Reliability and validity in clinical psychology

To be reliable as a diagnostic classification system, there would have to be consistency with the DSM. This means that the DSM is reliable if the clinicians using it consistently arrive at the same diagnoses as each other. The term inter-rater reliability is used to describe the extent to which different clinicians agree on the same diagnosis for the same patient.

Traditionally, reliability is calculated mathematically, often using a measure known as positive predictive value (PPV). The PPV of a disorder shows the reliability, taking the example of depression, of the DSM with that disorder; so if depression had a PPV value of 80, this means that 80% of diagnosed depression patients will have the same subsequent diagnosis when re-assessed. However, there may also be a cultural element to reliability, for example with Cooper et al. (1972) who showed the same patient interview tapes to various American and British psychiatrists, and American clinicians diagnosed schizophrenia twice as often as the British, and the British clinicians diagnosed depression twice as often as the American.

Validity is the extent to which a measure of a psychological variable measures what it sets out to measure. Essentially this means the correct variable (in clinical psychology, this variable will be a mental disorder) is measured, by arriving at the correct diagnosis. Needless to say, if the DSM were not reliable, it would not be valid either. This is because if it is unreliable it means inconsistent diagnoses are made, and so it must not be valid either as surely the correct diagnosis is being made.

- **construct validity** refers to whether or not a classified disorder is actually a good indicator of what you’re really trying to measure, for example, in operationalising (see below) a disorder such as depression by drawing a list of symptoms and features, you begin to lose some understanding of the real nature of the disorder so the DSM becomes invalid
- **concurrent validity** is when the results of a study matches the result of another study done at the same time, so this would mean that if a diagnosis made using the DSM arrives at the same mental disorder that another diagnosis has, it is likely to have concurrent validity
- **predictive validity**, on the other hand, is when the results of a test or study match the result of another done at a different time, so instead of looking at two diagnoses to see if they back each other up, the comparison is made over two different time periods, for example the DSM could be used for a diagnosis, then some time later another measure (perhaps a doctor’s view or observations by mental health personnel) would agree with the diagnosis to be valid
- **convergent validity** is when a test result converges (gets close to) the result of another test measuring the same thing – a correlational test would be carried out to see the convergence (the difference between convergent validity and concurrent/predictive is that convergent validity must be measuring the same thing, it can be different measures in concurrent and predictive validity)

Operationalising mental disorders

Mental health disorders need to be operationalised is they are to be definable within the DSM. This would mean arriving at lists of symptoms and behaviours to make the disorder measurable. It has been argued, however, that in operationalising a concept such as depression, something is lost from the understanding of the nature of the whole experience of depression, which means that the DSM is not a valid tool in that is lacks construct validity, in that the constructs which are drawn up may not actually be sufficient to represent the disorder.

A further possible problem with validity and the DSM is that although taking into account axes IV and V in the later versions of the DSM (personal and social factors and how well the patient is functioning in society), taking such factors into account when diagnosing can actually lead to an invalid diagnosis. For example, someone diagnosed with depression may not be functioning well in society at all, but this might not be due to their diagnosed disorder but actually for another reason, possibly unemployment (for example), and so such a diagnosis would not be valid.

It has also been argued that since there has been significant change in the content of the DSM with regards to categorising certain disorders that the manual is invalid. For example, homosexuality and epilepsy have both been considered mental disorders and been included in the DSM at one stage, but are no longer, which might suggest that the DSM has low validity.
Studies evaluating the DSM in terms of reliability

The course requires that you know studies to evaluate the use of the DSM as a classification system in terms of reliability and validity. Outlined here are three studies supporting the DSM in terms of reliability, and one criticising the DSM.

Goldstein (1988)
[For the full study by Goldstein, see 3.9 Goldstein (1988) in this section] Goldstein tested the DSM for reliability using the at-the-time current version, DSM-III. One of the aims of her study was to test the DSM-III, comparing the results of the re-diagnoses of 199 patients who had been originally diagnosed with schizophrenia using DSM-II. Experts carried out a re-diagnosis of a random sample of eight patients using a single-blind technique (not allowing the experts to know the hypothesis, so their answers are not biased, whereas Goldstein herself was aware of the hypothesis).

She found that 169 of the 199 patients diagnosed according to DSM-II as having some form of schizophrenia met the DSM-III criteria too, so reliability was seen as good with the DSM. Of the patients assessed by the clinical experts as well, she found high levels of inter-rater reliability.

Brown et al. (2001)
In 2001, Brown et al. studied anxiety and mood disorders in 362 outpatients in Boston, to test reliability of the DSM-IV and patients underwent two independent interviews using anxiety disorder interview schedules for DSM-IV, known as the life-time version.

Brown found good-to-excellent reliability for most of the DSM-IV categories (most of the disagreements tended not be on what the symptoms were, but simply if there were enough of them). However, they found some boundary problems with certain disorders, which made it hard to diagnose patients with disorders if they were at boundary level. Overall, the study highlights some problems with the DSM but generally proves it to be a reliable tool.

Stinchfield (2003)
Stinchfield studied both reliability and validity by looking at the diagnosis of pathological (behaviour which is considered abnormal due to its extreme or excessive nature) gambling. 803 people were studied from the general population, and 259 people on a gambling treatment programme, all from Minnesota, and they were assessed using a questionnaire of 19 items used to measure the DSM-IV diagnostic criteria for pathological gambling.

The DSM criteria were used to sort those linked to pathological gambling and those who were not. There were other validity measures as well. It was found that the DSM-IV criteria were both reliable and valid.

Kirk and Kutchins (1992)
In a review paper, Kirk and Kutchins argued that methodological problems with studies conducted to test the reliability of the DSM up until 1992 had limited the generalisability of their findings. For example, they argued that there had been insufficient training and supervision of interviewers, and studies tended to take place in specialised research settings, and so could lack validity as well as reliability.

Assessment of Kirk and Kutchins’ points
- Due to the specialised settings the findings may not be valid as they may not relate to the conclusions of real clinicians
- Their review paper was published in 1992, before the studies of Brown and Stinchfield, who showed the DSM-III and DSM-IV to be reliable, so it is suggested that reliability of the DSM has improved since then
- The accuracy of their criticisms is debatable, as the above three studies all found the DSM to be a reliable tool
- Some of the points about interviewing – such as that different interviewers may affect the situation and lead to different data – might be important, when considering generalising findings from studies; Goldstein, however, did not use interviewing to test reliability, she used re-diagnosis using secondary data and still found reliability

Another study you could use for evaluation is Nicholls et al. (2000), which showed that using the DSM-IV there was not good inter-rater reliability for the diagnosis of eating disorders in children.

Studies evaluating the DSM in terms of validity

The studies on the following page have evaluated the use of the DSM in terms of validity. It is recommended that you learn at least two studies from this selection.
In order to be valid, the diagnoses must identify a distinct condition that has different symptoms from other conditions and that is likely to progress in a certain way and respond to one treatment over another. A valid diagnosis for a mental disorder is more difficult than for a physical disorder because of a lack of objective physical signs. To be valid, the DSM also has to be reliable.

**Kim-Cohen et al. (2005)**

In 2005, Kim-Cohen et al. undertook a longitudinal study of conduct disorder in five year olds, to test the concurrent, convergent and predictive validity of the DSM-IV. There were 2,232 children involved in another, existing, longitudinal study which was used as a focus. The children’s mothers were interviewed and teachers received postal questionnaires asking about conduct disorder symptoms over the previous six months. They found that 6% were diagnosed as having conduct disorder (displaying three or more symptoms), and 2% of the children with severe conduct disorder (five or more symptoms). Children diagnosed with conduct disorder were more likely to describe themselves as having antisocial behaviours than comparison children. Also, during observational assessments these children were more likely to behave disruptively. Different measures were said to show the diagnoses had validity, as different data sources were used to check validity.

**Hoffmann (2002)**

Hoffmann studied prison inmates to look at diagnoses of alcohol abuse, alcohol dependency and cocaine dependency, to see if differences would occur in a computer-prompted structured interview, compared to the DSM-IV-TR criteria. It was found that the DSM-IV-TR diagnosis was valid and that the interview data supported the idea that dependence was more a severe syndrome than abuse. The symptoms from the automated interview matched those of the DSM criteria.

**Lee (2006)**

Lee studied the DSM-IV-TR diagnosis of ADHD to see if it would be suitable for Korean children, and looked at gender differences in the features of ADHD in the DSM. The DSM lists eighteen criteria for ADHD linked to children’s behaviour. In total, 48 primary school teachers rated the behaviour of 1,663 children (904 of which were boys, the remaining girls) using a questionnaire. Lee looked for concurrent validity by comparing the DSM-IV-TR criteria with criteria arising from the questionnaire, and compared DSM behavioural and psychological characteristics with those found in an ADHD test. Previous studies had showed that ADHD children had oppositional deficit disorder, ODD, as well, having problems with peers and discipline. Lee decided that finding the same correlation would support the diagnosis and show the DSM to be a valid tool. The same relationship was observed, and so it was said that the DSM-IV-TR had concurrent validity. Also found it to be reliable, as the correlation could check for similar diagnoses. However, the study found that for girls, the DSM-IV-TR symptoms and diagnoses were less compatible than they were for boys, which was a weakness found with the DSM as a diagnostic tool.

**Culture affecting diagnosis**

Whilst all of these above studies show the DSM to be a reliable and valid tool, the area it receives most criticism is in its usefulness across different cultures. There are two schools of thought, outlined below:

**Culture doesn’t affect diagnosis: mental disorders are ‘scientific’**

- The DSM was developed in the USA and is used widely across many other cultures – this is a valid use if mental disorders are clearly defined with specific features and symptoms
- In other words, this school of thought suggests mental disorders are scientifically defined illnesses that are explained in a scientific way and therefore culture does not affect diagnosis as it should be the same cross-culturally
- The study of Lee (2006) can be used to support this, as the DSM-IV-TR was used deliberately in a non-western culture to see if ADHD diagnoses were valid in Korea

**Culture does affect diagnosis: a spiritual model**

- There have been studies which have shown that culture can affect diagnosis, for example, hearing voices in western cultures is normally an abnormal sign of, for example, schizophrenia, whereas in other countries this may be seen as a positive characteristic, such as a sign of being connected to spirits
- Depending on cultural interpretations of what is being measured, the DSM is not always shown to be valid
Cultural differences in the symptoms of schizophrenia

You are required to know how culture affects the symptoms of schizophrenia, the required mental disorder you should know about for your course. It has been reported that catatonic schizophrenia (see 3.5 Schizophrenia) is on the decline, and this could be because of health measures that prevent the development of this type of schizophrenia.

Auditory hallucinations were reported to doctors by patients more in Mexican-born Americans than in non-Mexican-born Americans. Burnham et al. (1987) looked at this using self-reports and interviewing, and checked the evidence and found that there was a difference. No other explanation could be found except that culture had led to the difference.

White Americans were reported, using patients’ records, as showing more grandiosity as a symptom compared with Americans of Mexican origin, again showing cultural differences.

However, it should be noted that schizophrenia in all countries has more similarities than differences.

Culture-bound syndromes

A culture-bound syndrome (CBS) is a disorder which is isolated to one culture, usually only diagnoses exclusively in one region or country. Psychiatrists tend to reject culture-bound syndromes, although some are listed in the DSM-IV. Two examples of culture-bound syndrome are described below:

- one example is hikikomori (literally meaning ‘to be pulled away’ and translated to English as ‘complete social withdrawal’), a condition which has attracted concern in Japan recently, affecting mainly male teens who are otherwise perfectly healthy. The condition makes them withdraw completely, locking themselves in their rooms sometimes for up to twenty years, in extreme cases leaving only occasionally to commit violent crimes, although must sufferers are not violent, just depressive. The Japanese government have described hikikomori as a social disorder rather than a mental disorder, and say it is representative of the economic downturn the country is going through.

- a second syndrome is amok which comes from the Malay people, first described in the fifteenth century as an ‘understandable’ reaction to frustration, but now recognised as a mental disorder requiring treatment. Although the disorder has been described elsewhere, it is most often found in Malay males, beginning with depressive brooding and often following violent outbursts using weapons, commonly homicidal. The turn of phrase ‘running amok’ or ‘running amuck’ came from this, describing the sensation of rage leading to a killing spree.