EVEN MORE APPLICATIONS AND EXAMPLES OF RESEARCH METHODS IN PSYCHOLOGY

Kevin Brewer

ISBN: 978-1-904542-28-5

PUBLISHED BY
Orsett Psychological Services,
PO Box 179,
Grays,
Essex
RM16 3EW

PRINTED BY
Print-Trek,
Upminster,
Essex
RM14 2AD

CONTENTS

STUDYING GENETIC ORIGINS OF BEHAVIOUR: FRUIT FLIES, KNOCKOUT MICE, HUMANS AND SLEEP

The twenty-first century seems to be the age of studying the genetic origins of behaviour. The ever-increasing knowledge about genes fuels more research into genetic origins.

Generally genetic origins of behaviour can be studied in three ways:

i) Manipulation of genes in fruit flies;
ii) Use of "knockout mice";
iii) Genetic history and human case studies.

Each method has advantages and disadvantages, or they can be combined to gain a more detailed picture of behaviour, in this case, sleep.

Generally animal models for human behaviour are used in two ways (Maxson 2003):

a) To identify and map genes with effects on human behaviour: eg alcoholism and mice;

b) To develop hypotheses about the biological causes of human behaviour: eg memory in aplysia (mollusc), fruit flies, and mice.

FRUIT FLIES

The fruit fly (Drosophila) has become the stalwart of genetic research in all areas [1], mainly because they have only four pairs of chromosomes (Taylor 2007) (table 1).

Fruit flies produce a large amount of saliva, and this contains the chromosomes which are one thousand times larger than normal (Brookes 2001).

Fruit fly studies have been most helpful in understanding the genes involved in circadian rhythms over the last thirty years [2]. These genes have names like "period", "timeless", "Clock", and "doubletime" (Wager-Smith and Kay 2000). Findings are then moved to mice, though the circadian process is genetically slightly different in mice to flies.

[1] Database of research at http://flybase.bio.indiana.edu.
[2] First gene ("period") identified by Konopka and Benzer 1971).

Recently, Ganguly-Fitzgerald et al (2006) found differences in length of sleep depending upon social activity during waking. Those flies with an enriched social environment slept longer, particularly in the day, than those with an impoverished one. The length of sleep was linked to pathways in the brain related to learning and memory, and to seventeen different genes (of forty-three tested).

Socially enriched environments contained thirty flies, while socially impoverished flies were kept alone. The effect of interaction on sleep did not occur when groups of vision, olfactory, or hearing impaired flies were used.

Yuan et al (2006) investigated the role of the neurotransmitter, serotonin, in the sleep. Fruit flies with three genetically altered expressions of serotonin receptors were used, and one of these mutations (d5-HT1A) had shorter and fragmented sleep compared to others. The significant differences ($p<0.01$) were:

- Less than 600 minutes per 24 hours of total sleep versus approximately 800 minutes in controls;
- Average length of sleep bouts less than 20 minutes versus nearly 30 minutes in controls;
- Equal amounts of daytime sleep, but less at night;
- Over twenty bouts of sleep per night versus fifteen in controls;
- In situations of constant darkness, 30% reduction in sleep amount.

In a slightly different type of study, using wild-type strains, Cirelli et al (2005) screened 9000 different lines of flies to find one that slept much less than the average (called "minisleep") (eg 4-5 hours per day versus 9-15 hours). The researchers isolated the cause to a recessive [3] gene mutation on the X chromosome through selective breeding over five generations.

Greenspan et al (2001) are optimistic about the use of fruit flies to study sleep: "Once again flies are proving that they are more like us than one might think" (p145).

[3] Recessive genes require both copies before manifesting the effect, while dominant genes only need one copy from either the biological mother or father.

ADVANTAGES

1. Similarity between fruit-fly and mammalian sleep.
Fruit flies show following criteria used to define mammalian sleep: reduced responsiveness to external stimuli, circadian rhythms of sleep and waking patterns, "rebound" after sleep deprivation (ie increased sleep to compensate for loss), age differences (eg young sleep longer, old have fragmented and less sleep), and caffeine increases waking (Greenspan et al 2001).

2. Possible to manipulate genes in a way not possible with humans.

3. Short life allows observation across whole lifespan and many generations.
Males become sexually mature 12 hours after birth and females three days with lifespan of approximately 15 days (five as adults) (Taylor 2007).

4. Limited number of genes to study, and large chromosomes in saliva.

5. Generates findings that can be tested and applied to humans, and opens the way to studying difficult questions.
"Genetics are the shock-troops of biology" (Edgar and Epstein 1965).

6. Can be controlled and kept in lab conditions.

7. Similarity of genes to humans: eg 74% of human disease-causing genes (Reiter 2003).

8. Similarity of biological pathways in invertebrates and vertebrates.

9. Prolific and easy to breed, particularly as genetically identical.

10. Processes found in flies and in humans suggest an evolutionary basis.

11. Allows testing of sleep mechanisms independently of circadian rhythms: ie mutants mean that surgery to parts of the brain is not needed as in mammals.

12. Ability to isolate genes in way not possible with knockout mice (Greenspan et al 2001).

DISADVANTAGES

1. Limited applicability of results to humans, particularly as only four pairs of chromosomes compared to 23 pairs in humans.

2. Ethics of using animals in such ways. However, this is less of an issue because few people feel as strongly against the use of fruit flies as with mammals.
Fruit flies with different genetic variations can be ordered from suppliers' catalogues by stock number and name: eg Yuan et al (2006) used catalogue number/name: e01363/5HT2RB among others.

3. There are genetic differences: eg 1 X sex chromosome produces a male and XX produces a female compared to XX (female) and XY (male) in humans.

4. Only small number of genes will make a difference if genes work by their interactions rather than individually.

5. Genetic manipulation can produce unexpected results, particularly if genes have more than one function or role.

6. Ignores the ability of humans to learn and adapt.

7. Problems in observing sleep in small insects. It requires using visual, infrared, and ultrasound equipment (Greenspan et al 2001).

8. Genes varying in animals may not vary in humans or vice versa: ie same genes but different roles in animals and humans (eg "dunce" gene involved in memory formation in flies, but in mood in humans; Davis 2005).

9. If sleep is a whole brain process, then isolating individual genes is of limited use.

Table 1 - Advantages and disadvantages of using fruit flies to study genetics of sleep in humans.

KNOCKOUT MICE

"Knockout mice" are those animals with specific genes "turned off" in order to see the effect. The gene has been inactivated by replacing it with an artificial piece of DNA. The observed effect is then used to understand the normal role of the gene (table 2).

Breeding programmes then produce more animals with that gene turned off. Newman (2007) noted that the International Mouse Knockout Consortium's desire to have 20 000 different knockout mice (each with one gene turned off) will need a breeding programme of seven million animals to maintain it.

Knockout mice are made by taking embryonic stem cells from a four day old embryo. An artificial piece of DNA is inserted into the cells in a process called gene targeting or homologous recombination, and then the cells are injected back into the embryo (National Human Genome Research Institute 2007). An alternative known as gene trapping places random DNA rather than non-active pieces into the cells.

In terms of studying sleep disorders, Chemelli et al (1999) turned off a gene related to orexin production and produced behaviour in the mice similar to narcolepsy (eg sudden onset of sleep; unusual EEG patterns).

HUMAN CASE STUDIES

Work with humans involves studying specific individuals or groups and then constructing a genetic history to isolate specific genes from blood samples (table 3).

ADVANTAGES

1. Humans and mice share many genes.

2. Saves time for researchers and allows them to focus on the particular gene of interest.

3. Involves studying live animals.

4. Can be controlled and bred relatively easily in labs.

5. Exact details of the effects of the genes can be ascertained from observation from life and from post-mortem studies of the brain.

6. Allows elaboration on observation from studying humans.

7. Can lead to discovery of causes of behaviour (which can then be tested on humans).

8.. Whole lifespan and subsequent generations can be studied.

9. Good way to test hypotheses when it is not known which gene involved in behaviour.

10. Processes found in mice and in humans suggest an evolutionary basis.

11. Human sleep disorders, like advanced sleep phase syndrome, show similarities to lab mice with mutations to a particular gene ("dbt"; "doubletime") controlling circadian rhythms (Wager-Smith and Kay 2000).

DISADVANTAGES

1. Turning off a gene and seeing the effect is not necessarily the same as understanding the gene turned on. Assumptions have to be made about what is happening (ie deducing from the phenotype [4]). Other genes may compensate for the loss in some way.

2. Turning off a gene can produce a different effect in mice than in humans, no effect in mice but in humans, or vice versa.

3. Effects observed may be due to the interaction of genes not just the one gene being turned off.

4. Genetic manipulation can produce unexpected results including pain and distress to the animal.

5. Ethics of using animals in this way, particularly as they are sold as products by bio-engineering companies.
For example, Tafti and Franken (2006) claimed the "creation" of four new strains of transgenic mouse to study sleep.

6. Genetically manipulated animals are not the same as "normal" ones.

7. Human sleep disorders may be genetically complex and not amenable to the single gene approach, and sleep may be a whole brain process.

---

[4] Phenotype is the actual behaviour manifest by the gene.

8. Ignores the flexibility of humans to learn.

9. Some knockouts are lethal (eg 15%; National Human Genome Research Institute 2007), and the mouse does not live to adulthood. Some genes may serve different functions in adulthood than in embryos or infants, and this cannot be established.

10. Gene trapping is a random process and there is no guarantee that anything will happen.

11. Different to flies in some biological processes (eg circadian rhythms), so there must be differences to humans.

12. Same gene but different roles in animals and humans, and genes may have more than one role.

Table 2 - Advantages and disadvantages of using knockout mice to study genetics of sleep in humans.

In the case of the sleep disorder, narcolepsy, for example, first degree relatives (eg mother, father, siblings) of patients with narcolepsy have up to forty times greater risk of developing the condition (Taheri and Mignot 2002). A number of genetic factors have been studied in such families.

Twin studies, both monozygotic (MZ; identical) and dizygotic (DZ; non-identical), can be used. If one twin has a particular behaviour or disorder, how often the other twin has the same thing is calculated as the concordance rate. With MZ twins, who have the same genes, a concordance rate of 100% (1.0) would mean every time one twin has the behaviour or disorder so does the other twin. Thus it would be entirely inherited. A concordance rate of 0% would be the complete opposite. Furthermore, the concordance rate for MZ twins should be higher than of DZ twins if the behaviour or disorder is inherited.

In reality, the concordance rates vary in MZ twins for different aspects of sleep: eg 80% (0.80) for awake-resting EEG patterns, 50% (0.50) for sleepwalking and night terrors (vs 10-15% for DZ twins) (Taheri and Mignon 2002).

Xu et al (2005) constructed a family tree for advanced sleep phase syndrome to show the genetic basis (PER 2 gene). The sufferers fell asleep early in the evening (eg 7-8pm) and woke early the next morning (eg 4-5am). Onset of the condition occurred between early childhood and the mid-teens. In the family, the grandmother was a sufferer, and so were three of her four daughters, and one granddaughter.

Restless legs syndrome (RLS) (or Ekbom's syndrome; Ekbom 1960) is the involuntary movement of the legs

during non-rapid eye movement (NREM) sleep which disrupts the sleep.

In family studies, working with a sufferer (proband), the researchers try to discover how many first degree relatives also have the same condition. A few of the studies use control groups. The frequency of RLS in families of sufferers varies between 40-90% depending on the study (Winkelmann and Ferini-Strambi 2006).

Linkage studies [5] initially suggested genes on chromosome 12 (known as RLS1) in a South Tyrol family and an Icelandic sample, and then chromosome 14 (RLS2) (an Italian family), and 9 (RLS3) (two US families) (Winkelmann 2006; Winkelmann and Ferini-Strambi 2006).

ADVANTAGES
1. Studying humans sleeping is the best way to study human sleep behaviour.
2. Individuals with sleep disorders are better to study than "manufactured" animal versions.
3. Study specialist populations/families where sleep disorders more common.
4. Patient can talk about the experience of sleep or the disorder.
5. Can study patients with EEG or neuroimaging.
6. Can study sleep as a whole brain process.
DISADVANTAGES
1. Not possible to genetically manipulate.
2. Dependent on cases available, particularly for rare sleep disorders.
3. Limited knowledge of cases before came to notice of medical staff or came to the study.
4. Limitations on how they can be studied (eg length of sleep deprivation) compared to animals.
5. Cannot really know what is happening at the microscopic level in the brain.
6. Limitations of EEG and neuroimaging.

Table 3 - Advantages and disadvantages of human case studies to study genetics of sleep in humans.

[5] Linkage studies segregates the family members with or without the condition. It is possible to focus upon a particular loci (position on the chromosome), and to see which copies (allele) exists there for ill family members as opposed to healthy ones.

Humans can also be studied by looking at post-mortem brains (table 4), particularly for those with sleep disorders: eg less orexin neurons in narcoleptics (Thannickel et al 2000).

ADVANTAGES

1. Detailed look inside the brain.

2. Study parts of brain in microscopic detail.

3. Studying humans not animals.

DISADVANTAGES

1. Death may cause change to brain.

2. Confounding variables, like drug addiction, may distort findings.

3. Have to wait for patient to die.

Table 4 - Advantages and disadvantages of post-mortem human case studies to study genetics of sleep in humans.

CONCLUSIONS

Taheri and Mignon (2002) saw the benefits of animal models to study the genetics of sleep, but as complementary to large scale human studies.

Animal models are most effective for understanding disorders. For example, work on the genes that control circadian rhythms in both flies and mice have helped in the understanding of human sleep-wake disorders, like delayed sleep phase syndrome (Wager-Smith and Kay 2000).

REFERENCES

Brookes, M (2001) Fruit fly genetics Guardian 29/3 (Accessed at www.guardian.co.uk on 19/10/07)

Chemelli, R.M et al (1999) Narcolepsy in orexin knockout mice: Molecular genetics of sleep regulation Cell 98, 437-451

Cirelli, C et al (2005) Reduced sleep in Drosophila Shaker mutants Nature 28/4, 1087-1092

Davis, R.L (2005) Olfactory memory function in Drosophila: From molecular to systems neuroscience Annual Review of Neuroscience 28, 275-302

Edgar, R.S & Epstein, R.H (1965) The genetics of a bacterial virus Scientific American 212, 70-76

Ekbom, K.A (1960) Restless legs syndrome Neurology 10, 868-873

Ganguly-Fitzgerald, I et al (2006) Waking experience affects sleep need in drosophila Science 313, 1775-1781

Greenspan, R.J et al (2001) Sleep and the fruit fly Trends in Neuroscience March, 142-145

Konopka, R.J & Benzer, S (1971) Clock mutants of Drosphila melanogaster Proceedings of the National Academy of Sciences, USA 68, 2112-2116

Maxson, S.C (2003) Animal models of human behaviour. In Cooper, D.N (ed) Encyclopaedia of the Human Genome vol 1 London: Nature

National Human Genome Research Institute (2007) Knockout Mice (Accessed at www.genome.gov on 19/10/07)

Newman, C (2007) Top research grants for 2007 Alternative News Autumn, 4-6

Reiter, L.T (2003) Drosophila as a model for human diseases. In Cooper, D.N (ed) Encyclopaedia of the Human Genome vol 2 London: Nature

Tafti, M & Franken, P (2006) Using mice to elucidate genes controlling sleep and wakefulness Journal of Sleep Research 15, supplement 1, s28

Taylor, M (2007) Animal cameo Feedback September, p5

Taheri, S & Mignet, E (2002) The genetics of sleep disorders Lancet Neurology 1, 242-250

Thannickal, T.C et al (2000) Reduced number of hypocretin neurons in human narcolepsy Neuron 27, 469-474

Wager-Smith, K & Kay, S.A (2000) Circadian rhythm genetics: From flies to mice to humans Nature Genetics September, 23-27

Winkelmann, J (2006) Genetics of restless legs syndrome Journal of Sleep Research 15, supplement 1, s28

Winkelmann, J & Ferini-Strambi, L (2006) Genetics of restless legs syndrome Sleep Medicine Reviews 10, 179-183

Xu, Y et al (2005) Functional consequences of a CKI mutation causing familial advanced sleep phase syndrome Nature 31/3, 640-644

Yuan, Q et al (2006) A sleep-promoting role for Drosophila serotonin receptor 1A Current Biology 16, 1051-1062

WATCHING PEOPLE HAVE SEX IN THE NAME OF SCIENCE: CONTROLLED OBSERVATION VERSUS PARTICIPANT OBSERVATION

INTRODUCTION

Observation is the cornerstone of much research. Marshall and Rossman (1989) described it as "a systematic description of events, behaviours and artifacts in the social setting under study".

It allows the researcher to see for themselves, and avoids the bias of participants' self-reports. The question is, then, where should the observation take place - in the researcher's or the participant's normal environment.

The controlled observation (CO) involves the participants coming to the lab and being observed there (non-participant observation), while participant observation (PO) has the researcher going to the participant's normal environment and becoming part of that social world. Both methods have been used to study sexual behaviour, and, as with all methods, there are strengths and weaknesses (table 5).

TYPE OF OBSERVATION	ADVANTAGE	DISADVANTAGE
Controlled	Allows controlled measurement of behaviour	Participants are taken out of their normal environment
Participant	Takes place in participant's normal environment	Effect on situation of presence of observer as participant

Table 5 - Main advantage and disadvantage of CO and PO methods.

CONTROLLED OBSERVATION

The CO is an observation that takes place in a laboratory situation. It is not an experiment because there is no manipulation of the independent variable.

Masters and Johnson (1966) were the first to study human sexual arousal with lab observations. A number of physiological measures were used including heart rate (electrocardiogram), muscle tension (electromyogram), and internal vaginal changes (using artificial coital

equipment). Initially a small group of prostitutes were used, but later 382 female and 312 male community volunteers were studied. The age ranged from 18 to 89 years. In total, over 10 000 "orgasmic sexual responses" were studied.

The researchers dealt with a number of concerns from the research:

- The motivation of volunteers to have sex "in public" - Extensive pre-study interviews;
- Confidentiality and anonymity - Masters and Johnson found that the part of the hospital where they had a research suite became very busy with passers-by in the corridor. There was a lot of curiosity to see who was taking part. So the research was switched to "outside office hours" to avoid sightseers
- The problem of performing "in public" - There were practice sessions in the research suite to overcome nervousness.

Table 6 summarises the strengths and weaknesses of the use of the CO to study sexual behaviour.

PARTICIPANT OBSERVATION

The PO method allows the researcher to observe the participants in their normal environments, while, often, hiding their identity as a researcher.

Patton (1980) distinguished four types of participant observation based on the relationship between being an observer and being a participant:

i) Full participant (sometimes called active participant) - The researcher is known as a group member only. Thus their identity as a researcher is known by no one in the group studied;

ii) Participant as observer - The researcher's identity is known to some members of the group, like the leaders, only.

iii) Observer as participant (sometimes called passive participation) - The researcher's identity is known to the group as they join in the group's activities;

iv) Full observer - This is non-participant observation.

ADVANTAGES	DISADVANTAGES
1. Allows physiological measures that cannot be gained in everyday situations. 2. Allows the use of sophisticated equipment to measure aspects of behaviour. 3. Data are collected as they happen, and from multiple sources (eg: machines and human observers). 4. More flexible than an experiment which has to maintain strict design controls in all conditions. 5. Stricter control than the naturalistic observation, and some form of replication possible. 6. Overcomes problems of honesty of self-reported questionnaires. 7. Can observe details that participant may not know themselves (eg: how body changes during orgasm). 8. Use of volunteers overcomes ethical issues of invasion of privacy.	1. Individuals know they are being studied, and this may change their behaviour in some way (known as participant reactivity). 2. Individuals who volunteer to be studied during sexual activity are not typical of the general population. 3. Low ecological validity. It is not the same as being in the normal environment. 4. Focus upon physiological workings of body and overt behaviour during sex, but not the meanings to the individual. 5. Expensive and time-consuming compared to questionnaires. 6. Ethics of watching individuals involved in sexual activity. 7. Not an experiment, so limited ability to establish cause and effect. 8. Practical problems like having sexual intercourse while attached to various physiological measures.

Table 6 - Advantages and disadvantages of the CO to study sexual behaviour.

A classic study using the PO method is Humphreys (1970). He studied the behaviour of men engaged in anonymous sex with other men in public toilets (called "tearooms" by these men).

Initially, he tried non-participant observation by visiting the toilets, but the "tearoom trade" stopped when he arrived. To overcome this problem, Humphreys became a "watchqueen" (a look-out for police officers and strangers as this behaviour was illegal at the time).

Thus Humphreys could be in the situation to observe it without causing the participants to behave differently. This was crucial because it is a very secretive behaviour not usually open to research scrutiny.

Table 7 summarises the advantages and disadvantages of using PO to study sexual behaviour.

ADVANTAGES	DISADVANTAGES
1. Allows researcher to study behaviour that is usually hidden and secret. 2. Because the researcher is accepted as a member of the group, there is less reactivity by participants. 3. Builds relationships that allows deeper insights. 4. High ecological validity. 5. Does not involve manipulation of variables as in an experiment. 6. Allows the researcher to see the behaviour in context and build a holistic picture rather than focusing on specific variables. 7. Can lead to more specific hypotheses for future research.	1. The presence of the observer as a participant may change the situation. 2. The observer can become involved in the situation, and thereby less objective. 3. The ethics of hiding identity of researcher from participants (ie deception). Also those being observed do not the right to non-participation in the research, or give informed consent. 4. If participants know researcher is in their midst, they may change their behaviour. 5. The ethics for the researcher of being involved in an illegal activity. 6. Sometimes information cannot be recorded until later, and the researcher may have forgotten something (risk of observer bias). 7. There is only the researcher's report of events, and no independent verification.

Table 7 - Advantages and disadvantages of PO to study sexual behaviour.

ETHICAL ISSUES

All research involves the need for awareness of ethical issues, and both these types of observation have particular ethical concerns (table 8). Participant observation has more concerns including deception as the researcher hides their identity, lack of informed consent, and no right to non-participation in the research.

ETHICAL ISSUE	CO	COVERT PO
Informed consent	Gained	Not gained
Deception	Not problem	Researcher hiding identity and purpose
Debriefing	Given	Not possible
Right of non-participation	Yes	No
Right of withdrawal	Yes	No
Confidentiality	Identity of participants hidden by use of numbers not names	Some; depends how well researcher gets to know participants
Invasion of privacy	Use of volunteers	Invasion
Risk and distress	Participants fully aware of what volunteering for	Embarrassment of being observed

Table 8 - Ethical issues of CO and PO when studying sexual activity.

Generally it is accepted that informed consent is not required for observations in a public place, but, though Humphreys was observing public toilets, the behaviour was private.

Many researchers argue that deception is a necessary evil in order to gain accurate results. This has led the American Psychological Association (1973) to list five conditions that may make deception acceptable:

- The result problem is of great importance;
- The research cannot be accomplished without deception;
- There is sufficient reason to believe that the

participants will not be distressed when later findings out about the deception;
- The participants still have the right to withdraw from the research at any time;
- The researcher takes full responsibility for removing any stressful after-effects of the research.

CONCLUSIONS

The type of observation used will depend upon what the researcher is seeking to discover. The controlled observation of Masters and Johnson is best for studying the physiological aspects of sexual behaviour, while PO allowed Humphreys to discover more about the secretive world of men who have sex with men (MSM).

REFERENCES

American Psychological Association (1973) APA Ethics Guide Washington DC: APA

Humphreys, L (1970) Tearoom Trade London: Duckworth

Marshall, C & Rossman, G.B (1989) Designing Qualitative Research London: Sage

Masters, W.H & Johnson, V.E (1966) Human Sexual Response Boston: Little Brown

Patton, M.Q (1980) Qualitative Evaluation Methods London: Sage

RELIABILITY AND VALIDITY OF PSYCHOMETRIC INSTRUMENTS: THE EXAMPLE OF STUDENT EVALUATION OF TEACHING EFFECTIVENESS

INTRODUCTION

Psychometric questionnaires are based upon the assumption that the answers given and the consequent scores gained are accurate measures of behaviour. In order to achieve this, such instruments undergo rigorous reliability and validity establishing.

This article outlines the main types of reliability and validity for psychometric instruments by looking at the example of student evaluation of teaching effectiveness (SETE). In the USA, in particular, students are asked to rate their lecturers in terms of quality of teaching, for example. Psychometric questionnaires used to do this have to show that the scores gained for lecturers are an accurate measure of teaching effectiveness rather than just the students' unsubstantiated opinions. This is the establishing of validity (ie the questionnaire measures what it claims to measure). While reliability is the consistency of the questionnaire.

Theall and Franklin (2001) are confident about students' ratings: "No one else is as qualified to report on what transpired during the term simply because no one else is present for as much of the term" (p48).

Marsh (1984) outlined the purpose of student evaluation of teaching as fourfold:

i) As a "diagnostic feedback to faculty about the effectiveness of their teaching" (p707);

ii) For use in teacher promotion decisions;

iii) As information for students in selection of courses;

iv) To show the "outcome on a process description for research or teaching" (p707).

RELIABILITY OF STUDENT RATINGS

Here reliability refers to the fact that the ratings will measure the same score every time, ie the same lecturer producing the same quality lecture on two occasions will receive the same rating by the same student.

Table 9 shows the types of reliability for psychometric questionnaires.

TYPE OF RELIABILITY	DESCRIPTION
Internal 1. Split-halves a. odd and even questions b. all possible split-halves 2. Parallel/multiple forms a. with time interval b. without time interval 3. Item analysis	 - Correlation between scores on two halves of test - Correlation between scores on two versions of same test - Ability of each item to discriminate between high and low scorers
External Test-retest - immediate or later	 - Correlation of scores between same test repeated at different times

Table 9 - Different types of reliability for psychometric instruments.

Doyle (1975) listed the sources of reliability errors:

i) Computational error - eg putting the wrong instructor's name on ratings summary.

ii) Rater's task - ie problem with nature of the questions used.

iii) Environment - physical or social environment.

iv) Rater - lacks motivation or memory problems, as well as:

- Halo effect: overall impression influences specific rating items.
- Leniency error: tendency to rate higher when known that ratings being used for promotion purposes.
- Central tendency: inclination for mid-point on scale.
- Proximity error: rate adjacent items similarly.
- Contrast error: projection of own deficiencies on to ratee.
- Logical error: rating traits that "ought" to go together.

The first study of reliability came from Guthrie (1927). Two hundred and eighty-five psychology students ranked lecturers at the University of Washington, and

then again two weeks later. A correlation of r = 0.89 was found.

In Britain, Foy (1969) followed up his study with Cooper (Cooper and Foy 1967), due to objections about the original findings on an ideal lecturer. A different group of students used the same check-list as the first study, and there was a correlation of 0.93 between the two ratings (1 in 2000 possibility of a chance correlation as high as that). This seems the most straightforward evidence of the reliability of an instrument.

Methods Used to Establish Reliability

1. Internal Consistency

The aim is to correlate various questions within the instrument. Using for example, odd-even or split-half, and coefficient alpha (Cronbach 1951) or Kuder-Richardson formulas (Kuder and Richardson 1937).

Feldman (1977) reports an extension of this approach, where two mean scores for a particular item can be obtained by randomly dividing a class in half. Most of the commonly used instruments report reliability coefficients over 0.50.

2. Test-Retest

Here the rating instrument is given to the same subjects at two different times. The aim being to correlate the two scores of each subject.

But the instructor may change between administrations of the instrument, and so a small correlation will suggest that the instrument is unstable. This method is also criticised for "being a test of the student's memory instead of being a measure of reliability" (Frey 1978 p85).

3. Mean Ratings

It is assumed that mean ratings of instructors should be different, because the instructors display different teaching behaviour. If the means are similar or identical, the ratings are seen as biased.

But the assumption that instructors do differ is open to question.

Frey (1978) used a variation of this method. He chose a sample of the data representing instructors who had taught three or more classes (with 10 + students in

each), which had filled in ratings. Variance estimates were calculated for differences among instructors, and differences among classes within instructors - inter rater agreement. A formula used showed the proportion of observed variance due to differences in instructor.

4. Analysis of Variance (ANOVA)

Proposed by Guilford (1954): rather than attempting to remove potential bias, it aims to identify the contribution of bias to the final rating, and adjust for it. Obviously, this has advantages because some potential biases cannot be easily separated, like the halo effect.

Treffinger and Feldhusen (1970), using this method, found that the halo effect only accounted for 10% of the variance in students' ratings (quoted in Doyle 1975 p43).

5.Inter-Rater Reliability

This looks at the consistency of ratings among people. Reliability here is when all raters in a group give the same pattern of responses. Usually estimated by intraclass correlation coefficients, ie the comparison of ratings within one class of one lecturer with ratings of different instructors. Because it is sensitive to the number of raters, Centra (1979) suggested intraclass correlations of .70s for ten raters through to .90s for twenty.

Feldman (1977) makes a number of points about interpreting the reliability coefficients:

i) "reliability coefficients of individual ratings indicate the degree of general or relative consistency among raters; they do not measure exact or absolute agreement" (p229);

ii) inter-rater agreement is only the degree to which independent raters give the same rating for the same lecturer;

iii) inter-rater reliablity is "the degree to which the ratings by different raters are proportional when expressed as deviations from their means" (p229);

iv) the reliability coefficients of average college student ratings may be high, but this does not mean that individual students within the classes are highly consistent in their ratings;

v) consistency in ratings among students may not be a good basis for estimating individual ratings or average

ratings reliability, particularly if the aim is to compare ratings across situations.

Guthrie (1927) suggested that student ratings agree at the end of the term because of greater exposure to the lecturer, or student gossip.

Overall, establishing reliability of a psychometric questionnaire is probably easier than establishing validity.

VALIDITY OF STUDENT RATINGS

Do students know a good lecturer, ie are student ratings actually measuring good teaching? Here validity means that the ratings are an accurate assessment of teaching quality, not other factors, like class size or personality of student. There are different types of validity (table 10).

TYPES OF VALIDITY	DESCRIPTION
Face	Based on commonsense; the items appear valid
Content	Sophisticated version of face validity; experts see the items as valid
Criterion a. Predictive b. Concurrent	 a. Correlation of test score with future performance b. Correlation of test score with another test of same thing
Construct	Correlation of test score with behavioural measure of same thing
Discriminant	Correlation of test score with different measures of same behaviour (some expected and some unexpected); extension of construct validity

Table 10 - Different types of validity for psychometric questionnaires.

McBean and Al-Nassri (1982) noted that "students strongly believed that student evaluations do measure teacher effectiveness ... while faculty only slightly agreed" (p278). This statement can be said to show face

validity. Some would argue, though, that this is only valid as an indicator of student satisfaction.

Criterion Validity

This concentrates on the relationship of ratings with other objective measures. The most common measure used is student learning (usually defined as the grade in the course examination).

In a now famous study in "Science", Rodin and Rodin (1972) found a negative correlation between the amount learned from classes, and their rating of the teacher. The Rodins used a subjective rating of the lecturer, and an objective measure of the amount of calculus learned. The conclusion of r = -0.75 correlation threatened the validity of students' evaluation ratings.

But subsequent studies have consistently found positive correlations. Frey (1978) outlined a number of problems with the Rodins study - for example, study based on teaching assistants rather than teachers who gave the main lectures. Further on in his article, after reviewing the studies since Rodins, Frey pointed out the need to study the "regular instructors", and to use " a rating form which emphasises the appropriate teaching traits" (p75).

Frey (1978) in testing the validity of the two dimensions of "skill" and "rapport" of the Instructional Rating Card (Frey et al 1975), correlated each with examination scores. Using a course divided into multiple sections, taught by different instructors, but with a common syllabus, textbook, and examination. The median correlations were different: for the "skill" factor, it was r = 0.81 but for "rapport" it was r = 0.29. "The two rating factors are clearly not the same in their ability to indicate which teachers were most effective in preparing their students for the final examination" (p87) [6].

Doyle (1983) has his problems with using a student achievement test as the criterion for establishing the validity of student ratings of instruction:

i) some characteristics of teaching are not linked to test scores - eg "clarity" and "rapport";

ii) it is assumed that the relationship is a linear one and thus the Pearson product-moment correlation can

[6] More recently, Spooren and Mortelmans (2006) have identified an underlying factor called "teacher professionalism".

be used. But it is possible that it is a non-linear relationship between student achievement and student ratings of instruction;

iii) which unit of analysis should be used:

a) pooled within-class analysis (individual ratings in each section of the course, and average across course);
b) between-sections analysis (mean ratings of evaluation items across course);
c) total-class approach (individual ratings).
Doyle prefers the first approach;

iv) if participants are randomly divided into sections of the course, then the generalizability of findings are limited.

Emery et al (2003) revisited many of these problems and others.

The main alternative to final grade is to use students' gains in knowledge. But there are problems in how to measure the gain.
Marsh and Overall (1980) tried to combine both criteria. They used final examination grade, ability to apply course material, and inclination to pursue the subject further. The first is seen as a cognitive criterion, while the other two are self-reported affective criteria. The students used were taking a course in computer programming. The authors, accepting methodological weaknesses, felt that more than one construct must be used to establish validity. "Therefore, because there is no universally accepted criterion of effective teaching, the validation of any teaching effectiveness measure must focus on a wide range of indicators" (p474).

Obviously, the higher the correlation, the better for validation. But validity will be specific to a particular situation, and "must always be evaluated in relation to a situation as similar as possible to the one in which the measure is to be used" (Thorndike and Hagen 1977 p69).

Construct Validity

For some researchers, criterion validity is not a satisfactory method to establish the validity of student ratings of instruction because effective teaching is a construct. Thus for them construct validation is the best method.
The main aim is to correlate multiple indicators of

effective teaching. For example, student ratings and various criteria assessed for convergent and discriminant validity.

Howard et al (1985) used this method to establish teaching effectiveness using student ratings, colleagues ratings, teacher self-ratings, former-student ratings, and trained observers. Ratings by current and former students were most effective.

The main criteria used in construct validation are self-evaluation by the lecturer, colleagues' evaluation, external observers, administrators, former students' evaluations, and the research productivity of lecturers.

1.Lecturer Self-Rating

There is a general tendency for instructors to rate themselves more favourably than their students do. But there is agreement on instructor's strengths and weaknesses.

Centra (1972) found differences also between faculties: instructors in natural sciences rated effort needed for their course less than did the students, while education, business, home economics, and nursing instructors were the opposite.

Marsh (1982), quoting his own studies, found correlations of r = 0.41 for undergraduate ratings, and r = 0.39 for postgraduate ratings, with lecturer's self-evaluation.

2.Ratings by Colleagues

In their early literature review, Costin et al (1971) found correlations between 0.30 and 0.63 for students' ratings and colleagues' ratings.

But in most cases, colleagues' ratings are not based on sitting through the lecture, but on "student hearsay, on the observation of the presumed effects of instruction ... and on inferences from their personal acquaintances (with the colleagues)" (Guthrie 1949 p113).

Ballard, Reardon and Nelson (1976) found correlations that ranged from 0.62 to 0.84. Studies based on colleagues actual visitation to the classroom are limited.

Furthermore, there is the problem that the presence of an observer can change the classroom situation - for example, by effecting the performance of the lecturer.

Murray (1980) argued that peer ratings are "less sensitive, reliable and valid" (p45) than student ratings.

3.Observation by External Observers

Murray (1980) felt that student ratings "can be accurately predicted from outside observer reports of specific classroom teaching behaviours" (p31). The feeling is that trained observers are best, and particularly if they concentrate on specific behaviour (eg clarity-related behaviour: number of false starts or halts in speech, redundantly spoken words, and tangles in words) (Marsh 1984).

4.Administrators' View

Cotsonas and Kaiser (1962) used clinical students in a medical school, and compared their ratings with departmental administrators. The former tended to stress the attitude towards students, and teaching skill, while the latter stressed knowledge. The authors suggested that the administrators noted the knowledge of the lecturer, and then assumed the other abilities ("halo effect"). It would also seem that the administrators took into account more than just classroom behaviour, but also their general judgments about the lecturer.

5.Retrospective Ratings of Alumni

Graduating students were asked to nominate "most outstanding" and "least outstanding" lecturers in their departments. Then undergraduates were asked to rate the nominated lecturers. Results indicated that the "most outstanding" lecturers were rated higher than the "least outstanding". A correlation of r = 0.82 between graduates' and undergraduates' choices of most and least outstanding (Marsh 1977).

Gaski (1987) urged caution when using former students' ratings for validity purposes because "the similarity between the student and former student teaching evaluations can be explained if the primary determinant of the former student ratings is former students' recollection of the assessment they made when they were current students of the given instructor one or two years earlier" (p329).

6.Research Productivity

Blackburn (1974) suggested research and effective teaching were opposites. For example, McDaniel and Feldhusen (1970) found significant a negative correlation between first authorship of books and students' ratings of teaching. But a significant positive correlation

between second authorship of professional articles and rating of teaching.

Marsh (1984) finds no correlation or a small positive correlation between the two. "Although these findings seem to neither support nor refute the validity of student ratings, they do demonstrate that measures of research productivity cannot be used to infer teaching effectiveness or vice versa" (p729).

Feldman (1989) undertook a detailed literature review of the North American studies comparing overall ratings of teaching effectiveness made by current and former students, lecturers' colleagues, administrators, external (neutral) observers, and teachers' self-evaluation. The results are summarised in table 11.

Feldman concluded that there was similarity between various raters, in this order: current students and colleagues; current students and administrators; colleagues and administrators (similar in relative assessment, but not in absolute assessment); self-evaluation and current students; self-evaluation and colleagues. For the other relationships, there were not enough studies to determine.

Berk (2005), more recently, extended this type of analysis using twelve sources of evidence.

Method Used	Current Students	Former Students	External Observers	Colleague	Adminis-trators
Current Students		+.69(6)*	+.50(5)*	+.55(14)*	+.39(11)*
Former Students			+.08(1)	+.33(1)	no cases
External Observers				-.12(1)	no cases
Colleague					+.48(5)*
Administrators					

(* = significant correlation p<0.001 two-tailed. The number in () is number of studies found)

Table 11 - Summary of the studies found by Feldman (1989) showing a correlation between different methods of assessing teaching effectiveness.

Use of MTMM

A number of criteria are used under the heading of the Multi-Trait Multi-Method (MTMM) approach (Campbell and Fiske 1959). The use of a number of methods to measure one trait/construct allows correlations to be

made; thus producing a MTMM matrix. It allows the estimation of variance due to traits or methods, and of unique or error variance.

It is possible to show convergent validity (correlation between items that should go together) and divergent validity (small or no correlation between items that should not go together). This method allows the research to estimate the effects of bias; for example, method bias (large correlation between variables because of the method used).

Murphy and Davidshofer (1988) summarised three points that a test will possess as established effectively by MTMM:

> 1.Scores on the test will be consistent with scores obtained using other measures of the same construct.
> 2.The test will yield scores that are not correlated with measures that are theoretically unrelated to the construct being measured.
> 3. The method of measurement employed by the test shows little evidence of bias (p106).

In their original article, Campbell and Fiske proposed a series of rules to follow for evaluating convergent and discriminant validity:

> 1. The convergent validity coefficients should be statistically significant and sufficiently different from zero to warrant further examination of the validity.
> 2. The convergent validities should be higher than correlations between different traits assessed by different methods.
> 3. The convergent validities should be higher than correlations between different traits assessed by the same method.
> 4. The pattern of correlations between different traits should be similar for each of the different methods (quoted in Marsh and Hocevar 1983 p233).

The above rules have been criticised. Firstly, over what constitutes a satisfactory result.

Secondly, the use of correlations based on observed variables to draw conclusions about underlying factors (Kenny and Kashy 1992).

Attempts have been made to establish validity by using large multi-section courses, where different groups of students are presented the same material by different instructors.

Ideally the following controls should be used:

- many sections to the course;

- random assignment of students to the sections;
- pre-test measures used;
- each section taught by separate instructors;
- the final examination graded externally;
- common textbooks among the sections (Marsh 1984).

Validity is then assessed by correlating the student ratings in each section.

CONCLUSIONS

Centra (2003) believed that SETE instruments are reliable and stable, and valid when compared with student learning.

The question of establishing validity has become a methodological issue debated in the literature, particularly around the use of criterion validity (established through multi-section courses) or construct validity (established using MTMM).

However, taking into account the weaknesses of the use of the different criteria, it is fair to say that student ratings of instruction are valid. But the criteria used are validity measures of what?

Feldman (1977) looked at the purpose of the ratings - if it is to obtain objective descriptions of teachers, there may be a problem, but not if it is to measure students' subjective responses.

REFERENCES

Ballard, M; Reardon, J & Nelson, J (1976) Student and peer rating of faculty Teaching of Psychology 3, 88-90

Berk, R.A (2005) Survey of twelve strategies to measure teaching effectiveness International Journal of Teaching and Learning in Higher Education 17, 1, 48-62

Blackburn, R.T (1974) The meaning of work in academia. In Doi, J (ed) Assessing Faculty Effort San Francisco: Jossey Bass

Campbell, D.T & Fiske, D.W (1959) Convergent and discriminant validation by the MTMM matrix Psychological Bulletin 56, 81-105

Centra, J.A (1972) Two Studies on the Utility of Student Ratings for Improving Teaching SIR Report no.2; Princeton, NJ: Educational Testing Service

Centra, J.A (1974) The relationship between student

and alumni ratings of teachers Educational and Psychological Measurement 34, 321-325

Centra, J.A (2003) Will teachers receive higher student evaluations by giving higher grades and less course work? Research in Higher Education 44, 5, 495-518

Cohen, P.A (1981) Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies Review of Educational Studies 51, 3, 281-309

Cooper, B & Foy, J (1967) Evaluating the effectiveness of lectures Universities Quarterly 21, 2, 182-185

Costin, F; Greenough, W.T & Menges, R.J (1971) Student ratings of college teaching: reliability, validity, and usefulness Review of Educational Research 41, 511-535

Cotsonas, N.J & Kaiser, H.F (1962) A factor analysis of students' and administrators' ratings of clinical teachers in a medical school Journal of Educational Psychology 53, 219-223

Cronbach, L.J (1951) Coefficient alpha and the internal structure of tests Psychometrika 16, 297-334

Doyle, K.O (1975) Student Evaluation of Instruction Lexington, Mass: Lexington Books

Doyle, K.O (1983) Evaluating Teaching Lexington, Mass: Lexington Books

Emery, C.R; Kramer, T.R & Tian, R.G (2003) Return to academic standards: A critique of student evaluation of teaching effectiveness Quality Assurance in Education 11, 1, 37-46

Feldman, K.A (1977) Consistency and variability among college students in rating their teachers and courses: a review and analysis Research in Higher Education 6, 223-274

Feldman, K.A (1989) Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers Research in Higher Education 30, 2, 137-194

Foy, J (1969) A note on lecturer evaluation by students Universities Quarterly 23, 3, 345-349

Frey, P.W (1978) A two-dimensional analysis of

student ratings of instruction Research in Higher Education 9, 69-91

Frey, P.W; Leonard, D.W & Beatty, W.W (1975) Student ratings on instruction: Validation research American Educational Research Journal 12, 435-447

Gaski, J.F (1987) On "construct validity of measures of college teaching effectiveness" Journal of Educational Psychology 79, 3, 326-330

Guilford, J.P (1954) Psychometric Methods (2nd ed) New York: McGraw-Hill

Guthrie, E.R (1927) Measuring student opinion of teachers School and Society 25, 175-176

Guthrie, E.R (1949) The evaluation of teaching Educational Record 30, 109-115

Howard, G.S; Conway, G.C & Maxwell, S.E (1985) Construct validity measures of college teaching effectiveness Journal of Educational Psychology 77, 2, 187-196

Kenny, D.A & Kashy, D.A (1992) Analysis of the MTMM matrix by confirmatory factor analysis Psychological Bulletin 112, 1, 165-172

Kuder, G.F & Richardson, M.W (1937) The theory of the estimation of test reliability Psychometrika 2, 151-160

McBean, E.A & Al-Nassri, S (1982) Questionnaire design for student measurement of teaching effectiveness Higher Education 11, 273-288

McDaniel, E.D & Feldhusen, J.F (1970) Relationships between faculty ratings and indexes of service and scholarship Proceedings of the 78th Annual Convention of the American Psychological Association 5, 619-620

Marsh, H.W (1977) The validity of students' evaluations: classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors American Educational Research Journal 14, 4, 441-447

Marsh, H.W (1982) SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching British Journal of Educational Psychology B52, 77-95

Marsh, H.W (1984) Students' evaluations of university teaching: dimensionality, reliability,

validity, potential biases, and utility Journal of Educational Psychology 76, 5, 707-754

Marsh, H.W (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions to future research International Journal of Educational Research 11, 253-388

Marsh, H.W & Hocevar, D (1983) Confirmatory factor analysis of MTMM matrices Journal of Educational Measurement 20, 3, 231-248

Marsh, H.W & Overall, J.U (1980) Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria Journal of Educational Psychology 72, 4, 468-475

Murphy, K.R & Davidshofer, C.O (1988) Psychological Testing: Principles and Applications Englewood Cliffs, NJ: Prentice Hall

Murray, H.G (1980) Evaluating University Teaching: A Review of Research Toronto: Ontario Confederation of University Faculty Associations

Rodin, M & Rodin, B (1972) Student evaluation of teachers Science 177, 1164-1166

Spooren, P & Mortelmans, D (2005) Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? Educational Studies 32, 2, 201-214

Theall, M & Franklin, J (2001) Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? New Directions in Institutional Research 27, 5, 45-56

Thorndike, R.L & Hagen, E (1977) Measurement and Evaluation in Psychology and Education (4th ed) New York: John Wiley

THE USE OF COMPLEX EXPERIMENTS IN SOCIAL PSYCHOLOGY

Students are taught about the experiment in social psychology based around a single independent variable (IV) and a single dependent variable (DV). In reality, research involves more complex experiments with multiple IVs at the same time, or many parts or stages to the experiment.

Complex experiments have their own strengths and weaknesses in relation to simple, single IV experiments (table 12).

ADVANTAGES	DISADVANTAGES
1. General advantages of the experiment - control of variables - establishing causality - replication 2. Allows study of complex behaviour beyond single variables 3. In reality, a number of factors influence behaviour and having more than one IV can reflect this 4. Performing one experiment with two IVs saves time and needs less participants than two experiments with one IV each 5. Complex experiments with many parts can hide the real purpose from participants and thereby reduce demand characteristics	1. General disadvantages of the experiment - artificial situation - demand characteristics - experimenter effects 2. Assumptions made about each IV when more than one IV involved 3. Concerns about validity of measures 4. Loss of precise control as complexity increases 5. Many parts to the experiment may produce problems, like fatigue or drop-outs, particularly if the experiment takes a long time

Table 12 - Advantages and disadvantages of using complex experiments in social psychology.

TWO EXAMPLES FROM THE RESEARCH INTO ATTITUDES AND DECISION-MAKING

1. Sanbonmatsu and Fazio (quoted in Fazio 1990)

This research is based around the accessibility of attitudes and how they influence decision-making and behaviour.

Fazio (1990) developed the idea of automatic accessibility of attitudes in his MODE model (motivation, opportunity and determinants). Motivation and opportunity determine whether spontaneous or deliberate processing occurs. In other words, whether there is consideration of the link between attitudes and behaviour depends on motivation (eg cost of inconsistency) and opportunity (eg immediate decision required).

Sanbonmatsu and Fazio designed a complex lab experiment to test this theory. Participants were told about two department stores, Smith's and Brown's. Most of the information about Smith's was positive, and most about Brown's was negative. But then participants were told about the camera department in each store using the opposite criteria (ie mostly negative about Smith's and positive about Brown's). The task was to buy a camera.

Under spontaneous processing, the participants would choose Smith's because the general information was positive, and the individual does not think about their behaviour in detail here. But under deliberate processing, Brown's would be the choice because the individual is thinking about their behaviour more (table 13).

	BROWN'S	SMITH'S
General information	negative	positive
Camera department	positive	negative
Type of processing in choice	deliberate	spontaneous

Table 13 - Deliberate and spontaneous processing.

But the experiment also varied the motivation (had to explain decision to group or not) and opportunity (make immediate decision or time for reflection). Only in the condition which allowed deliberate processing did participants choose Brown's store to buy their camera (table 14).

	CONDITION 1	CONDITION 2
Motivation	high	low
Opportunity	reflection	immediate decision
Type of processing	deliberate	spontaneous
Decision	Brown's	Smith's

High motivation = explain decision to group; low = no explanation for decision

Table 14 - Different conditions in the experiment by Sanbonmatsu and Fazio.

Evaluation

i) The IVs were:

- Nature of information: positive general information and negative camera department, negative general information and positive camera department, and two controls (positive/positive and negative/negative);
- High and low motivation;
- Opportunity for reflection or not.

ii) Assumptions are made about the cognitive processes involved based on the decision taken. Individuals may have arrived at certain decisions but for different reasons to those suggested by the researchers.

iii) Use of artificial scenarios. Individuals may behave differently in a real-life situation of buying a camera. Furthermore, attitudes to particular stores may not be simply positive or negative, but a combination based on other factors as well. In other words, attitudes are on a dimension rather than being a dichotomy.

iv) The more complex the experiment becomes, there is a greater chance of participants realising what is happening and demand characteristics occurring.

v) Statistical analysis in complex experiments usually involves analysis of variance (ANOVA) rather than two-condition tests of difference like the Mann-Whitney U test.

2. Perugini (2005)

Recent research into attitudes has distinguished between explicit and implicit attitudes (Wilson et al 2000). Explicit attitudes are those that an individual expresses and is aware of, while implicit attitudes are automatic (and often outside of conscious awareness).

One commonly used way of measuring the latter is the Implicit Association Test (IAT) (Greenwald et al 1998).

This is a computer-based test using response times to pressing certain keys when words appear on the screen. Pairs of words (which contrast like dog and spider) are presented on the screen, and the task is to press a right side key for dog or pleasant words and a left side key for spider or unpleasant words. The keys are then changed, so that right is for dog or unpleasant and left for spider and pleasant, and the same words are presented again.

If the participant is slower in their reaction time in the second version (ie spider/pleasant and dog/unpleasant) compared to the first version, this is taken as evidence of an implicit positive attitude towards dogs relative to spiders.

Perugini (2005) performed two separate experiments to test implicit and explicit attitudes and behaviour [7].

This involved measuring implicit and explicit attitudes towards smoking among twenty-five smokers and twenty-three non-smokers at the University of Essex. Explicit attitudes towards smoking and exercise were measured by a semantic differential scale containing eleven pairs of adjectives (eg bad-good, calming-stressful, glamorous-ugly).

Implicit attitudes were measured by the IAT (using five steps - three being practice ones). In the first experimental step, the left key was used for words related to smoking (eg ashtray, lighter) or pleasant (eg rainbow, smile), and the right key for exercise (run, swim) or unpleasant (eg pain, vomit) related words (figure 1).

In the final step, the right and left keys were switched. The IAT was calculated by taking the difference in reaction times between the two experimental steps, and the score showed an implicit positive attitude towards smoking compared to exercise.

Smokers had both more positive implicit and explicit attitudes towards smoking than non-smokers. But only explicit attitudes predicted whether the individual was a

[7] Only experiment 1 is described here.

smoker or not (behaviour). This was especially true for non-smokers.

SMOKING RELATED WORDS press "q"	EXERCISE RELATED WORDS press "p"
PLEASANT ASSOCIATED WORDS press "z"	UNPLEASANT ASSOCIATED WORDS press "m"

Figure 1 - How computer screen looked in Perugini (2005) experiment.

Evaluation

i) This experiment is complex not because of the number of IVs, but because of the complexity of the procedure for measuring implicit attitudes. Participants spent time learning how to respond on the computer and their mean reaction times were measured. Having five steps in the IAT allows the possibility of order effects.

This is where performance on the first part of an experiment influences performance on the second part through fatigue or boredom (slowing down performance) or practice (improving performance).

ii) Concerns over the validity of the IAT. This is whether the test is measuring what it claims to measure ie hidden attitudes. The assumption is made that slower reaction time is evidence of implicit attitudes.

iii) Measuring implicit attitudes also has two other problems (MacDonald et al 2005):

a) The scales used have low reliability compared to explicit attitude scales;

b) The context in which the attitude object is perceived can influence the implicit attitude in tests like the IAT.

iv) The use of a computer programme in the experiment gives the researcher accurate reaction time measurements which could not be made by hand. It also allows the collection of a large amount of data, and analysis for the difference in reaction times.

v) Because participants were recruited openly as smokers and non-smokers, it gives the possibility that they could have realised the purpose of the experiment and changed their behaviour. However, it is claimed that the IAT overcomes deliberate manipulation of behaviour by participants.

REFERENCES

Fazio, R.H (1990) Multiple processes by which attitudes guide behaviour: The MODE model as an integrative framework Advances in Experimental Social Psychology 23, 75-109

Greenwald, A.G; McGhee, D.E & Schwartz, J.K.L (1998) Measuring individual differences in implicit cognition: The implicit association test Journal of Personality and Social Personality 1464-1480

MacDonald, G; Nail, P.R & Levy, D.A (2004) Expanding the scope of the social response context model Basic and Applied Social Psychology 26, 1, 77-92

Perugini, M (2005) Predictive models of implicit and explicit attitudes British Journal of Social Psychology 44, 29-45

Wilson, T.D; Lindsey, S & Schooler, T (2000) A model of dual attitudes Psychological Review 107, 101-126

STUDYING RECALL: FOUR DIFFERENT METHODS

Memory consists of components like acquisition and storage, but the ultimate aspect of memory is recall. This is the ability to retrieve information from the memory stores when required. It has been studied in a number of ways and this article focuses upon four of them:

i) Experiments
ii) Quasi-experiments
iii) Case studies
iv) Qualitative methods

EXPERIMENTS

Experiments are the backbone of cognitive psychology and of the study of recall. Its popularity is based upon the ability to isolate variables and establish cause and effect (table 15).

A "true" experiment, as opposed to a quasi-experiment, has three elements:

i) The random assignment of participants to conditions;
ii) Standardised procedures;
iii) Control over variables.

ADVANTAGES	DISADVANTAGES
- Only method to establish cause and effect relationship between independent variable and dependent variable - Isolate and control over variables and participants - Replication possible - Measure recall precisely and allows for statistical analysis	- Low ecological validity ie artificial recall tasks - Narrowness of independent variable and dependent variable ie memory studied out of context - Demand characteristics - Experimenter effects

Table 15 - Advantages and disadvantages of the experiment for studying recall.

The experiment allows the researcher to measure recall accurately. This is done in a number of ways:

- Free recall - recall of items in any order unaided;

- Cued recall - recall of items in any order aided by prompts (cues);
- Recognition - recall of items from presented list that includes new items as well as items seen before;
- Serial reproduction - recall of items in order presented;
- Sternberg paradigm - recall of items after and before item presented in test.

In each case, there will be a clear system of scoring, like twenty words to remember. The quantitative data is then analysed using statistical tests.

Example: Very Long-Term Recall - Stanhope et al (1993)

This experiment was centred around the recall of information about Charles Dickens' novel, "Hard Times", by Open University students. One hundred and fifty two volunteers, were recruited through the Open University student newspaper, who had studied the novel as part of an Arts course. The sample ranged from three to thirty-nine months since taking the course.

The volunteers were sent a test about the novel which they self-administered at home. Recall was scored in two key ways:

a) Name and role recall of all characters (free recall)

There were three possibilities: only name, only role, or both recalled. Roles were better recalled than names, and memory for both declined over time (1st twenty-seven months) and stabilised.

b) Fact verification of fifty-four items (eg "Stephen's wife is a drunkard") as true or false (recognition)

Participants were asked to rate their confidence level for each answer also. Recall was best for highly important events and worst for low importance.

Qualitative data was also collected at the beginning of the test in the form of a short paragraph about most vivid memories of the novel.

Evaluation

1. Sending participants questionnaires to fill in themselves was a lessening of control because the experimenters could not observe the process and verified individual answering of the questions from memory.

2. Postal questionnaires allowed the recruitment of a variety of participants from throughout the UK (age range 24-74 years) who had studied the same material on the

Arts course.

3. Of the 152 volunteers, only 140 (92%) returned the test completed. Overall, 80% (111) of the volunteers were female. The question is always about those who did not volunteer - is their recall similar to the sample?

4. More ecologically valid use of memory than random word lists.

5. Detailed piloting of test materials. A group of former Open University Arts students were tested beforehand, and a matched group of non-Arts students tried the fact verification test. This control group was used to establish the chance (guessing) level score. The fact verification test was developed from a pool of 227 statements rated by four Open University Arts tutors as suitable.

QUASI-EXPERIMENTS

Quasi-experiments are almost experiments, but do not have the randomisation of participants, for example, as with gender or age as the independent variable. It is as close as possible to an experiment when an experiment is not possible (table 16).

ADVANTAGES	DISADVANTAGES
- As close as possible to experiment (eg age differences in recall) - Controls independent variable rather than manipulates it - Gives clues to cause and effect if not the exact relationship - Better for real life studies or where the experiment cannot be used	- Because not an experiment, caution about claiming causality - Not necessarily replicable - Often does not have a baseline measure - No randomisation of participants

Table 16 - Advantages and disadvantages of using quasi-experiments to study recall.

Example: Autobiographical Memory - Herlihy et al (2002)

The researchers were interested in the assumption

that if recall is different in two separate interviews it is taken as a sign of fabrication. Specifically, "the assumption that discrepancies in asylum seekers' accounts of persecution mean that they are fabricating their stories" (p324).

Twenty-seven Kosovan Albanians and sixteen Bosnians seeking asylum in the UK (living in London) were invited to take part. They were interviewed on two occasions (varying between three to thirty-two weeks) by the same researcher (and interpreter) as part of a diagnostic interview and treatment for post-traumatic stress disorder. Two key events (one traumatic and one non-traumatic) were used as the focus in the second interview.

Discrepancies between the two accounts of the events were calculated, and rated by independent judges (for inter-rater reliability). The discrepancy rate was "the number of discrepant details between answers at the two interviews (including new information) divided by the total number of units of information in the first interview", and was 0.32 overall. Not surprisingly, there were significantly more discrepancies for peripheral details.

Discrepancies were linked to presence and degree of post-traumatic stress disorder rather than intention to deceive, argued the researchers.

Evaluation

1. No way of knowing if the events described ever happened as recalled. However, this was not the focus of the researchers.

2. Sample based upon those available. Overall, 23 participants were men and twenty women. But four Bosnians did not undertake the second interview.

3. Interviews by same researcher (Jane Herlihy) each time.

4. Important area of research with implications for assessment of asylum seekers, which had a good level of control of variables. The research can suggest possible reasons for discrepancies in interviews but not the cause because it was not a "true" experiment.

5. The gap between the two interviews was longer for the Bosnian sample than the Kosovan group (mean 159 days vs 29 days). The difference in time was due to practical reasons. It was not deliberate, and would have been unethical to be so. But a "true" experiment would manipulate and control all variables even this one.

The length of time between interviews did produce more discrepancies for individuals with higher levels of post-traumatic stress disorder in the whole sample.

CASE STUDIES

Case studies involve detailed collection of information about an individual or a small group of individuals. They provide richer data than experiments (table 17). Case studies collect both quantitative and qualitative data.

ADVANTAGES	DISADVANTAGES
- Builds up detailed picture of the whole individual - Outstanding or rare cases can be studied, like individuals with exceptional memory - Helps to discover how past influences present - Not artificial	- Not possible to generalise findings - Not possible to establish cause and effect - No replicability - Researcher may miss certain details and overemphasise others

Table 17 - Advantages and disadvantages of using case studies to study recall.

Example: Exceptional or Extraordinary Memory - Luria (1968)

Luria studied "S.S" (Solomon Shereshevski), a thirty year old Russian man (at beginning of study) with a remarkable memory. He was studied since the 1920s.

Among his feats of memory, he could repeat back a list of up to seventy words (compared to less than twenty as the average) as well as repeat the list backwards or name words at any point in the list. He even remembered the list sixteen years later.

While he had perfect recall for 50 digits given three minutes to study and 40 seconds to recall, both immediately and "several months" later.

He claimed to be able to remember details of every conversation and book read, even in childhood.

There was a downside to having such a memory which included living in an inner world and appearing as a "dull, awkward, somewhat absent-minded fellow".

Evaluation

1. Though "S.S" had an exceptional memory, he also improved it more by using a mnemonic technique called method of loci. This involves remembering items by placing them (in the mind) along a familiar path (eg the way home from work), and replaying the path when recall is needed.

Wilding and Valentine (1994) interviewed 10 competitors at the Second World Memory Championships in 1993. They distinguished three groups:

- Individuals with ordinary memories who has improved with the efficient use of mnemonic strategies;
- Individuals with naturally exceptional memories;
- Individuals who combined both the above.

2. "S.S" also had synaesthesia, which is where stimuli in one sense are experienced by another sense (eg colours look hot, sounds feel).

3. He was studied in different ways including lab-based tests, and his own introspections.

4. He was studied for thirty years from the time when he was a journalist through to a stage performer as a mnemonist.

5. He was poor at using memory in same way as other people. For example when shown the matrix in figure 2, "he proceeded to recall the entire series of numbers through his customary devices of visual recall, unaware that the numbers in the series progressed in a simple logical sequence" (Luria 1969 quoted in Brace and Roth 2002 p158). Most other people would recall the information based upon the pattern rather than simple memory.

1	2	3	4
2	3	4	5
3	4	5	6
4	5	6	7
5	6	7	8
6	7	8	9

Figure 2 - Extract of matrix used to test memory of "S.S".

QUALITATIVE METHODS

Traditional research on recall collects quantitative data, but it is possible to have qualitative data as well (case studies) or instead of (table 18).

ADVANTAGES	DISADVANTAGES
- Holistic focus, including social context, rather than just memory by itself - Studies memory in real-life use, including social situations - More detailed than quantitative data - Can explore the meaning of memory to the individual	- Not possible to test data statistically - Cannot generalise or replicate - Subjectivity of researcher - Tends not to measure recall precisely

Table 18 - Advantages and disadvantages of using qualitative methods and data to study recall.

Example: Joint or Collective Remembering - Edwards and Middleton (1986)

Edwards and Middleton applied the method and theory of discourse analysis to produce a discursive model of remembering (working within the framework of social constructionism).

They believed in the:

> [I]mportance of studying remembering as a social activity governed by the settings in which it occurs, serving a potentially large set of personal and interpersonal functions in which the significance of past experiences for current purposes is generally of greater importance than accuracy and completeness..(p423)

They asked eight 1st year psychology students at Loughborough university to recall the film "E.T", which they had all recently seen at the cinema: "Try and remember as much as you can of the film 'E.T' what happens in it, what the plot is, whatever is particularly memorable of it". The discussion was recorded, a verbatim transcript made, and discourse analysis applied.

From this analysis, three processes were highlighted:

i) Framing and orientation - the establishment of

criteria for the joint recall (framing), and how individuals locate themselves in the unfolding account (orientation);

ii) Correspondence functions - putting the ideas into words (semantic function), and structuring the order of events (continuity function);

iii) Validation function - how the joint account is agreed.

Evaluation

1. The emphasis throughout this work, and social constructionism generally, is away from memory as an individual process to a:

> [W]ider, distinctively social context of establishing a mutuality of understanding between people, a shared version of past experience communicated through language (Edwards and Middleton 1986 p441).

2. This is a real use of memory in a social situation. But Roediger and Wheeler (1992) argued that, in fact, it was as artificial as an experiment in its own way.

3. This type of research moves away from traditional information-processing models of memory, and the artificial recall of word lists.

4. It cannot distinguish between what are individual memories and group memories. But, apart from in specialist situations like examinations, individuals work together to recall information, argued Edwards and Middleton.

5. There is no truth of memory (ie what did really happen) with this approach because within social constructionism, memories are treated like other utterances (eg expressions of opinion) as actions and are used to achieve certain things. To remember an event is to tell a story.

For example, "family memories" are narratives which emphasise the common family identity as they talk about holiday snapshots (Edwards and Potter 1992). But the memories are not fixed as correct or not (accurate/not accurate). They are negotiated. Collective remembering is a negotiation process, where there is no absolute truth of what did or did not happen.

REFERENCES

Brace, N & Roth, I (2002) Memory: Structures, processes and skills. In Miell, D; Phoenix, A & Thomas, K (eds) Mapping Psychology 2 Milton Keynes: Open University

Edwards, D & Middleton, D (1986) Joint remembering: Constructing an account of shared experience through conversational discourse Discourse Processes 9, 4, 423-459

Edwards, D & Potter, J (1992) Discursive Psychology London: Sage

Herlihy, J; Scragg, P & Turner, S (2002) Discrepancies in autobiographical memories - implications for the assessment of asylum seekers: Repeated interviews study British Medical Journal 324, 324-327

Luria, A.R (1968) The Mind of a Mnemonist New York: Basic Books

Roediger, H & Wheeler, J (1992) Discursive remembering: A brief note Psychologist October, 452-453

Stanhope, N; Cohen, G & Conway, M (1993) Very long-term retention of a novel Applied Cognitive Psychology 7, 239-256

Wilding, J & Valentine, E (1994) Memory champions British Journal of Psychology 85, 231-244