders", and we'll talk about those in just a moment.

In the meantime, let's consider what a proper experiment might look like. Recall that our concern was that smokers and non-smokers might differ in lots of ways. The solution, as long as you have no ethics, is to *control* who smokes and who doesn't. Specifically, if we randomly divide young non-smokers into two groups and force half of them to become smokers, then it's very unlikely that the groups will differ in any respect other than the fact that half of them smoke. That way, if our smoking group gets cancer at a higher rate than the non-smoking group, we can feel pretty confident that (a) smoking does cause cancer and (b) we're murderers.

Non-experimental research

Non-experimental research is a broad term that covers "any study in which the researcher doesn't have as much control as they do in an experiment". Obviously, control is something that scientists like to have, but as the previous example illustrates there are lots of situations in which you can't or shouldn't try to obtain that control. Since it's grossly unethical (and almost certainly criminal) to force people to smoke in order to find out if they get cancer, this is a good example of a situation in which you really shouldn't try to obtain experimental control. But there are other reasons too. Even leaving aside the ethical issues, our "smoking experiment" does have a few other issues. For instance, when I suggested that we "force" half of the people to become smokers, I was talking about *starting* with a sample of non-smokers, and then forcing them to become smokers. While this sounds like the kind of solid, evil experimental design that a mad scientist would love, it might not be a very sound way of investigating the effect in the real world. For instance, suppose that smoking only causes lung cancer when people have poor diets, and suppose also that people who normally smoke do tend to have poor diets. However, since the "smokers" in our experiment aren't "natural" smokers (i.e., we forced non-smokers to become smokers, but they didn't take on all of the other normal, real life characteristics that smokers might tend to possess) they probably have better diets. As such, in this silly example they wouldn't get lung cancer and our experiment will fail, because it violates the structure of the "natural" world (the technical name for this is an "artefactual" result).

One distinction worth making between two types of non-experimental research is the difference between **quasi-experimental research** and **case studies**. The example I discussed earlier, in which we wanted to examine incidence of lung cancer among smokers and non-smokers without trying to control who smokes and who doesn't, is a quasi-experimental design. That is, it's the same as an experiment but we don't control the predictors (IVs). We can still use statistics to analyse the results, but we have to be a lot more careful and circumspect.

The alternative approach, case studies, aims to provide a very detailed description of one or a few instances. In general, you can't use statistics to analyse the results of case studies and it's usually very hard to draw any general conclusions about "people in general" from a few isolated examples. However, case studies are very useful in some situations. Firstly, there are situations where you don't have any alternative. Neuropsychology has this issue a lot. Sometimes, you just can't find a lot of people with brain damage in a specific brain area, so the only thing you can do is describe those cases that you do have in as much detail and with as much care as you can. However, there's also some genuine advantages to case studies. Because you don't have as many people to study you have the ability to invest lots of time and effort trying to understand the specific factors at play in

each case. This is a very valuable thing to do. As a consequence, case studies can complement the more statistically-oriented approaches that you see in experimental and quasi-experimental designs. We won't talk much about case studies in this book, but they are nevertheless very valuable tools!

1.6

Assessing the validity of a study

More than any other thing, a scientist wants their research to be "valid". The conceptual idea behind **validity** is very simple. Can you trust the results of your study? If not, the study is invalid. However, whilst it's easy to state, in practice it's much harder to check validity than it is to check reliability. And in all honesty, there's no precise, clearly agreed upon notion of what validity actually is. In fact, there are lots of different kinds of validity, each of which raises it's own issues. And not all forms of validity are relevant to all studies. I'm going to talk about five different types of validity:

- Internal validity
- External validity
- Construct validity
- Face validity
- Ecological validity

First, a quick guide as to what matters here. (1) Internal and external validity are the most important, since they tie directly to the fundamental question of whether your study really works. (2) Construct validity asks whether you're measuring what you think you are. (3) Face validity isn't terribly important except insofar as you care about "appearances". (4) Ecological validity is a special case of face validity that corresponds to a kind of appearance that you might care about a lot.

1.6.1 Internal validity

Internal validity refers to the extent to which you are able draw the correct conclusions about the causal relationships between variables. It's called "internal" because it refers to the relationships between things "inside" the study. Let's illustrate the concept with a simple example. Suppose you're interested in finding out whether a university education makes you write better. To do so, you get a group of first year students, ask them to write a 1000 word essay, and count the number of spelling and grammatical errors they make. Then you find some third-year students, who obviously have had more of a university education than the first-years, and repeat the exercise. And let's suppose it turns out that the third-year students produce fewer errors. And so you conclude that a university

education improves writing skills. Right? Except that the big problem with this experiment is that the third-year students are older and they've had more experience with writing things. So it's hard to know for sure what the causal relationship is. Do older people write better? Or people who have had more writing experience? Or people who have had more education? Which of the above is the true *cause* of the superior performance of the third-years? Age? Experience? Education? You can't tell. This is an example of a failure of internal validity, because your study doesn't properly tease apart the *causal* relationships between the different variables.

1.6.2 External validity

External validity relates to the **generalisability** or **applicability** of your findings. That is, to what extent do you expect to see the same pattern of results in "real life" as you saw in your study. To put it a bit more precisely, any study that you do in psychology will involve a fairly specific set of questions or tasks, will occur in a specific environment, and will involve participants that are drawn from a particular subgroup (disappointingly often it is college students!). So, if it turns out that the results don't actually generalise or apply to people and situations beyond the ones that you studied, then what you've got is a lack of external validity.

The classic example of this issue is the fact that a very large proportion of studies in psychology will use undergraduate psychology students as the participants. Obviously, however, the researchers don't care *only* about psychology students. They care about people in general. Given that, a study that uses only psychology students as participants always carries a risk of lacking external validity. That is, if there's something "special" about psychology students that makes them different to the general population in some *relevant* respect, then we may start worrying about a lack of external validity.

That said, it is absolutely critical to realise that a study that uses only psychology students does not necessarily have a problem with external validity. I'll talk about this again later, but it's such a common mistake that I'm going to mention it here. The external validity of a study is threatened by the choice of population if (a) the population from which you sample your participants is very narrow (e.g., psychology students), and (b) the narrow population that you sampled from is systematically different from the general population *in some respect that is relevant to the psychological phenomenon that you intend to study*. The italicised part is the bit that lots of people forget. It is true that psychology undergraduates differ from the general population in lots of ways, and so a study that uses only psychology students *may* have problems with external validity. However, if those differences aren't very relevant to the phenomenon that you're studying, then there's nothing to worry about. To make this a bit more concrete here are two extreme examples:

- You want to measure "attitudes of the general public towards psychotherapy", but all of your participants are psychology students. This study would almost certainly have a problem with external validity.
- You want to measure the effectiveness of a visual illusion, and your participants are all psychology students. This study is unlikely to have a problem with external validity

Having just spent the last couple of paragraphs focusing on the choice of participants, since that's a big issue that everyone tends to worry most about, it's worth remembering that external validity is a broader concept. The following are also examples of things that might pose a threat to external validity, depending on what kind of study you're doing:

- People might answer a "psychology questionnaire" in a manner that doesn't reflect what they would do in real life.
- Your lab experiment on (say) "human learning" has a different structure to the learning problems people face in real life.

1.6.3 Construct validity

Construct validity is basically a question of whether you're measuring what you want to be measuring. A measurement has good construct validity if it is actually measuring the correct theoretical construct, and bad construct validity if it doesn't. To give a very simple (if ridiculous) example, suppose I'm trying to investigate the rates with which university students cheat on their exams. And the way I attempt to measure it is by asking the cheating students to stand up in the lecture theatre so that I can count them. When I do this with a class of 300 students 0 people claim to be cheaters. So I therefore conclude that the proportion of cheaters in my class is 0%. Clearly this is a bit ridiculous. But the point here is not that this is a very deep methodological example, but rather to explain what construct validity is. The problem with my measure is that while I'm *trying* to measure "the proportion of people who cheat" what I'm actually measuring is "the proportion of people stupid enough to own up to cheating, or bloody minded enough to pretend that they do". Obviously, these aren't the same thing! So my study has gone wrong, because my measurement has very poor construct validity.

1.6.4 Face validity

Face validity simply refers to whether or not a measure "looks like" it's doing what it's supposed to, nothing more. If I design a test of intelligence, and people look at it and they say "no, that test

doesn't measure intelligence", then the measure lacks face validity. It's as simple as that. Obviously, face validity isn't very important from a pure scientific perspective. After all, what we care about is whether or not the measure *actually* does what it's supposed to do, not whether it *looks like* it does what it's supposed to do. As a consequence, we generally don't care very much about face validity. That said, the concept of face validity serves three useful pragmatic purposes:

- Sometimes, an experienced scientist will have a "hunch" that a particular measure won't work. While these sorts of hunches have no strict evidentiary value, it's often worth paying attention to them. Because often times people have knowledge that they can't quite verbalise, so there might be something to worry about even if you can't quite say why. In other words, when someone you trust criticises the face validity of your study, it's worth taking the time to think more carefully about your design to see if you can think of reasons why it might go awry. Mind you, if you don't find any reason for concern, then you should probably not worry. After all, face validity really doesn't matter very much.

- Often (very often), completely uninformed people will also have a "hunch" that your research is crap. And they'll criticise it on the internet or something. On close inspection you may notice that these criticisms are actually focused entirely on how the study "looks", but not on anything deeper. The concept of face validity is useful for gently explaining to people that they need to substantiate their arguments further.

- Expanding on the last point, if the beliefs of untrained people are critical (e.g., this is often the case for applied research where you actually want to convince policy makers of something or other) then you *have* to care about face validity. Simply because, whether you like it or not, a lot of people will use face validity as a proxy for real validity. If you want the government to change a law on scientific psychological grounds, then it won't matter how good your studies "really" are. If they lack face validity you'll find that politicians ignore you. Of course, it's somewhat unfair that policy often depends more on appearance than fact, but that's how things go.

1.6.5 Ecological validity

Ecological validity is a different notion of validity, which is similar to external validity, but less important. The idea is that, in order to be ecologically valid, the entire set up of the study should closely approximate the real world scenario that is being investigated. In a sense, ecological validity is a kind of face validity. It relates mostly to whether the study "looks" right, but with a bit more rigour to it. To be ecologically valid the study has to look right in a fairly specific way. The idea

behind it is the intuition that a study that is ecologically valid is more likely to be externally valid. It's no guarantee, of course. But the nice thing about ecological validity is that it's much easier to check whether a study is ecologically valid than it is to check whether a study is externally valid. A simple example would be eyewitness identification studies. Most of these studies tend to be done in a university setting, often with a fairly simple array of faces to look at, rather than a line up. The length of time between seeing the "criminal" and being asked to identify the suspect in the "line up" is usually shorter. The "crime" isn't real so there's no chance of the witness being scared, and there are no police officers present so there's not as much chance of feeling pressured. These things all mean that the study *definitely* lacks ecological validity. They might (but might not) mean that it also lacks external validity.

1.7

Confounds, artefacts and other threats to validity

If we look at the issue of validity in the most general fashion the two biggest worries that we have are *confounders* and *artefacts*. These two terms are defined in the following way:

- **Confounder**: A confounder is an additional, often unmeasured variable[*6] that turns out to be related to both the predictors and the outcome. The existence of confounders threatens the internal validity of the study because you can't tell whether the predictor causes the outcome, or if the confounding variable causes it.
- **Artefact**: A result is said to be "artefactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artefact describes a threat to your external validity, because it raises the possibility that you can't generalise or apply your results to the actual population that you care about.

As a general rule confounders are a bigger concern for non-experimental studies, precisely because they're not proper experiments. By definition, you're leaving lots of things uncontrolled, so there's a lot of scope for confounders being present in your study. Experimental research tends to be much less vulnerable to confounders. The more control you have over what happens during the study, the more you can prevent confounders from affecting the results. With random allocation, for example,

---

[*6] The reason why I say that it's unmeasured is that if you *have* measured it, then you can use some fancy statistical tricks to deal with the confounder. Because of the existence of these statistical solutions to the problem of confounders, we often refer to a confounder that we have measured and dealt with as a *covariate*. Dealing with covariates is a more advanced topic, but I thought I'd mention it in passing since it's kind of comforting to at least know that this stuff exists.

confounders are distributed randomly, and evenly, between different groups.

However, there are always swings and roundabouts and when we start thinking about artefacts rather than confounders the shoe is very firmly on the other foot. For the most part, artefactual results tend to be a concern for experimental studies than for non-experimental studies. To see this, it helps to realise that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behaviour in a more naturalistic context. By working in a more real-world context you lose experimental control (making yourself vulnerable to confounders), but because you tend to be studying human psychology "in the wild" you reduce the chances of getting an artefactual result. Or, to put it another way, when you take psychology out of the wild and bring it into the lab (which we usually have to do to gain our experimental control), you always run the risk of accidentally studying something different to what you wanted to study.

Be warned though. The above is a rough guide only. It's absolutely possible to have confounders in an experiment, and to get artefactual results with non-experimental studies. This can happen for all sorts of reasons, not least of which is experimenter or researcher error. In practice, it's really hard to think everything through ahead of time and even very good researchers make mistakes.

Although there's a sense in which almost any threat to validity can be characterised as a confounder or an artefact, they're pretty vague concepts. So let's have a look at some of the most common examples.

1.7.1 History effects

History effects refer to the possibility that specific events may occur during the study that might influence the outcome measure. For instance, something might happen in between a pre-test and a post-test. Or in-between testing participant 23 and participant 24. Alternatively, it might be that you're looking at a paper from an older study that was perfectly valid for its time, but the world has changed enough since then that the conclusions are no longer trustworthy. Examples of things that would count as history effects are:

- You're interested in how people think about risk and uncertainty. You started your data collection in December 2010. But finding participants and collecting data takes time, so you're still finding new people in February 2011. Unfortunately for you (and even more unfortunately for others), the Queensland floods occurred in January 2011 causing billions of dollars of damage and killing many people. Not surprisingly, the people tested in February 2011 express quite different beliefs about handling risk than the people tested in December 2010. Which (if any) of these reflects the "true" beliefs of participants? I think the answer is

probably both. The Queensland floods genuinely changed the beliefs of the Australian public, though possibly only temporarily. The key thing here is that the "history" of the people tested in February is quite different to people tested in December.

- You're testing the psychological effects of a new anti-anxiety drug. So what you do is measure anxiety before administering the drug (e.g., by self-report, and taking physiological measures). Then you administer the drug, and afterwards you take the same measures. In the middle however, because your lab is in Los Angeles, there's an earthquake which increases the anxiety of the participants.

1.7.2 Maturation effects

As with history effects, **maturational effects** are fundamentally about change over time. However, maturation effects aren't in response to specific events. Rather, they relate to how people change on their own over time. We get older, we get tired, we get bored, etc. Some examples of maturation effects are:

- When doing developmental psychology research you need to be aware that children grow up quite rapidly. So, suppose that you want to find out whether some educational trick helps with vocabulary size among 3 year olds. One thing that you need to be aware of is that the vocabulary size of children that age is growing at an incredible rate (multiple words per day) all on its own. If you design your study without taking this maturational effect into account, then you won't be able to tell if your educational trick works.
- When running a very long experiment in the lab (say, something that goes for 3 hours) it's very likely that people will begin to get bored and tired, and that this maturational effect will cause performance to decline regardless of anything else going on in the experiment

1.7.3 Repeated testing effects

An important type of history effect is the effect of **repeated testing**. Suppose I want to take two measurements of some psychological construct (e.g., anxiety). One thing I might be worried about is if the first measurement has an effect on the second measurement. In other words, this is a history effect in which the "event" that influences the second measurement is the first measurement itself! This is not at all uncommon. Examples of this include:

- *Learning and practice*: e.g., "intelligence" at time 2 might appear to go up relative to

time 1 because participants learned the general rules of how to solve "intelligence-test-style" questions during the first testing session.

- *Familiarity with the testing situation*: e.g., if people are nervous at time 1, this might make performance go down. But after sitting through the first testing situation they might calm down a lot precisely because they've seen what the testing looks like.

- *Auxiliary changes caused by testing*: e.g., if a questionnaire assessing mood is boring then mood rating at measurement time 2 is more likely to be "bored" precisely because of the boring measurement made at time 1.

1.7.4 Selection bias

Selection bias is a pretty broad term. Suppose that you're running an experiment with two groups of participants where each group gets a different "treatment", and you want to see if the different treatments lead to different outcomes. However, suppose that, despite your best efforts, you've ended up with a gender imbalance across groups (say, group A has 80% females and group B has 50% females). It might sound like this could never happen but, trust me, it can. This is an example of a selection bias, in which the people "selected into" the two groups have different characteristics. If any of those characteristics turns out to be relevant (say, your treatment works better on females than males) then you're in a lot of trouble.

1.7.5 Differential attrition

When thinking about the effects of attrition, it is sometimes helpful to distinguish between two different types. The first is **homogeneous attrition**, in which the attrition effect is the same for all groups, treatments or conditions. In the example I gave above, the attrition would be homogeneous if (and only if) the easily bored participants are dropping out of all of the conditions in my experiment at about the same rate. In general, the main effect of homogeneous attrition is likely to be that it makes your sample unrepresentative. As such, the biggest worry that you'll have is that the generalisability of the results decreases. In other words, you lose external validity.

The second type of attrition is **heterogeneous attrition**, in which the attrition effect is different for different groups. More often called **differential attrition**, this is a kind of selection bias that is caused by the study itself. Suppose that, for the first time ever in the history of psychology, I manage to find the perfectly balanced and representative sample of people. I start running "Dani's incredibly long and tedious experiment" on my perfect sample but then, because my study is incredibly long and tedious, lots of people start dropping out. I can't stop this. Participants absolutely have the

right to stop doing any experiment, any time, for whatever reason they feel like, and as researchers we are morally (and professionally) obliged to remind people that they do have this right. So, suppose that "Dani's incredibly long and tedious experiment" has a very high drop out rate. What do you suppose the odds are that this drop out is random? Answer: zero. Almost certainly the people who remain are more conscientious, more tolerant of boredom, etc., than those that leave. To the extent that (say) conscientiousness is relevant to the psychological phenomenon that I care about, this attrition can decrease the validity of my results.

Here's another example. Suppose I design my experiment with two conditions. In the "treatment" condition, the experimenter insults the participant and then gives them a questionnaire designed to measure obedience. In the "control" condition, the experimenter engages in a bit of pointless chitchat and then gives them the questionnaire. Leaving aside the questionable scientific merits and dubious ethics of such a study, let's have a think about what might go wrong here. As a general rule, when someone insults me to my face I tend to get much less co-operative. So, there's a pretty good chance that a lot more people are going to drop out of the treatment condition than the control condition. And this drop out isn't going to be random. The people most likely to drop out would probably be the people who don't care all that much about the importance of obediently sitting through the experiment. Since the most bloody minded and disobedient people all left the treatment group but not the control group, we've introduced a confound: the people who actually took the questionnaire in the treatment group were *already* more likely to be dutiful and obedient than the people in the control group. In short, in this study insulting people doesn't make them more obedient. It makes the more disobedient people leave the experiment! The internal validity of this experiment is completely shot.

1.7.6 Non-response bias

Non-response bias is closely related to selection bias and to differential attrition. The simplest version of the problem goes like this. You mail out a survey to 1000 people but only 300 of them reply. The 300 people who replied are almost certainly not a random subsample. People who respond to surveys are systematically different to people who don't. This introduces a problem when trying to generalise from those 300 people who replied to the population at large, since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every question. If (say) 80 people chose not to answer one of your questions, does this introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal; probably the pages just fell off. However, if the

question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of **missing data**. If the data that is missing was "lost" randomly, then it's not a big problem. If it's missing systematically, then it can be a big problem.

1.7.7 Regression to the mean

Regression to the mean refers to any situation where you select data based on an extreme value on some measure. Because the variable has natural variation it almost certainly means that when you take a subsequent measurement the later measurement will be less extreme than the first one, purely by chance.

Here's an example. Suppose I'm interested in whether a psychology education has an adverse effect on very smart kids. To do this, I find the 20 psychology I students with the best high school grades and look at how well they're doing at university. It turns out that they're doing a lot better than average, but they're not topping the class at university even though they did top their classes at high school. What's going on? The natural first thought is that this must mean that the psychology classes must be having an adverse effect on those students. However, while that might very well be the explanation, it's more likely that what you're seeing is an example of "regression to the mean". To see how it works, let's take a moment to think about what is required to get the best mark in a class, regardless of whether that class be at high school or at university. When you've got a big class there are going to be *lots* of very smart people enrolled. To get the best mark you have to be very smart, work very hard, and be a bit lucky. The exam has to ask just the right questions for your idiosyncratic skills, and you have to avoid making any dumb mistakes (we all do that sometimes) when answering them. And that's the thing, whilst intelligence and hard work are transferable from one class to the next, luck isn't. The people who got lucky in high school won't be the same as the people who get lucky at university. That's the very definition of "luck". The consequence of this is that when you select people at the very extreme values of one measurement (the top 20 students), you're selecting for hard work, skill and luck. But because the luck doesn't transfer to the second measurement (only the skill and work), these people will all be expected to drop a little bit when you measure them a second time (at university). So their scores fall back a little bit, back towards everyone else. This is regression to the mean.

Regression to the mean is surprisingly common. For instance, if two very tall people have kids their children will tend to be taller than average but not as tall as the parents. The reverse happens with very short parents. Two very short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than

from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement people tended to do worse, but after the negative reinforcement they tended to do better. But notice that there's a selection bias here! When people do very well, you're selecting for "high" values, and so you should *expect*, because of regression to the mean, that performance on the next trial should be worse regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artefact caused by regression to the mean (**Kahneman1973**).

1.7.8 Experimenter bias

Experimenter bias can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behaviour" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not, for example the right answer to the questions being asked or knowledge of the expected pattern of performance for the condition that the participant is in. The classic example of this happening is the case study of "Clever Hans", which dates back to 1907 (**Pfungst1911**; **Hothersall2004**). Clever Hans was a horse that apparently was able to read and count and perform other human like feats of intelligence. After Clever Hans became famous, psychologists started examining his behaviour more closely. It turned out that, not surprisingly, Hans didn't know how to do maths. Rather, Hans was responding to the human observers around him, because the humans did know how to count and the horse had learned to change its behaviour when people changed theirs.

The general solution to the problem of experimenter bias is to engage in double blind studies, where neither the experimenter nor the participant knows which condition the participant is in or knows what the desired behaviour is. This provides a very good solution to the problem, but it's important to recognise that it's not quite ideal, and hard to pull off perfectly. For instance, the obvious way that I could try to construct a double blind study is to have one of my Ph.D. students (one who doesn't know anything about the experiment) run the study. That feels like it should be enough. The only person (me) who knows all the details (e.g., correct answers to the questions, assignments of participants to conditions) has no interaction with the participants, and the person who does all the talking to people (the Ph.D. student) doesn't know anything. Except for the reality that the last part is very unlikely to be true. In order for the Ph.D. student to run the study effectively they need to have been briefed by me, the researcher. And, as it happens, the Ph.D. student also knows me and knows a bit about my general beliefs about people and psychology (e.g., I tend to think humans are much smarter than psychologists give them credit for). As a result of all this,

it's almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect. Suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there's a thing called the "Pygmalion effect", where if you expect great things of people they'll tend to rise to the occasion. But if you expect them to fail then they'll do that too. In other words, the expectations become a self-fulfilling prophesy.

1.7.9 Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter's knowledge or desires for the experiment are communicated to the participants, and that these can change people's behaviour (**Rosenthal1966**). However, even if you manage to stop this from happening, it's almost impossible to stop people from knowing that they're part of a psychological study. And the mere fact of knowing that someone is watching or studying you can have a pretty big effect on behaviour. This is generally referred to as **reactivity** or **demand effects**. The basic idea is captured by the Hawthorne effect: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a study that took place in the "Hawthorne Works" factory outside of Chicago (**Adair1984**). This study, from the 1920s, looked at the effects of factory lighting on worker productivity. But, importantly, change in worker behaviour occurred because the workers *knew* they were being studied, rather than any effect of factory lighting.

To get a bit more specific about some of the ways in which the mere fact of being in a study can change how people behave, it helps to think like a social psychologist and look at some of the *roles* that people might *adopt* during an experiment but might *not adopt* if the corresponding events were occurring in the real world:

- The *good participant* tries to be too helpful to the researcher. He or she seeks to figure out the experimenter's hypotheses and confirm them.
- The *negative participant* does the exact opposite of the good participant. He or she seeks to break or destroy the study or the hypothesis in some way.
- The *faithful participant* is unnaturally obedient. He or she seeks to follow instructions perfectly, regardless of what might have happened in a more realistic setting.
- The *apprehensive participant* gets nervous about being tested or studied, so much so that his or her behaviour becomes highly unnatural, or overly socially desirable.

1.7.10 Placebo effects

The **placebo effect** is a specific type of demand effect that we worry a lot about. It refers to the situation where the mere fact of being treated causes an improvement in outcomes. The classic example comes from clinical trials. If you give people a completely chemically inert drug and tell them that it's a cure for a disease, they will tend to get better faster than people who aren't treated at all. In other words, it is people's belief that they are being treated that causes the improved outcomes, not the drug.

However, the current consensus in medicine is that true placebo effects are quite rare and most of what was previously considered placebo effect is in fact some combination of natural healing (some people just get better on their own), regression to the mean and other quirks of study design. Of interest to psychology is that the strongest evidence for at least some placebo effect is in self-reported outcomes, most notably in treatment of pain (**hrobjartsson2010**).

1.7.11 Situation, measurement and sub-population effects

In some respects, these terms are a catch-all term for "all other threats to external validity". They refer to the fact that the choice of sub-population from which you draw your participants, the location, timing and manner in which you run your study (including who collects the data) and the tools that you use to make your measurements might all be influencing the results. Specifically, the worry is that these things might be influencing the results in such a way that the results won't generalise to a wider array of people, places and measures.

1.7.12 Fraud, deception and self-deception

> It is difficult to get a man to understand something, when his salary depends on his not understanding it.
> – Upton Sinclair

There's one final thing I feel I should mention. While reading what the textbooks often have to say about assessing the validity of a study I couldn't help but notice that they seem to make the assumption that the researcher is honest. I find this hilarious. While the vast majority of scientists are honest, in my experience at least, some are not.[*7] Not only that, as I mentioned earlier, scientists

---

[*7]Some people might argue that if you're not honest then you're not a real scientist. Which does have some truth to it I guess, but that's disingenuous (look up the "No true Scotsman" fallacy). The fact is that there are lots of people who are employed ostensibly as scientists, and whose work has all of the trappings of science, but who are outright fraudulent. Pretending that they don't exist by saying that they're not scientists is just muddled thinking.

are not immune to belief bias. It's easy for a researcher to end up deceiving themselves into believing the wrong thing, and this can lead them to conduct subtly flawed research and then hide those flaws when they write it up. So you need to consider not only the (probably unlikely) possibility of outright fraud, but also the (probably quite common) possibility that the research is unintentionally "slanted". I opened a few standard textbooks and didn't find much of a discussion of this problem, so here's my own attempt to list a few ways in which these issues can arise:

- **Data fabrication**. Sometimes, people just make up the data. This is occasionally done with "good" intentions. For instance, the researcher believes that the fabricated data do reflect the truth, and may actually reflect "slightly cleaned up" versions of actual data. On other occasions, the fraud is deliberate and malicious. Some high-profile examples where data fabrication has been alleged or shown include Cyril Burt (a psychologist who is thought to have fabricated some of his data), Andrew Wakefield (who has been accused of fabricating his data connecting the MMR vaccine to autism) and Hwang Woo-suk (who falsified a lot of his data on stem cell research).

- **Hoaxes**. Hoaxes share a lot of similarities with data fabrication, but they differ in the intended purpose. A hoax is often a joke, and many of them are intended to be (eventually) discovered. Often, the point of a hoax is to discredit someone or some field. There's quite a few well known scientific hoaxes that have occurred over the years (e.g., Piltdown man) and some were deliberate attempts to discredit particular fields of research (e.g., the Sokal affair).

- **Data misrepresentation**. While fraud gets most of the headlines, it's much more common in my experience to see data being misrepresented. When I say this I'm not referring to newspapers getting it wrong (which they do, almost always). I'm referring to the fact that often the data don't actually say what the researchers think they say. My guess is that, almost always, this isn't the result of deliberate dishonesty but instead is due to a lack of sophistication in the data analyses. For instance, think back to the example of Simpson's paradox that I discussed in the beginning of this book. It's very common to see people present "aggregated" data of some kind and sometimes, when you dig deeper and find the raw data yourself you find that the aggregated data tell a different story to the disaggregated data. Alternatively, you might find that some aspect of the data is being hidden, because it tells an inconvenient story (e.g., the researcher might choose not to refer to a particular variable). There's a lot of variants on this, many of which are very hard to detect.

- **Study "misdesign"**. Okay, this one is subtle. Basically, the issue here is that a researcher designs a study that has built-in flaws and those flaws are never reported in the paper. The data that are reported are completely real and are correctly analysed, but they are produced by a study that is actually quite wrongly put together. The researcher really wants to find a

particular effect and so the study is set up in such a way as to make it "easy" to (artefactually) observe that effect. One sneaky way to do this, in case you're feeling like dabbling in a bit of fraud yourself, is to design an experiment in which it's obvious to the participants what they're "supposed" to be doing, and then let reactivity work its magic for you. If you want you can add all the trappings of double blind experimentation but it won't make a difference since the study materials themselves are subtly telling people what you want them to do. When you write up the results the fraud won't be obvious to the reader. What's obvious to the participant when they're in the experimental context isn't always obvious to the person reading the paper. Of course, the way I've described this makes it sound like it's always fraud. Probably there are cases where this is done deliberately, but in my experience the bigger concern has been with unintentional misdesign. The researcher *believes* and so the study just happens to end up with a built in flaw, and that flaw then magically erases itself when the study is written up for publication.

- **Data mining & post hoc hypothesising**. Another way in which the authors of a study can more or less misrepresent the data is by engaging in what's referred to as "data mining" (**Gelman2014**). As we'll discuss later, if you keep trying to analyse your data in lots of different ways, you'll eventually find something that "looks" like a real effect but isn't. This is referred to as "data mining". It used to be quite rare because data analysis used to take weeks, but now that everyone has very powerful statistical software on their computers it's becoming very common. Data mining per se isn't "wrong", but the more that you do it the bigger the risk you're taking. The thing that is wrong, and I suspect is very common, is *unacknowledged* data mining. That is, the researcher runs every possible analysis known to humanity, finds the one that works, and then pretends that this was the only analysis that they ever conducted. Worse yet, they often "invent" a hypothesis after looking at the data to cover up the data mining. To be clear. It's not wrong to change your beliefs after looking at the data, and to reanalyse your data using your new "post hoc" hypotheses. What is wrong (and I suspect common) is failing to acknowledge that you've done. If you acknowledge that you did it then other researchers are able to take your behaviour into account. If you don't, then they can't. And that makes your behaviour deceptive. Bad!

- **Publication bias & self-censoring**. Finally, a pervasive bias is "non-reporting" of negative results. This is almost impossible to prevent. Journals don't publish every article that is submitted to them. They prefer to publish articles that find "something". So, if 20 people run an experiment looking at whether reading *Finnegans Wake* causes insanity in humans, and 19 of them find that it doesn't, which one do you think is going to get published? Obviously,

it's the one study that did find that *Finnegans Wake* causes insanity.[*8] This is an example of a *publication bias*. Since no-one ever published the 19 studies that didn't find an effect, a naive reader would never know that they existed. Worse yet, most researchers "internalise" this bias and end up *self-censoring* their research. Knowing that negative results aren't going to be accepted for publication, they never even try to report them. As a friend of mine says "for every experiment that you get published, you also have 10 failures". And she's right. The catch is, while some (maybe most) of those studies are failures for boring reasons (e.g. you stuffed something up) others might be genuine "null" results that you ought to acknowledge when you write up the "good" experiment. And telling which is which is often hard to do. A good place to start is a paper by **Ioannidis2005** with the depressing title "Why most published research findings are false". I'd also suggest taking a look at work by **Kuhberger2014** presenting statistical evidence that this actually happens in psychology.

There's probably a lot more issues like this to think about, but that'll do to start with. What I really want to point out is the blindingly obvious truth that real world science is conducted by actual humans, and only the most gullible of people automatically assumes that everyone else is honest and impartial. Actual scientists aren't usually *that* naive, but for some reason the world likes to pretend that we are, and the textbooks we usually write seem to reinforce that stereotype.

1.8

Summary

This chapter isn't really meant to provide a comprehensive discussion of psychological research methods. It would require another volume just as long as this one to do justice to the topic. However, in real life statistics and study design are so tightly intertwined that it's very handy to discuss some of the key topics. In this chapter, I've briefly discussed the following topics:

- *Introduction to psychological measurement* (Section 1.1). What does it mean to operationalise a theoretical construct? What does it mean to have variables and take measurements?
- *Scales of measurement and types of variables* (Section 1.2). Remember that there are *two* different distinctions here. There's the difference between discrete and continuous data, and there's the difference between the four different scale types (nominal, ordinal, interval and

---

[*8]Clearly, the real effect is that only insane people would even try to read *Finnegans Wake*.

ratio).

- *Reliability of a measurement* (Section 1.3). If I measure the "same" thing twice, should I expect to see the same result? Only if my measure is reliable. But what does it mean to talk about doing the "same" thing? Well, that's why we have different types of reliability. Make sure you remember what they are.

- *Terminology: predictors and outcomes* (Section 1.4). What roles do variables play in an analysis? Can you remember the difference between predictors and outcomes? Dependent and independent variables? Etc.

- *Experimental and non-experimental research designs* (Section 1.5). What makes an experiment an experiment? Is it a nice white lab coat, or does it have something to do with researcher control over variables?

- *Validity and its threats* (Section 1.6). Does your study measure what you want it to? How might things go wrong? And is it my imagination, or was that a very long list of possible ways in which things can go wrong?

All this should make clear to you that study design is a critical part of research methodology. I built this chapter from the classic little book by **Campbell1963**, but there are of course a large number of textbooks out there on research design. Spend a few minutes with your favourite search engine and you'll find dozens.