Does measuring people change them?

David P. French and Stephen Sutton call for more understanding of reactivity

People who are aware that they are in a psychological study may not behave in their normal way. For this reason, ‘unobtrusive’ measurement has been advocated, whereby people are not aware they are being studied (Webb et al., 1966). A good recent example of this was a study that varied messages about hand-washing outside a motorway service station, and assessed the number of people who entered the toilets and number of uses of soap dispensers using electronic sensors (Judah et al., 2009). This procedure allowed different messages to be presented at different times of day, in a randomly determined order, and the effects on handwashing to be objectively assessed, with the people whose behaviour was manipulated unaware that an experiment was taking place.

Although such approaches have much to recommend them and are probably under-used, in many instances psychologists wish to measure constructs such as emotions or beliefs, which are difficult to assess unobtrusively with good levels of validity. Consequently, self-report measures are widespread.

Unfortunately, the mere act of measurement may be sufficient to affect the people who complete the measures. This possibility was noted in relation to mental testing over 40 years ago: ‘we can perhaps repeat a measurement once or twice, but if we attempt further repetitions, the examinee’s response changes substantially because of fatigue or practice effects’ (Lord & Novick, 1968, p.13). Despite this observation, psychologists (and others) usually assume that the process of participants being interviewed or completing questionnaires does not result in them having different thoughts, feelings or behaviour as a consequence. However, there is increasingly strong evidence that this assumption is not always valid: the process of psychological measurement can affect people’s thoughts, feelings and behaviour, and is therefore ‘reactive’ (French & Sutton, 2010).

Emotion

One area where the evidence of measurement reactivity is compelling relates to emotional reactions in people completing measures concerning personally salient illness. For example, one study involving women with breast cancer placed a measure of general anxiety at either the beginning or the end of a questionnaire, according to experimental condition (Johnston, 1999). The questionnaire assessed ‘demographic and clinical factors, social support, and attitudes for attending a social support group for women with breast cancer’ (p.78). Women who completed the anxiety measure at the end of the questionnaire had significantly higher anxiety scores than those who completed it at the beginning, with the most plausible explanation being that the other questionnaire items raised the anxiety of these women.

A reduction in negative emotion between the first and subsequent occasions of measurement has also been observed. For example, one study asked undergraduate students to complete a battery of emotion measures on two occasions, one week apart (Sharpe & Gilbert, 1998). A significant reduction in depression scores was observed on the second occasion of measurement. Similar results were obtained for other measures of affect, and this drop in scores upon repeated completion of measurements was replicated in a second sample.

These two measurement artefacts have the potential to bias the conclusions drawn from research studies, if ignored (French & Sutton, 2010). One such area of research concerns emotional reactions to the results of health screening tests (Johnston et al., 2004). It has been observed on many occasions that receipt of a positive screening test result is associated with elevated anxiety in the short term, with this anxiety returning to normal levels in the longer term (Shaw et al., 1999). Given that few studies examining this issue obtain true baseline measures of anxiety (i.e. before screening), it is not possible to be sure if the higher anxiety in the short-term is generated by receipt of a positive screening test result or if the higher scores are an artefact of measurement, due to participants completing questions that require them to consider the potentially distressing consequences of the illness. Equally, the reduction in anxiety over the longer term may reflect a coming to terms with the initially distressing result or it may be an instance of the observation that people’s negative affect scores tend to drop on the second occasion of measurement.

Recent research suggests that it is unlikely that measurement reactivity is responsible for the typical pattern of distress scores observed after receiving a positive screening test (i.e. increased anxiety in the short term returning to normal levels in the long term) (Shaw et al., 1999). Specifically, postal

References


questionnaires concerning diabetes have not elicited any discernible effects on anxiety on subsequent occasions of measurement, relative to people who have not previously completed such measures (French et al., 2009). Possible reasons for the postal questionnaire study not finding the effects on emotion scores obtained in earlier studies include the absence of an interviewer, respondents having more time to complete the measures (and for any distress to dissipate), and those who are most distressed being more easily able to drop out of the study.

**Behaviour**

Say you provide people with pedometers, and conclude that people given pedometers are subsequently more physically active (Bravata et al., 2007). Is it the process of measurement that is having these effects or the associated use of effective behaviour change techniques such as goal setting (Michie et al., 2009)? Similar findings include the observation that completing an alcohol disorders screening questionnaire affects reporting of alcohol consumption two to three months later (McCambridge & Day, 2007), and being asked to complete a questionnaire relating to blood donation, along with postal reminders and thank-you letters, appears to lead to higher rates of objectively assessed blood donation behaviour (Godin et al., 2008).

If measurement leads to such changes in health-related behaviours, then it may be more difficult for deliberate interventions to have additional effects. For example, a recent study (ProActive) that attempted to bring about increases in physical activity found that all three experimental groups increased their physical activity by the equivalent of 20 minutes brisk walking per day (Kinmonth et al., 2008). The authors attributed this, at least in part, to a measurement reactivity effect: by assessing motivation using questionnaires, objective behaviour using heart-rate monitors and a treadmill exercise test, and a variety of physiological measures, participants may not only have become more convinced of the importance of physical activity but may also have become aware of their own low levels of activity and gained some insight into the psychological processes by which an increase might be brought about. The combined effects of these measures may have been sufficient to preempt any effects of the behaviour change intervention that differed between experimental conditions. If this reasoning is correct, future interventions to increase healthy behaviour could profit from a greater understanding of the processes underlying measurement reactivity.

**Dealing with measurement reactivity effects**

Exactly how measurement has the reactivity effects described is currently poorly understood (French & Sutton, 2010); we await more thorough theorising of the multiple mechanisms that may be involved and empirical tests of such theories. However, most researchers will be most interested in how to avoid such reactivity effects biasing their own research findings. There are a number of steps researchers can take with regard to measurement reactivity effects, some of which are helpful in identifying these effects, and some are helpful with avoiding their impact.

Given sufficient resources, the ideal approach involves the use of ‘Solomon designs’, where people are randomised not only to receive an intervention or not, but also to receive pre-test measures or not (e.g. Spence et al., 2009). The use of this design thereby not only examines the effects of an intervention, which is usually what researchers are interested in, but also any reactivity to measurement, and crucially any interaction between intervention and measurement, such as was proposed above for the ProActive study.

More generally, the order of any outcome measures that are thought particularly likely to be affected by order effects could be counter-balanced for order. This is common practice within many laboratory studies, but less common in more ‘applied’ field studies. However, although this will control for order effects (see Schuman & Presser, 1981), this does not control for any changes in measurement across multiple periods of follow-up. Thus, although this simple approach may be useful for some purposes, it does not control for all forms of measurement reactivity. For this to be possible, a greater understanding of the multiple mechanisms involved in producing reactivity effects is required, to be able to control for each possible source of reactivity.

Another simple method that has been proposed is to place the most reactive measure at the beginning of the questionnaire (Johnston et al., 2004). Thus, given that asking people questions about their illness appears to result in higher anxiety scores (Johnston, 1999), it may be sensible to place the anxiety measure at the beginning of a set of questionnaire measures. It is important to note that counterbalancing of the order of measures, an approach which is useful for...
detection of measurement reactivity effects, would only reduce the impact of such effects in this example, not eliminate them completely.

The impact of measurement reactivity can also be reduced by requiring participants to only complete a measure on one occasion: if there is good reason to believe that measures are reactive, this approach removes any possibility of reactivity effects on subsequent occasions of measurement. This approach does not, however, avoid reactivity effects within a single occasion of measurement, such as completing questions about an illness resulting in higher anxiety. Further, assessing different samples at single (different) timepoints is less statistically efficient, and therefore requires larger sample sizes. Equally, experiments that use post-test-only designs do not allow us to assess the possibility that there is a difference between samples at baseline.

It is sensible to be particularly cautious when the effects of measurement are likely to be similar to those of the phenomenon being studied. For example, questions testing knowledge should be carefully designed not to provide information about the topic being assessed. The use of the same items on multiple occasions would be particularly unwise. Another example would be when an intervention encourages self-monitoring of own performance against a criterion such as a personal goal (see Michie et al., 2009). Given that psychological measurement appears to encourage such monitoring, and potentially similar effects as the deliberate self-monitoring intervention, it may be prudent to consider post-test-only designs in such situations.

A final area that warrants caution is where the research involves assessing beliefs about an issue that the participant has not previously thought about. ‘Non-attitudes’ can be said to be present when a sample of respondents choose the ‘don’t know’ option in response to a question designed to assess attitude, whereas a similar sample choose an apparently meaningful option to the same question, when the ‘don’t know’ option is not included (Schuman & Presser, 1981).

More recently, studies have looked at this in more depth by asking people to ‘think aloud’ whilst they complete questionnaires about issues such as physical activity and drinking alcohol (e.g. Darker & French, 2009; French et al., 2007). These studies have shown that when people are asked to complete questions about issues they have not previously considered, they provide questionnaire answers that are generated on the spot, on the basis of inferences from what they do know.

It is clear from examination of the literature on measurement reactivity effects that they are poorly understood. The best future defence against research conclusions being biased by these effects is to increase our understanding of why they are likely to arise, and under what circumstances.

David P. French is at the Applied Research Centre in Health and Lifestyle Interventions, Faculty of Health and Life Sciences, Coventry University
david.french@coventry.ac.uk

Stephen Sutton is at the Institute of Public Health, University of Cambridge
srs34@medschl.cam.ac.uk

“measurement reactivity effects are poorly understood”