Chapter 6: Psychometrics: designing and road testing new measurement scales

Professor Patrick Callaghan

Chapter overview

Measurement scales (questionnaires) are often used in quantitative research to summarise the experiences of a group of participants, for example the number and range of symptoms they report, or their level of satisfaction with their care. We can use these questionnaires once to get a snapshot of people’s scores at one point in time, or we can ask people to complete them on more than one occasion to see if their scores change. For example, if we compare people’s symptom scores before and after treatment we can get an idea of whether or not the treatment they were given helped. To do this, it is essential that the scale or questionnaire that we are using has been designed to measure the outcome that we want to measure, and that it has been road-tested on a similar group of people to make sure they understand it, like it and can complete it in a way that works.

Most existing scales have been designed by clinicians, academics and researchers and often focus on people’s ‘clinical’ outcomes, e.g. their symptoms. These clinical measures are often criticised by service users, especially if they do not tap into the priorities of the service users themselves (Crawford, 2011). As a result, we have seen increasing emphasis placed on the development of Patient Reported Outcomes Measures (PROMS). These tend to be less focussed on symptoms and more on the everyday experiences of people using services. They are much more likely to be designed and developed in collaboration with service users. The EQUIP research project developed a good quality PROM for assessing user and carer involvement in care planning, the first such measure of its kind in mental health.
Learning objectives

By the end of this chapter you should be able to:

1. Understand more about the origins of measurement scales in research
2. Understand what might influence our choice of measurement scales
3. Begin to understand how new measurement scales are designed, developed and evaluated.

Introduction

Psychometrics is the term used to describe the science of psychological testing and is concerned with the measurement of mental and behavioural processes. Objective measurement is at the heart of psychometrics. In quantitative research this commonly involves the use of measurement scales, often referred to as questionnaires. The use of measurement scales is widespread in quantitative research. However, prior to their use, researchers must ensure that such scales are robust, i.e. they are reliable and valid.

Using scales that are unreliable or invalid is a major threat to the integrity of quantitative research.

This chapter will examine the origins of measurement scales in research by considering the science of psychological testing. In particular the chapter will provide a brief definition of a measurement scale, outline why scales are used, examine the design and evaluation of scales, discuss what the responses to scales mean, outline advantages and limitations of their use, and provide examples of measurement scales developed and used in the EQUIP project and other published mental health research.
Measurement scales

A measurement scale is a device for measuring pre-specified outcomes e.g. a person’s reported state of mind, behaviour, performance, attitudes, intentions, abilities, personality, beliefs, cognitive functioning or style, preferences or coping style. The term measurement scale is often used interchangeably with rating scale, test, inventory, questionnaire or measure. Measurement scales can be used to look at relationships between different characteristics. For example, in a recent study exploring the relationship between resilience and depression, the researchers used a specific measurement scale – The Connor-Davidson-Resilience Scale [CD-RISC] - to quantify each participant’s resilience (Smith, 2009). They can also be used to provide an assessment of people as a baseline against which to measure the success of interventions, e.g. the use of the Positive and Negative Symptoms Scale (PANSS) to assess the effect of Cognitive Behaviour Therapy on psychotic symptoms. Scales may also be used to measure the behaviour of others, for example rating the severity of observed aggression using the Staff Observation Aggression Scale – Revised (SOAS-R), or as an assessment of performance, e.g. using an appraisal tool.

Clinical vs patient-reported measures

Outcome measures can have clinical utility and meaning (e.g. classification of illness severity), or they can be used to understand more about patient experience and the treatment outcomes that are most important to patients. Patient-reported outcomes (PROMs) and patient reported experience measures (PREMs) collect information directly from patients, without interpretation by health professionals. This means that they often reflect the priorities and treatment outcomes that are most relevant to patients themselves. PROMs can include measures of symptoms, daily activities, and functioning. They can also be used to measure patient satisfaction or treatment preferences.

Modern healthcare systems place considerable emphasis on delivering high quality, patient-centred care, and as such PROMs have enormous potential to trigger changes in healthcare delivery. Developing meaningful and useful PROMs and adopting these tools in research studies represents a major step forward in assessing the outcomes of new interventions and approaches to care.
The evaluation of measurement scales

Measurement scales are critically important to data collection and interpretation, and must accurately measure what they propose to measure, i.e. they must have sound psychometric properties. The psychometric properties of a scale are usually established during its design and development phases.

Using scales whose reliability and validity are weak is a major threat to the validity of research, is unethical, and may produce false results. Therefore, measurement scales need to be reliable and valid. There are some key issues to consider when assessing measurement scales for use in a research study.

1. Minimising Error

All measurement scales are subject to error of some degree or another. Two types of errors are possible: systematic and random. Systematic error is less of a problem because it is predictable and affects all scales, like the error associated with a watch that loses a few seconds every 24 hours. Random error, however, is a major problem as it occurs randomly and does not happen with every participant who uses the scale. Sources of random error may include participants being upset, tired, or ill when completing the measure, or because questions are difficult to understand or answer, and instructions on how to use the measure are poor. The way in which different questions relate to each other and influence how a particular person might answer them can also cause random error.

For any measure a person has a true score; the score that a person obtains on a scale on any occasion is an obtained score (a close estimate). Testing participants on several occasions to get more than one obtained score and taking their average score gives us the best estimate and moves us closer to their true score, i.e. it reduces random error.
2. Establishing cut-off scores

A cut-off score is the minimum score used to decide a person’s status. For example, the Beck Depression Inventory has three cut-off scores, used to indicate mild, moderate or severe depression levels. The Angoff method (Goodwin, 1996) is the most widely used method in determining cut-off scores and involves asking experts to rate the probability that the item will be an accurate measure of the concept. If the wrong cut-off score is used, a person’s outcome status may not be correctly recorded.

3. Ensuring reliability

Reliability is the degree of stability with which a scale measures what it is designed to measure. There are several different types of reliability that you might come across:

- internal consistency
- test-retest
- parallel form
- inter-rater

Internal consistency reliability refers to how well individual items (or questions) in a scale measure the same thing. Internal consistency is determined by calculating a Cronbach’s Alpha score between 0 and 1. A score closer to one indicates that the scale has a high internal consistency; a minimally acceptable internal consistency is around 0.75.
Test-retest reliability refers to the likelihood that a scale will measure the same score in the same participant, on more than one occasion, assuming that their circumstances have not changed. It is measured on a scale from 0 (low reliability) to 1 (high reliability) and involves linking the sets of scores of participants taking the same test at two different times. For test-retest reliability to be meaningful, it should have been tested in a representative sample of at least a 100 participants, 3 months apart (Kline, 2000a).

Parallel form reliability is tested using different versions of the same measure. Participants’ scores are linked and compared between the different versions to check if their responses are stable.

Inter-rater reliability is the relationship between ratings of two independent raters of the same behaviour. To assess the reliability of scales we use scores from 0 = low reliability to 1 = the highest level of reliability.

4. Ensuring Validity

Validity is the degree to which a scale measures what it was designed to measure. Different types of validity exist and include face, concurrent, content, construct, incremental and predictive (criterion) validity. As with reliability, large, representative samples are needed to establish validity.

**Face validity** is not really a form of validity as it only measures the participants’ perception of what a scale is measuring. Nevertheless, it is important because if a participant questions the face validity of a scale they may not cooperate and this will jeopardise attempts to establish ‘true’ validity. Face validity can be established by simply asking participants what they think the scale is measuring.

**Concurrent validity** is assessed by linking scores on one scale with another measure of the same construct. This is problematic because it requires a benchmark measure of the construct (e.g. finding a scale you know measures the construct accurately) and these are extremely difficult to find.
Of course, if a benchmark measure exists it would call into question the need for a new measure. For example, in mental health we can establish concurrent validity by asking people to complete the Beck Depression Inventory and another depression measure and see how the scores to both compare.

**Predictive validity** occurs when people’s scores on one scale predict other related things. For example, because we know anxiety and depression are often strongly related, we could assess whether high scores on the Beck Depression Inventory (BDI) will predict the likelihood that a person will also be anxious.

**Construct validity** is the extent to which items on a measurement scale are accurate indicators of the variable being measured. To understand construct validity a bit more, it is helpful to know what is meant by constructs. Constructs are abstract attributes that do not exist in a physical sense, such as intelligence, happiness, motivation, attitudes, self-esteem, but which are common areas of assessment for researchers and clinicians. Thus, reliable and valid measures of these constructs are necessary. The construct is the theory that underlies the measure. For example, The Eysenck Personality Questionnaire has construct validity because it is based on Eysenck’s personality theory.

**Content validity** is the extent to which a measure evaluates all the different aspects of a construct. For example a depression scale would ideally measure a person’s feelings, thoughts and behaviours to judge whether or not a person is living with depression, and/is how severe this depression is. Content Validity can be established by asking a panel of experts in the field what they think should be included. McDowell (2006) suggests that any effort to improve content validity must consider how well items are presented to participants and how well responses are recorded as well as the content of the questions themselves.

**Factor Analysis (FA)** a type of statistical procedure, can also be useful for assessing content validity as it allows you to determine whether items on a scale are measuring the same underlying construct. Items that measure the same thing ‘group together’ in the test. It is these groups, also known as ‘factors’, which give the test its name. FA has two main types: Exploratory and confirmatory. Exploratory Factor Analysis (EFA) is used when researchers develop a new scale and wish to test the number and meaning of factors that the items on the new scale measure. Confirmatory Factor Analysis (CFA) which looks at whether the factors identified for an existing scale in one population, also remain valid when the scale is tested on a different population.
Interpreting people’s responses to measurement scales

Measurement scales generate a score that is obtained by assessing a person’s response, or an assessor’s response, to the items on the scale. Responses are often anchored by a numerical value and this value often indicates the strength of the response. For example, a pain score might measure the severity of a person’s pain on a scale from 0 (none) to 10 (unbearable), or an opinion-based score might measure a person’s level of agreement from 0 (strongly disagree) to 5 (strongly agree). Respondents’ scores might also be used to categorise their responses, e.g. responses to the Beck Depression Inventory (BDI) generate a score that categorises depression as mild, moderate or severe.

Responses to measurement scales are seldom used as the sole criterion in diagnosis; they may be a useful adjunct to other forms of assessment, such as a diagnostic interview. For example, responses to the Eating Disorder Inventory (EDI) do not indicate that the respondent has an Eating Disorder; the responses indicate any similarity in scores between the respondent and the responses of those who have been diagnosed with an Eating Disorder. Because of the dangers in misinterpreting scores, many scales and tests are authorised for use only by those with a minimum level of training.

The advantages and limitations of using a measurement scale

Most measurement scales are relatively easy to use and may be a useful adjunct to other forms of assessment in helping clinicians arrive at accurate diagnoses. Scores generated from measurement scales provide a baseline, or benchmark against which the efficacy and effectiveness of interventions may be assessed. Measurement scales may be useful in tracking how much peoples’ symptoms change over time or with treatment, or to provide feedback to people on the level of performance they attain on a particular task.

Responses to measurement scales are usually restricted to answering the items on the scale; therefore they may not reflect the totality of the respondents’ views. Often measurement scales require that the user is trained in their use and this can be costly and time consuming. Also, it is difficult to interpret what respondents’ scores on scales actually mean in practice; if a person’s symptom score decreases by 1, is this meaningful? How much might a numerical score need to change before the person themselves feels they are better? It is also not easy to arrive at meaningful cut-off scores in order to categorise responses. For example, on the BDI, the categorization of someone as mildly, moderately or severely depressed hinges on small differences in scores. Depression in reality is seldom so easy to categorise.
Evaluating and Quantifying User and Carer Involvement in Mental Health Care Planning (EQUIP): Co-development of a new Patient-Reported Outcome Measure (PROM)

Items for the EQUIP PROM were developed from 74 interviews and 9 focus groups conducted with service users, carers and mental health professionals recruited from two large NHS Trusts. From these data, 70 items (potentially relevant questions) were developed.

First, face validity was examined with a mixed sample of 16 members of a Service User and Carer Advisory Group (SUCAG). Nine items were rejected by this panel as not being useful or relevant to what was being measured. The remaining 61 items comprised the first draft scale.

Members of the SUCAG were asked to comment on potential response formats for the scale. Consensus was reached for a 5-point Likert scale, with named anchors of ‘Strongly disagree’ at one end and ‘Strongly agree’ at the other. A middle neutral value with the category label “Neither agree nor disagree” was included (Figure 17 the EQUIP PROM).

### Figure 17

<table>
<thead>
<tr>
<th>Item</th>
<th>Strongly agree</th>
<th>Neither agree nor disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>The care plan has a clear objective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am satisfied with the care plan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am happy with all of the information on the care plan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The contents of the care plan were agreed on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Care is received as it is described in the care plan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The care plan is helpful</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>My preferences for care are included in the care plan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The care plan is personalised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The care plan addresses important issues</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The care plan helps me to manage risk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The information provided in the care plan is complete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The care plan is worded in a respectful way</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important decisions are explained to me</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The care plan caters for all the important aspects of my life</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Following initial development with service users and carers, the 61-item scale was piloted with a sample of 402 service users and carers. The completed questionnaires were anonymised and the data were entered into a computer package. The team applied different forms of psychometric testing to identify unnecessary questions, remove questions that were difficult for people to answer and reduce the length of the scale. This analysis showed that there were 14 important items that needed to be retained in the final scale.

Additional psychometric testing confirmed that the final 14-item PROM was scientifically reliable and valid, and that it could be used by any mental health service seeking to assess the involvement of service users and carers in care planning.

**Conclusion**

Psychometrics allow researchers to illustrate how mental health care and outcomes associated with it can be measured. In order to measure mental health care and its outcomes using the principles and science of psychometrics, a measure must be a reliable and valid. The reliability and validity of measurement scales is an essential requirement to establish the accuracy of scales and reduce threats to research integrity. Psychometric testing was used in the EQUIP research project to develop a co-produced, robust Patient-Related Outcome Measure for assessing user and carer involvement in care planning, the first such measure of its kind in mental health.

**PPI stories from EQUIP**

Next, Joe and Andrew share with you their experiences of being involved in the EQUIP PROM.

**Joe’s story**

From the outset of EQUIP the service users and carers who were involved in the programme felt that there were no adequate measures for assessing involvement in mental health care planning.

As a group we reviewed lots of existing measures. We had some intense conversations about the wording of questionnaires, and these discussions highlighted how loaded some words can be.
The term ‘Service-users’ is a good example. Are we users? Sometimes it can sound pejorative, as in he’s just a user, a taker never a giver. The word ‘relapse’ is another. For some people it fittingly described what can happen but for others it felt like a rebuke. Getting the language of the questionnaire right was so important, and like a lot of things in life there were no simple right/wrong answers.

As a group we felt strongly that we needed to devise a new measure, one that was both relevant and acceptable to service users. Our opinions fed straight into the design of the EQUIP programme – the team included a whole study dedicated to developing and testing a new questionnaire to measure service user involvement in care planning.

We then devised a new Patient Reported Outcome Measure (or PROM) – a questionnaire, completed by a service user, to measure quality involvement in care planning. We met as a group to draft the questions and discuss how they should be worded.

As PPI representatives and researchers, we used our existing networks, including social media and our contacts with local service user and carer groups, to get as many people as possible to complete our new questionnaire. This gave us lots of data and meant that we could validate the measure properly.

We have been able to develop a short 14-item questionnaire that is valid and reliable, and can be used by researchers and health services to measure service-user involvement in care planning.

I thoroughly enjoyed this experience, and am proud of our measure. I’ve since gone onto to assist in the design of other PROMs for mental health service users.

Andrew’s story

We knew that developing a new patient-reported outcome measure for service involvement in care planning was important.

To begin with, I co-facilitated focus groups and interviews with service users, carers, and mental health professionals. With people’s consent, we recorded these discussions and transcribed what everybody said. By reviewing the written transcripts, we were able to identify the key components and priorities for service-user involvement in care planning.

We then devised a new Patient Reported Outcome Measure (or PROM) – a questionnaire, completed by a service user, to measure quality involvement in care planning. We met as a group to draft the questions and discuss how they should be worded.

As PPI representatives and researchers, we used our existing networks, including social media and our contacts with local service user and carer groups, to get as many people as possible to complete our new questionnaire. This gave us lots of data and meant that we could validate the measure properly.

We have been able to develop a short 14-item questionnaire that is valid and reliable, and can be used by researchers and health services to measure service-user involvement in care planning.

I thoroughly enjoyed this experience, and am proud of our measure. I’ve since gone onto to assist in the design of other PROMs for mental health service users.
Reflective Exercise

- Give examples of three different types of measurement scales that might be used in mental health.
- Define, in your own words the terms reliability and validity.
- Why is it important to ensure reliability and validity of measurement scales?

Allied EQUIP Papers


References and Further Reading


PLOS One 3:e0149973. Doi:10.1371/journal.pone.0149973.