The Development and Validation of Mood Scales Suitable for Use with Stroke Patients with Aphasia

Paul David Barrows, PGDip BSc(Hons)

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

ABSTRACT

About a third of stroke survivors have some degree of depression. Depression has a significant impact on recovery after stroke, and identification is important so it can be treated. A common symptom of stroke is aphasia, where comprehension or expression of language is significantly impaired. Communication problems following stroke have been shown to be a major predictor of depression after stroke, yet these problems often make assessment of mood using conventional, language-based measures difficult or impossible.

Though some non-verbal, self report mood measures exist, their utility is limited and evidence base lacking. The aim of this study was to design, create and validate a non-verbal mood assessment instrument suitable both as a general outcome measure and as a screening measure for depression in stroke patients with aphasia. A series of four judgement experiments were conducted based on 22 photographic sittings, and a series of scales were developed. The resulting prototype instrument *Dynamic Visual Analogue Mood Scales* (D-VAMS) is a tablet/computer-based instrument consisting of seven bipolar scales comprising images of human faces whose expressions are modulated by sliders.

The instrument was then validated in a sample of 46 stroke survivors recruited from online, from stroke clubs and via NHS rehabilitation services. Good construct validity was demonstrated by high correlations between word and face versions of the seven D-VAMS scales (r=.73 to r=.79), however discriminant validity was poor, with

substantial cross-correlations between scores for all of the face scales (r=.58 to r=.88). Internal consistency of D-VAMS was very high, with a Cronbach's α of 0.95. A Principal Components Analysis revealed one factor accounting for 80% of the variance, corresponding to pleasantness or unpleasantness of mood.

Excellent criterion validity was evidenced by strong correlations between D-VAMS and Hospital Anxiety and Depression Scale (HADS) depression subscale (HADS-D) scores (r=.73). Excellent test-retest reliability (r=.89), and high sensitivity and specificity against HADS-D cut-offs of 4–7 were also found.

The findings suggest that the D-VAMS is a valid, brief measure of pleasantness of mood in a range of 0–100 which is suitable for use as a general outcome measure for stroke survivors with aphasia, and which may serve as an indirect, simplified measure of depression. Though D-VAMS may also be useful as a screening measure for depression following stroke, further validation is needed to examine how it performs in people during the acute stage after stroke. Some supervision may be required for people unfamiliar with using a tablet or PC interface.

ACKNOWLEDGEMENTS

This project would not have been possible without the support and guidance of my supervisors Dr Shirley Thomas and Professor Nadina Lincoln, and other staff at University of Nottingham School of Medicine's Department of Rehabilitation and Ageing, including Joanna Zuranska, Vanessa Ziegler, Louise Sabir and Helen Taylor, whose invaluable administrative work allowed things to proceed smoothly despite unexpected obstacles.

My thanks go to Professor Marion Walker MBE, Ozzie Newell, MBE, and the panel members of the Nottingham Stroke Research Partnership for their support of this project, their invaluable feedback on the D-VAMS prototype, and their suggestions for improvements to the validation study in which the instrument was assessed.

Thanks also go to Kate Hooban, Rachel Illingworth and Shirley Mitchell of Nottingham CityCare Partnership Community Stroke Discharge and Rehabilitation Service for their help in recruiting stroke survivors, and to the many service users who agreed to take part. Thanks are also owed to Bob Logan and Michael Casajuana for their help and cooperation in recruitment at Nottingham Stroke Club and Sherwood Stroke Club, and to the many members who volunteered their time to participate.

I also owe a great debt of gratitude to Dr David Stanley of University of Guelph for his invaluable guidance in the use and interpretation of Principal Components Analyses on the data from the facial expression judgement experiments described herein. I thank Professor James Russell, Professor Edward Diener, and Professor Robert Thayer for their prompt permission to adapt illustrations from their papers, and Timothy Wright for his technical advice on photographic equipment.

Finally, I thank my family and friends for always being there.

ABBREVIATIONS

ADL	Activities of Daily Living
ADRS	Aphasic Depression Rating Scale
BASDEC	Brief Assessment Depression Cards Schedule
BDI	Beck Depression Inventory
CES-D	Centre for Epidemiological Studies Depression Scale
CMA	Circumplex Model of Affect
CVA	Cardiovascular Accident (stroke)
CVD	Cardiovascular Disease
DSM	Diagnostic and Statistical Manual of Mental Disorders
DISCs	Disc Intensity Scale Circles
DT	Distress Thermometer
D-VAMS	Dynamic Visual Analogue Mood Scales
ECT	Electroconvulsive Therapy
FAI	Frenchay Activities Index
FIM	Functional Independence Measures
GDS	Geriatric Depression Scale
GRS	Graphic Rating Scale
HADS	Hospital Anxiety and Depression Scale
HDRS	Hamilton Depression Rating Scale
JHFI	John Hopkins Functional Inventory
LOC	Locus of Control
MADRS	Montgomery-Asberg Depression Scale
MDS	Multi-dimensional Scaling
MMSE	Mini Mental State Examination
MSQ	Mental Status Questionnaire
NGRS	Numbered Graphic Rating Scale
PANAS	Positive and Negative Affect Schedule
PCA	Principal Components Analysis
POMS	Profile of Mood States
PSD	Poststroke Depression
PSDS	Poststroke Depression Rating Scale
RPM	Raven Progressive Matrices
SAD-Q	Stroke Aphasic Depression Questionnaire
SODS	Signs of Depression Scale
VADS	Visual Analogue Dysphoria Scale
VAMS	Visual Analogue Mood Scales
VAS	Visual Analogue Scale
VASES	Visual Analogue Self-Esteem Scales
WDI	Wakefield Depression Inventory
ZDS	Zung Self-rating Depression Scale

Table of Contents

1. Introduction: Stroke

1.1 What is Stroke?

The World Health Organisation defines stroke as "rapidly developing clinical signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer or leading to death, with no apparent cause other than of vascular origin" (Hatano, 1976, p. 541). Stroke, or *cerobrovascular accident* (CVA) occurs when the blood supply to parts of the brain is disrupted, resulting in a sudden loss of brain function due to the consequent oxygen deprivation. The symptoms of stroke are varied, but may include paralysis or weakness down one side of the body (*hemiplegia*, or *hemiparesis*), blindness in one side of the visual field (*hemianopsia*) or difficulty speaking or comprehending language (*aphasia*). Loss of balance and coordination or loss of consciousness may also occur. If the interruption of blood supply to the affected areas is only temporary – in what is known as a *transient ischaemic attack* (TIA), then little or no damage may occur and full function can be recovered. However if the affected areas are deprived of oxygen for a long enough period, then necrosis sets in, resulting in lesions composed of dead brain matter that can no longer function. Depending on the extent of the injury, function can be recovered as the brain adapts, but many stroke survivors are left severely disabled.

Stroke comes in two primary forms, *ischaemic* and *haemorrhagic*. Ischaemic strokes result from an interruption to blood flow caused by a blood vessel becoming blocked or partially occluded by a clot or *embolism*. This starves parts of the brain of

oxygen until the blockage resolves spontaneously or is treated by medical intervention. Haemorrhagic stroke is caused by bleeds around or within the brain matter itself. In the most common type of haemorrhagic stroke, a thin-walled blood vessel within the brain may burst, leaking blood into the surrounding tissue as a *cerebral haemorrhage*. In a less common form of haemorrhagic stroke, a blood vessel on the surface of the brain itself may rupture, creating a *subarachnoid haemorrhage*. In both of these cases the result is a leakage of blood which not only prevents oxygen reaching parts of brain served by the ruptured blood vessel, but results in the critical medical complication of a pool of blood (the haemorrhage) building up within or upon the surface of the brain.

Ischaemic strokes account for approximately 80% of incidents, while haemorrhagic strokes are rarer, with intracerebral haemorrhages occurring in around 15% of stroke incidents and the remaining 5% occurring in the context of a subarachnoid haemorrhage. Stroke of any kind is an emergency and must be treated immediately. The first few hours after stroke are critical, and timely diagnosis and early treatment is essential to prevent or ameliorate damage which may otherwise result in severe disability or death.

1.2 Morbidity, Mortality and Risks

After heart disease, stroke is the single most common cause of death worldwide, and is responsible for 7% of deaths annually in the UK. (The Stroke Association, 2015) . It is estimated that stroke occurs approximately 152,000 time a year in the UK (Townsend et al., 2012), or once every three-and-a-half minutes.

Of those people who have a stroke, about third are likely to die within the first ten days, about third can be expected to make a recovery within one month and about a third will likely be left severely disabled (Bosanquet & Franks, 1998). In the decade following the year 2000, stroke accounted for an average of around 7.5% of deaths in men and 10% of deaths in women (Office for National Statistics, 2010). It is estimated that there are over 1.1 million people living with stroke in the UK (Scarborough et al., 2009), but by 2020 it is expected to double, mainly due to the increasing proportion of older people.

Factors contributing to the risk of stroke are similar to those responsible for an elevated likelihood of cardiovascular disease (CVD). Modifiable risk factors include high blood pressure, smoking, lack of physical exercise, a diet high in saturated fat and high cholesterol levels. Non-modifiable risk factors include age and a family history of coronary heart disease or stroke. Age is the single most important risk factor, with the risk of stroke doubling every decade after the age of 55; by the age of 75, 1 in 6 men and 1 in 5 women will have had a stroke (The Stroke Association, 2015).

1.3 Consequences of Stroke

Stroke causes a larger range of disabilities and has a greater impact on disability than other chronic medical condition (Adamson et al., 2004), and can include a wide range of physical and psychological impairments. In addition to permanent hemiparesis, hemiplegia, or hemianopsia, stroke patients may also have unilateral *neglect*, in which there is a denial of or reduced insight into the physical problems caused by stroke.

A range of cognitive problems may also be experienced. About a third of stroke patients have some degree of *aphasia* (Hackett & Pickles, 2014), in which a person's ability to communicate using spoken or written language may be impaired or absent. Aphasia typically accompanies damage to the left side of the brain and can include difficulty with or absence of speech (*expressive aphasia*) or impairment of the ability to comprehend spoken language (*receptive aphasia*). The ability to interpret and express written language may also be affected (*alexia* and *agraphia*). The patient may also experience memory problems, difficulties recognising objects or places; it may also impact upon executive function, impairing a person's ability to reason and organise their behaviour. Changes in affect are also common, with patients experiencing depression, elevated levels of anxiety or other mood problems.

In England at least 450,000 people are left severely disabled as a result of stroke (The Stroke Association, 2006). The health consequences of stroke are a substantial burden on health services, and are responsible for a substantial proportion of acute hospital stays, accounting for over 2.6 million bed days per year, with an average length of stay of 28 days per stroke patient (Department of Health, 2005).

2. Poststroke Depression

When a person suffers a neurological injury such as stroke, mood disturbances often result which can impact significantly upon a patient's prognosis. Though such mood problems can take many forms, and include bipolar or anxiety-related disorders, the mental health issue that has been of most clinical interest – and consequently the subject of much research – has been that of depression.

For many years it was thought that depression following stroke was no different to more common, reactive forms of depression. The appearance of depression after stroke was commonly interpreted as an expected psychological response to a life event that impacts on a person's ability to function, and which consequently impairs their subsequent quality of life. However, in an exploratory study by Folstein et al. (1977), depressive symptoms of stroke patients were compared with those of patients admitted with other medical conditions. Though both groups appeared to have similar levels of functional ability, the prevalence of depression was substantially higher in the group of stroke patients (45% versus 10%). The results seemed to suggest that there are complications specific to stroke which make these patients particularly susceptible to low mood.

These seemed to suggest that depression following stroke may have an aetiology and manifestation distinct from ordinary depression, and that it may – at least in part – be a direct result of neurological damage resulting from the stroke event rather than simply a consequence of concomitant psychosocial factors. The term

poststroke depression (PSD) came into usage as this notion gained traction, and researchers began to more intensively study factors relating to depression after stroke. Opinion regarding the aetiology of PSD, however, became polarised, with positions on this issue dividing into two separate camps. On the one hand there are those whose primary interest is in prospective biological mechanisms, such as the impact of cerebral lesions upon aminergic pathways and other neural circuits involved in mood regulation (Robinson & Szetela, 1981; Robinson et al., 1984a; Narushima et al., 2003; Bhogal et al., 2004), while on the other hand there are those who prefer to focus on reactive psychosocial and stress mechanisms (Wade et al., 1987; M. Kauhanen et al., 1999; Pohjasvaara et al., 2001; Hackett & Anderson, 2005; B. S. Townend et al., 2007a). This divide is reflected in the approaches that researchers have taken to disentangling the complex interplay of factors at work in the emergence and course PSD, and continues to be the source of much debate and controversy.

2.1 The Measurement of Depression

The study of PSD, however, has been hampered both by the overarching debate as to how depression should best be defined and assessed, and the question of whether depression resulting from stroke should be treated as a distinct syndrome with its own unique characteristics and assessment criteria, or simply as a 'ordinary' depression occurring in a particular – but not especially distinctive – context.

The way in which depression should be measured is also highly contingent upon the purpose for which it is being measured. While a doctor or psychiatrist may wish for a screening measure from which the risk of depression may be inferred, or a checklist of symptoms for the purposes of a medical diagnosis, researchers examining

the progress of patients under particular conditions or the impact of particular causes or interventions on outcome measures with regard to mental health, may prefer measures more suitable for tracking change over time.

In the early studies of depression after stroke, systematic criteria were not used for the diagnosis of depression. This changed with the introduction of the Diagnostic and Statistical Manual DSM-III and DSM-III-R (American Psychiatric Association, 1980, 1987), which offered a clear set of criteria for depression. Generally PSD studies have used this to classify patients as having major or minor depression, using either the DSM-III or DSM-II-R criteria for *major depression* (excluding the non-organic criterion), or *dysthymia* (excluding the 2 year and non-organic factor criteria). With the release of DSM-IV and DSM-IV-TR (American Psychiatric Association, 1994, 2000), accompanying medical conditions were also included in the typology of depression, making stroke one of the few medical conditions documented as being a direct cause of depression. The DSM-IV categorises poststroke depression as a "mood disorder due to a general medical condition" (American Psychiatric Association, 1994, p. 366) with specifiers of depressive features, major depressive-like episodes, manic features or mixed features. *Minor depression* is defined by DSM-IV/DSM-IV-TR as "either a sad or 'depressed' mood or loss of interest or pleasure in nearly all activities" involving "at least two but less than five additional symptoms" of major depression. (American Psychiatric Association, 1994, p. 719).

With the advent of DSM-V (American Psychiatric Association, 2013), the distinction between major and minor forms of depression gave way to a typology that reflected the duration or recurrence of depression rather its intensity. DSM-IV's *major depressive disorder* was retained as depressed mood or a loss of interest or

pleasure in daily activities consistently for at least a 2 week period, while *dysthymia* was consolidated with *chronic major depressive disorder* under the new name *persistent depressive disorder*. The essential feature of *persistent depressive disorder (or dysthymia)* is a depressed mood that occurs for most of the day, for more days than not for at least 2 years. At this point, PSD became part of the general category *depressive disorder due to another medical condition.*

Though questions have been raised by some about the suitability of some DSM criteria for assessing depression related to stroke (Gainotti et al., 1997b), the diagnosis of depression by a professional using an interview in conjunction with the DSM is now widely regarded as the gold standard. However, for much research the use of purely dichotomous classifications for depressive symptomatology is limiting, and over the years a number of instruments have emerged to offer a scalar measure of the degree or seriousness of depressive symptoms, yielding data that can be used as an approximate interval measure of depression and enable changes in depressive symptoms to be monitored.

Some of the scales use a similar, interview format but where symptoms are more specifically evaluated and graded and then processed into a final, numeric score. Where the patient is unable to clearly communicate, a proxy – such as a family member, friend, or nurse – is often used to relay their own observations in lieu of the patients own testimony, or, less frequently, the patient is assessed purely based on their observed behaviour. Scales based on this approach are broadly referred to as *observer rated*. Other scales, however, have been designed specifically to eliminate the need for a third-party assessment, and comprise a form with a number of simple

questions, to which the patient is asked to respond in multiple-choice format. These are referred to as *self-report* or *self-rating* scales.

The Hamilton Depression Rating Scale (HDRS) (Hamilton, 1960) is the most widely used instrument for quantifying depression in research. The scale is intended to be administered by a trained clinician and consists of 21 items addressing specific symptoms, which are graded by responses on a 5-point scale. The total score, in the range of 0-52, gives an indication of the severity of depression in the patient. Scores of 7-17 are considered to signify "mild" depression, 18-24, "moderate" depression, and any score above 25 to mean "severe" depression. The scale takes around 30-45 minutes to administer.

Like the HDRS, the Montgomery-Asperg Depression Rating Scale (MADRS) (Montgomery & Asberg, 1979) is an observer rated depression scale intended to be administered by a trained clinician. The MADRS however, was developed in response to the need for instruments that are more sensitive to change. The MADRS was developed to specifically address the needs of research in which the ability to measure changes in depression across time – such as the response of patients to antidepressants – could be more accurately gauged. The scale consists of 10 items describing symptoms which are graded on a 6-point scale. The total score is in the range of 0-60, with any score of 7 or above considered to be associated with the presence of some degree of depression. Administration time is similar to the HDRS, with the assessment taking around 20-60 minutes to complete.

In much research, it is not practical to involve experienced clinicians in diagnosing or quantifying depression using these types of instruments, therefore self-report measures have been developed by which patients can be quickly assessed by

means of a short questionnaire. These types of measures have the advantage of being much quicker and easier to administer. The most widely used of these self-report measures is the Beck Depression Inventory II (BDI-II) (Beck et al., 1961; Beck et al., 1996), a 21-item questionnaire with responses scored on a 4-point scale. The total score is in the range of 0-63. Scores of 10-18 are considered to signify "mild" depression, 19-29, "moderate" depression, and any score above 30 to mean "severe" depression. The scale takes only 5-10 minutes to complete, so is particularly desirable in studies where a brief assessment is required.

Also widely used is the Zung Self-Rating Depression Scale (ZDS) (Zung, 1965). The scale is similar to the BDI-II, but is considered more useful in patients with neurological disorders (Carota & Bogousslavsky, 2003). It consists of 20 items scored on a 4-point scale, with a total score in the range of 20-80. Scores of 50-59 are considered to signify "minimal to mild" depression, 60-69, "moderate to severe" depression, and any score above 70 to mean "severe" depression. Like the BDI-II, it is reasonably quick to administer, taking approximately 10-15 minutes to complete.

The Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983) was developed as a response to the widespread prevalence of mood disorders comorbid to physical illness and disability, and to complaints of somatic symptoms that have no basis in any organic pathology. Like the BDI and the ZDS, the HADS is a self-assessment instrument, however it was designed for use in a non-psychiatric hospital setting to allow medical staff to quickly assess psychological distress. In order to avoid confusion of physical symptoms with those of a psychological origin, it specifically avoids somatic symptomatology common to both psychological and physical conditions, and is restricted to purely 'psychic symptoms'. The HADS

consists of 14 items divided into two sub-scales of 7 items each, one for anxiety related symptoms (HADS-A), and the other for symptoms relating to the presence of depression (HADS-D). Both scales have a score in the range of 0-21, with a score of 8-10 representing borderline cases and a score of 11 or more indicating the presence of clinically significant symptoms. Excellent validity of the HADS has been reported compared to the HDRS (Carota & Bogousslavsky, 2003), and the HADS is widely used in studies which include hospitalised patients. It is particularly quick to administer, taking only around 5 minutes to complete.

Because of the confounding somatic factors introduced by the physical symptoms of stroke, a number of scales have also been developed to control for the symptoms common to both stroke and depression, such as vegetative and neuropathological symptoms. These adapted scales will be discussed in Chapter 3.

2.2 Prevalence of PSD

In a recent systematic review and meta-analysis Hackett and Pickles (2014) examined data from 61 population-based studies on the proportional frequency of depression up to five years following stroke. Only studies using prospective, consecutive recruitment of patients from within a clearly defined geographic area and region of time were included, and studies with selection criteria or patients characteristics which were not deemed generalisable were excluded. The measures used to assess frequency of depression included the Beck Depression Inventory (BDI) (Beck et al., 1961), the Geriatric Depression Scale (GDS) (Yesavage et al., 1983), the Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983) and the Hamilton

Depression Rating Scale (HDRS) (Hamilton, 1960), with DSM criteria providing the most common criterion for the presence of depression.

Pooled data indicated that depression is present in 31% (95% CI 28–35%) of stroke patients up to five years after stroke. This figure did not differ significantly from the one arrived at in an earlier review (Hackett et al., 2005) in which a pooled figure of 33% (95% CI 29–36%) was calculated. Subgroup analyses were also performed based on recruitment method (population-based, hospital-based, and rehabilitation-based) and timing of mood assessment (0-1 month, 2-5 months, 6-9 months, 1 year, 2-4 years, 5 years), but the results showed little variation across method or timescale, although "a lower proportion had depression between one and less than 5 years" (Hackett & Pickles, 2014, p. 6).

In a recent cohort study of patients followed up for 30 months after a minor stroke (Altieri et al., 2012), meanwhile, 41% were diagnosed with PSD at some point during follow-up, with 2.9% meeting the DSM-IV criteria for major depression; 22% were depressed within the first month post-stroke.

Though a figure in the region of 30% also appears to emerge with notable frequency in studies elsewhere (Singh et al., 2000; Cully et al., 2005; Paolucci et al., 2006; Pinoit et al., 2006; Linden et al., 2007; Bergersen et al., 2010; Ostir et al., 2011; Hommel et al., 2015), substantial variations have also been found. A review by Robinson & Spaletta (2010), for example, found such large variations in prevalence between studies, that they could venture an estimate of overall prevalence of depression as somewhere within the range of 20% to 60%, while an earlier literature review (Gordon & Hibbard, 1997) could only offer a figure of between 25% and 79%.

Prevalence figures vary for a number of methodological reasons. Firstly, methods for assessing depression differ from study to study. In addition to "gold standard" criteria such as those of the DSM, which themselves are subject to revision with each new release, many different self-rating scales and observer rating scales are in use to offer a quantified measure of depression. Defining and measuring depression, therefore, does not enjoy a clear, unambiguous consensus. Different assessment methods evaluate and prioritise symptomatology in different ways, yielding different classifications and different diagnostic conclusions (Pohjasvaara et al., 1998). Furthermore, where a categorical diagnosis has been used, though some studies have adhered to traditional diagnostic subdivisions of 'major' and 'minor' depression, others have used a simple, binary classification of 'depressed' and 'non-depressed' based on cut-off scores deemed appropriate for a given scale.

Secondly, there is the matter of the timing of the assessment; the magnitude of PSD has been found to differ substantially depending on the time elapsed since the index event (Whyte & Mulsant, 2002). There have been numerous studies suggesting that the onset and progression of PSD follow a changing pattern of incidence and prevalence across time following the index stroke event, and vary according to a multitude of factors.

Thirdly, there is the matter of the population under examination in a particular study. Different studies use patients in different settings; some may be hospitalised, while others may be outpatients or participants in a rehabilitation program. They may have varying levels of functional impairment and may be receiving different types and levels of support. Furthermore the selection criteria for studies may omit certain patients in an attempt to eliminate biases or confounding factors, or simply for

practical reasons. Patients with a history of depression prior to stroke are often omitted, for example, while patients with significant communication problems due to aphasia are usually excluded due to the practical difficulties in assessing them. People with haemorrhagic – rather than ischaemic – stroke are also frequently excluded.

Finally, there are complexities in recognising and diagnosing depression symptoms in stroke patients due to the presence of symptoms that are directly due to the stroke itself (Stern, 1999; Hackett et al., 2005; E. Townend et al., 2007b). Certain symptoms, such as hemiplegia and fatigue may mask or mimic the effects of depression, making accurate diagnosis difficult. These issues will be covered in more depth in due course.

2.3 Onset and Natural History of PSD

In trying to understand the aetiology of PSD, many studies have examined cohorts of patients across a period of months or years in order to study the pattern of onset and the changes of prevalence and incidence of PSD across time. Such studies have usually examined PSD in relation to cognitive impairment, physical disability, lesion location and various demographic factors in order to better understand what factors may influence outcome after stroke.

As reported earlier, the recent systematic review and meta-analysis by Hackett and Pickles (2014), in which frequency of depression was examined up to five years following stroke, found that a lower proportion had depression between one and less than 5 years. But with just two categories of time periods (1 year, 2-4 years) spanning the time between 9 months and 5 years, it is impossible to accurately gauge the trajectory of prevalence changes within this period. An earlier review with less

stringent inclusion criteria (Hackett et al., 2005) found no evidence of a pattern of prevalence changes across time. However they conceded that the complex, heterogenous nature of the many studies made comparison difficult.

The presence of such a pattern is particularly important to understanding its aetiology, as a biphasic pattern is consistent with the existence of a distinct, biological component to PSD, in which the acute effects of lesions on brain function play a role. Proponents of this view have hypothesised that an acute syndrome triggered by damage to brain circuits involved in mood regulation contribute to an initial increase in depressive symptoms, but that as the brain heals, these symptoms subside, but there then follows a second phase of depression in which psychosocial causes predominate (Robinson et al., 1984b; Robinson, 2003; Robinson & Spalletta, 2010). While the pattern of prevalence in Hackett and Pickles (2014) hints at a drop in prevalence consistent with a biphasic pattern, studies elsewhere offer a clearer picture of PSD prevalence changes following stroke.

Despite the lack of unequivocal consensus, some studies have found tentative evidence supporting a biphasic pattern to the onset of PSD, with the period of highest risk appearing in the first few weeks following stroke, and prevalence falling during the first year and then rising again around 12-18 months later.

Some of these studies suggested a steady decline in depression scores during the first year following stroke which is consistent with this biphasic pattern (Paolucci et al., 2006; B. S. Townend et al., 2007a; Ostir et al., 2011), however the time-span covered did not allow the question of a second phase to be examined. Others, though, have yielded evidence supporting this second phase. Berg et al. (2003) reported that the prevalence of depression was highest at the acute stage, but then tailed off

between 3 and 6 months, only to rise again at 12 months. While in perhaps the strongest evidence to date, a meta-analysis of 14 studies by Whyte and Mulsant (2002) found that peak prevalence of major depression appeared to be 3-6 months after stroke, after which it declined to around 50% of initial rates at about 12 months, only to rise back close to its peak levels between 18 and 24 months poststroke, where it remained high up to 3 years after stroke.

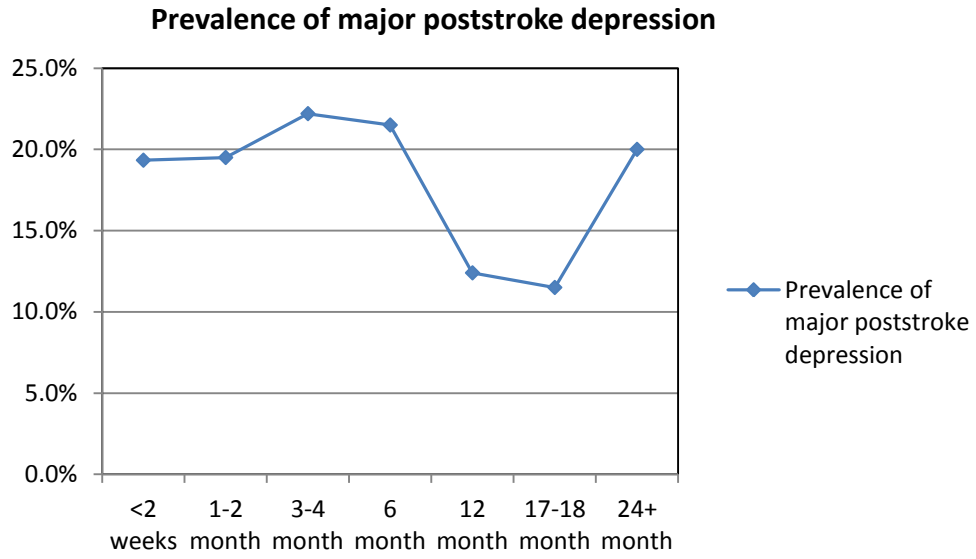


Fig 1. *Prevalence of PSD over time, from data in Whyte & Mulsant (2002)*

A similar pattern was also in evidence in a 3-year cohort study by Astrom et al. (1993) which found that the incidence of major depression was 25% at the acute stage and 31% at 3 months, dropping to 16% by 12 months post-stroke. At 2 years it was 19% but by 3 years it had risen to 29%. Prevalence of depression was highest immediately after stroke, particularly in community settings, and many stroke patients developed depression hours to days after stroke, many also became depressed much

later, supporting the findings of an earlier study by Robinson (1984b) in which 30% of patients not initially manifesting PSD symptoms became depressed 6 months later. They also noted that while almost half of stroke patients developing major depression within 3 months of stroke remain depressed at one year (Astrom et al., 1993; Wade et al., 1987), onset of PSD within days of stroke was more likely to remiss spontaneously, whereas patients with onset at more than 7 weeks had a lower rate of spontaneous recovery (Andersen et al., 1994). The authors concluded that there may be some utility in sub-typing major poststroke depression into early versus late onset, though it is difficult to reach any firm conclusion on what periods defines each.

2.4 Consequences of PSD

Even when considered in isolation, depression has been associated with a number of negative outcomes, such as increased disability (Lenze et al., 2001), impaired cognitive function (2000) and increased mortality (Schulz et al., 2000; Reynolds et al., 2008; Ellis et al., 2010). When depression occurs as a result of stroke, a similar range of problems can therefore be expected, but ones that are exacerbated by the added complexities of stroke symptomatology. PSD has been shown to impact significantly on rehabilitative outcomes, and many studies have now shown that depression following stroke is associated with similar negative outcomes regarding functional impairment and mortality.

2.4.1 Physical Disability. Much of the research into the impact of PSD has been focussed on physical function. One of the most significant early findings to emerge from research into PSD was the fact that depression following stroke does not

appear to be solely a consequence of functional impairments resulting from the stroke itself (Folstein et al., 1977).

A recent review (Kutlubaev & Hackett, 2014) examined 14 studies totalling 4498 participants from the previous ten years, in which the association between PSD and stroke outcome were examined, alongside the results of an earlier review (Hackett & Anderson, 2005). Though follow-up from assessment to outcome varied considerably across the studies, depression was consistently shown to impact negatively on functional outcome following stroke. Depression was also found to predict poorer quality of life, poor life satisfaction, less efficient use of rehabilitation services and need for institutional care.

Elsewhere, Robinson et al. (1983) found that measures of physical function by the John Hopkins Functional Inventory (JHFI) had significant negative correlations with three independent depression measures, with significant increases in correlations between JHFI scores and measures of depression throughout a six month follow-up, (Robinson et al., 1984b).

Wade et al. (1987) likewise found significant negative correlations between scores of physical function by the Barthel Index (BI) and Frenchay Activities Index (FAI) and Wakefield Depression Inventory (WDI) scores during the year following stroke, while Parikh et al. (1990) found that depressed patients as determined by DSM-III criteria had lower JHFI scores and showed significantly less functional recovery over two years than non-depressed patients.

Bacher et al. (1990) found that physical function as assessed by BI improved throughout follow-up but was significantly higher for those who were not depressed initially, while Morris et al. (1992) found that depressed patients as assessed by DSM-

III criteria improved less than non-depressed patients in functional status. The 3-year, longitudinal study by Astrom et al. (1993), discussed earlier, yielded similar findings; at one year follow-up, patients who were depressed at the acute stage did not recover in functional ability as assessed by ADL measures, whereas non-depressed patients recovered significantly.

Kauhanen et al. (1999) found that patients classified as depressed by DSM-III criteria performed worse on almost every measure of physical function than non-depressed patients. A study by Pohjasvaara et al. (2001) yielded similar findings, while also shedding light on the direction of causation between PSD and physical function. Depression at 3 months was found to be an independent predictor of poor functional outcome at 15 months (OR 2.5), yet functional outcome at 3 months did not predict depression at 15 months, thus favouring the interpretation of PSD as a cause of rather than a consequence of functional impairment.

If PSD is indeed a significant factor in functional outcome, then this may be a key intervention point in ameliorating the adverse consequences of stroke, as it suggests that treating depressive symptoms may improve a patient's prognosis. To examine this hypothesis, Chemerinski (2001) examined differences in recovery in PSD patients with and without remission of depression over the first 3 to 6 months poststroke. Group comparison showed that patients with remission of depressive symptoms showed significantly greater functional recovery than patients without remission at follow-up.

This growing evidence of the impact of PSD upon functional outcome was underlined in a meta-analysis of 13 earlier studies by Carota & Bogousslavsky (2003). The results revealed that out of 11 factors examined, only functional outcome was

reliably found to be associated with PSD, with 10 out of 11 studies examining this finding a significant relationship.

The large, multicenter study by Paolucci et al., (2006) mentioned earlier – as well as indicating patterns of prevalence over time – also examined physical function outcomes throughout the 9 month follow-up. The results, similarly, revealed that depressed patients had significantly more physical impairment at baseline and follow-up than those without PSD, while patients without PSD and patients with early-onset showed significant improvement from baseline BI scores at follow-up. In another study by Van de Port et al. (2006), activity level, fatigue and depression at 1 year after stroke were found to be statistically significant predictors of mobility decline between 1 and 3 years poststroke.

One of the problems with hospital or rehabilitation-based studies is that it is difficult to know whether factors associated with the selection and treatment of the patients may impact upon the results, and whether the findings of such studies are generalisable to the population at large. Goodwin and Devanand (2008) sought to address this by examining the results of a survey conducted on a large, nationally representative sample of U.S. residents between the age of 25 and 74 years. A multivariate regression analysis showed that both stroke and depression independently predicted poorer function on all measures of ADL, and revealed a significant association between major depression and stroke in the previous 12 months (OR 3.5, CI=1.4–8.9), with 29.2% of stroke patients having experienced depression in the previous 12 months, a figure consistent with many other prevalence studies. These findings offer some reassurance that studies of treatment-seeking samples reflect findings within the general population.

West et al. (2010) examined the association between psychological symptoms after stroke and functional outcomes in a cohort of recovering stroke patients across one year. After adjusting for age, sex and initial disability levels, regression analysis revealed that persistent psychological symptoms in the first 26 weeks were associated with much poorer physical function at 52 weeks.

Even when the assessment of depressive symptoms is greatly simplified, a relationship between depressed mood and subsequent level of disability has still proved evident. As part of the *North Manhattan Stroke Study*, a population-based, follow-up study spanning 7 years, depressed mood was assessed 7-10 days poststroke using just the first question of the HDRS (Willey et al., 2010). When controlling for other factors in a multivariate logistic regression analysis, depressed mood was found to be associated with greater likelihood of severe disability as compared with no disability at all at 1 year (OR 2.91; CI=1.07–7.91) and 2 years (OR 3.72; CI=1.29–10.71) poststroke.

Though studies overwhelmingly support the existence of an association between PSD and outcomes on measures of physical function, not all have found a relationship. Robinson et al.'s (1984a) study of lesion location and depression following stroke found no relationship between JHFI scores and overall depression scores. Sinyor et al. (1986), in a study of rehabilitative outcome on 65 patients, found that self-report depression scores did not correlate with physiotherapy or occupational therapy function scores upon admission; furthermore, when patients were divided into depressed and non-depressed groups, both showed significant improvement of function from admission to discharge but no significant differences were found between the degree of improvement in the two groups.

Paolucci et al. (1999), in an earlier study of the impact of PSD on rehabilitation outcomes in 470 stroke patients, concluded that PSD had a low impact on rehabilitation and functional outcome as a whole, however, as discussed earlier, their subsequent study (Paolucci et al., 2006) yielded much stronger evidence that PSD impacts on rehabilitative outcomes.

2.4.2 Cognitive Impairment. Many of these studies of the impact of depression on physical function, however, also included at least one measure of cognitive function.

Robinson et al.'s (1983) two year follow-up study also found correlations between his three, independent depressions measures, and scores of cognitive function as measured by the Mini Mental State Examination (MMSE), while a subsequent study (Robinson et al., 1984b) also yielded significant increases in correlations between cognitive impairment and depression scores between 3 and 6 months. A significant correlation between MMSE and overall depression scores was also found, but only in patients with left hemisphere lesions (r=-0.43). Likewise, Wade et al.'s (1987) community study also found a significant relationship between depression scores (WDI) and cognitive function as measured by Raven Progressive Matrices (RPM), with WDI and RPM scores showing a significant negative correlations throughout the one year follow-up. Bacher et al.'s (1990) study also examined cognitive function as measured by the MMSE; as with physical function, cognitive function did not change significantly throughout the one year follow-up, but MMSE scores showed a significant negative correlation with initial depression scores, suggesting that depression predicted cognitive impairment.

Morris et al.'s (1992) study of a cohort of 49 stroke patients at two rehabilitation hospitals, also examined cognitive function as an outcome measure using the Mental Status Questionnaire (MSQ) across a 14 month follow-up. The results revealed impairments of cognitive recovery in patients classified as depressed, with the depressed group showing less cognitive recovery than the non-depressed group. Despite near identical initial MSQ scores, depressed patients showed a small decline (-1%) from baseline, whereas non-depressed patients showed a modest improvement (11%). In Kauhanen et al.'s (1999) one year follow-up study of 106 stroke patients admitted to a hospital stroke unit, patients classified as depressed by DSM-III criteria scored poorer on nonverbal problem solving, verbal and visual memory, attention and psychomotor speed. Most of these differences also markedly increased in significance between 3-months and one year poststroke.

As in studies such as that of Pohjasvaara et al. (2001), some researchers have also critically examined the direction of causality at play in assessing relationships between PSD and cognitive impairment, as it is not clear from many study findings whether depression is a symptom of or cause of cognitive impairment.

Murata et al. (2000) set out to shed light on the causal relationship between cognitive impairment and PSD by examining the relationships between improvements in cognitive function and recovery from major depression after stroke. Though the generally prevailing view of cognitive impairment is as one of PSD's many comorbid symptoms – sometimes known as "dementia of depression" or *pseudodementia*, an obvious alternative exists; as with physical disability, some have surmised that cognitive impairment may be *a cause of*, rather than *a consequence of* PSD.

If cognitive impairment *causes* PSD then the treatment of depression after stroke would not be expected to be accompanied by a recovery of cognitive function. However, if cognitive impairment is a *consequence* of PSD, then patients with major depression who show improvement in mood after stroke should also show greater improvement in cognitive function than those who showed no improvement in mood; this latter hypothesis is what the authors proposed.

By dividing patients into groups with and without major depression, and into sub-groups of those showing significant improvements in depression scores and measures of cognitive function respectively, they were able to examine this hypothesis. The results confirmed the predicted relationship: patients with major depression whose mood improved also showed significantly greater cognitive improvement than any other group. Furthermore, depression scores in the group of patients with major depression improved despite worsening scores of cognitive functioning. The authors concluded that these findings favour the view that cognitive impairment is more frequently a consequence rather than a cause of PSD.

The evidence, therefore, overwhelmingly suggests a strong relationship between cognitive impairment and PSD, with few studies failing to find an association. Only two negative findings are immediately evident: Lipsey et al. (1983) found no significant correlation between MMSE scores and any depression score. While findings from Robinson's two year follow-up study (Robinson et al., 1984b) were mixed; though there was a correlation between severity of depression and MMSE scores 2 weeks after stroke. this correlation dropped dramatically at 3 months but then increased significantly between 3 and 6 months post-stroke.

2.4.3 Mortality. Clearly**,** the most serious potential consequence of stroke or its concomitant impairments is death. Any condition that significantly impacts upon the mortality rate of its patients is a cause for great concern, and underlines the importance of screening, early detection and timely intervention. Many researchers have therefore focussed their efforts on the question of whether PSD – in addition to its numerous adverse functional consequences – may also be responsible for higher mortality amongst stroke patients.

In a systematic review and meta-analysis, Bartoli et al. (2013) examined a total of 13 studies from 1993 to 2011 in which mortality risk was assessed in stroke survivors with and without depression following stroke. The studies included a total of 59,598 participants, around 10% of whom had depression after stroke. The authors calculated a pooled OR of 1.22 (95% CI=1.02–1.47) and an HR of 1.52 (95% CI =1.02–2.26) for mortality at follow-up. They also noted that the association between PSD and mortality was contingent on the duration of the time studied. While short term studies (<2 years) showed no statistically significant relationship, medium term studies (2–5 years) did, while long term studies (>5 years) showed a weaker relationship.

In a five year, prospective cohort study of 7,381 older adults living in the community, Reynolds et al. (2008) studied the impact of depression on active life expectancy and disability, both independently of and comorbid to the presence of chronic diseases, including cancer, diabetes, heart disease and stroke. The findings revealed that the presence of stroke and depressive symptoms reduced mean active life expectancy by 10.8 years in men (from 11.5 to 0.7 years) and 8.0 years in women (from 12.3 to 4.3 years). However the study also found similar reductions of total and

active life expectancy in patients with depressive symptoms who had other chronic diseases, so – in contrast to Folstein's (1977) study – did not suggest that stroke patients were particularly vulnerable to the effects of comorbid depressive symptoms in respect of life expectancy. The findings of a cohort study of 129 stroke patients (B. S. Townend et al., 2007a) also indicated an increased risk in those with depression following stroke, with HADS scores at baseline significantly predicting mortality at 3 months.

Elsewhere, Lewis et al. (2001) examined the impact of depression and anxiety symptoms as well as some of the attitudes and psychological dispositions that accompany them, including helplessness/hopelessness, anxious preoccupation, fatalism, and denial/avoidance. Though fatalism (OR 1.07, CI=1.01–1.14) and helplessness-hopelessness (OR 1.07, CI=1.01–1.13) were associated with decreased survival 3 to 5 years after stroke, HADS depression/anxiety scores were not, indicating that while negative attitudes were associated with increased mortality, the presence of depression or anxiety was not.

On balance, then, studies largely bear out the conclusions of Bartoli's review, which is that depression after stroke significantly – but not drastically – elevates the risk of death 2–5 years after stroke, probably by somewhere in the region of 22%.

2.5 Predictors of PSD

To fully appreciate the aetiology of PSD, an understanding of its consequences must go hand-in-hand with an understanding of its causes and risk factors. Understanding the causal factors for a condition is of particular clinical value because it can help identify potential points of intervention prior to the condition's emergence and

broaden the base of preventative strategies available. PSD and accompanying functional consequences are known to result from stroke, but key to preventing and treating depression and other mood disorders following stroke is understanding the factors that increase or decrease their likelihood.

The aetiology of PSD is of particular interest to theorists because it is the key to constructing and testing theories about its possible underlying mechanisms. While some favour psychosocial accounts, framing depressive symptoms as a response to disability and social isolation, others have sought biological explanations for mood disturbance emerging in the context of this type of acute brain injury, with many studies examining neuroanatomical correlates of depressive symptoms to see if PSD might be accounted for by damage or disruption to specific neurological mechanisms.

2.5.1 Functional Impairment and Psychosocial Factors. Just as PSD has been shown to be a predictor in the patient's subsequent level of functional impairment, there has also been much research examining the reverse position, that is, the impact of stroke's concomitant physical disability and impairment on the subsequent manifestation of PSD. Astrom et al.'s (1993) 3-year longitudinal study of patients admitted to a stroke unit yielded some predictive associations regarding function and subsequent depressive symptoms. The most important predictor of depression at 3 months was dependence in ADL, while beyond 12 months the lack of social contact appeared to the most significant determinant. Of particular interest is that one of the two most important predictors of immediate major depression was aphasia, implicating communication difficulties as a key factor in PSD outcomes.

Pohjasvaara et al. (1998) studied a consecutive series of 486 patients with ischaemic stroke at 3 months poststroke. A history of previous depression was found

to be significantly associated with a diagnosis of depression (OR 2.3; CI=1.3–4.4) and with a diagnosis of major depression (OR 3.4; CI=1.7–6.7). The study also found that dependence in ADL correlated with a diagnosis of depression (OR 1.8; CI=1.1–3.1) and major depression (OR 2.9; CI=1.6–5.5), however it was not possible to infer a causal relationship between ADL and depression because measures were taken at the same time point.

In a study of functional and neuroanatomical correlates of PSD, Singh et al. (2000) found a significant negative correlation between functional independence measures (FIM) and depression scores at both 3 (r=0.38) and 12 months poststroke (r=0.36). FIM scores at 1 month were also found to be the most significant predictor of depression scores at 3 months, with the authors concluding that the degree of functional independence is the greatest risk factor for PSD. In an 18-month follow-up study of 100 stroke patients, Berg et al. (2003) found that functional impairment and severity of stroke were the most significant predictors of depression from 6-12 months after stroke, and in a review of 13 studies of examining a total of 11 factors, Dieguez (2004) concluded that functional outcome and stroke severity were the only reliable predictors of PSD.

As part of a longitudinal prospective study, Whyte and Mulsant (2004) examined the relationship between stroke and ADL measures and outcome measures of depression. Though stroke survivors were found to have a greatly elevated risk of depressive symptoms (OR 6.3; CI=1.7–23.2), the outcome was independent of other predictors, including physical disability.

Hackett et al. (2005), in a review of 20 studies, concluded that severity of stroke, physical disability and cognitive impairment were the factors most

consistently associated with depression, with at least 16 of the studies supporting this. However no clear conclusion could be drawn about the characteristics of patients at most risk of depression after stroke.

In a study of patients recruited to a randomised controlled trial of Cognitive Behavioural Therapy (Thomas & Lincoln, 2006), 123 depressed patients with mild to moderate disabilities were assessed for ADL, depression, communication difficulty and locus of control (LOC) after stroke and then again 6 months later. ADL was assessed using the extended activities of daily living scale (EADL), communication problems were assessed with the Sheffield Screening Test for Acquired Language Disorders (SST) and LOC was measured using the Recovery locus of control scale (RLOC). Patients with severe depression at follow-up had significantly higher depression scores and significantly lower language and LOC scores at baseline, while patients severely depressed at recruitment scored significantly lower on communication and ADL at follow-up. A logistic regression showed that both SST (OR 0.69, CI=0.51–0.93) and RLOC (OR 0.99, CI=0.81–0.99) scores were significant predictors of depression at 6 months as measured by the BDI. The authors concluded that communication impairment is the strongest predictor of the presence and severity of depression after stroke, with locus of control also being a significant determinant. Physical disability in terms of ADL function at baseline, however, did not appear to predict the level of depression 6 months later.

More evidence of the impact of physical function on PSD outcomes emerged in a study of early poststroke mood disorder by E. Townend et al. (2007a), in which 127 stroke patients were examined at initial presentation, then at follow-up 1 and 3 months later. Multivariate regression revealed that, at 1 month, ADL disability,

change in stroke severity scores and social support were independently associated with depressive symptoms, while at 3 months, disability, social support and institutionalisation significantly predicted depressive symptoms.

Following their findings that communication difficulties are a likely predictor of subsequent depressive symptoms, Thomas & Lincoln (2008) ran another study examining the relationship between communication impairment, ADL and emotional distress at 1 and 6 months poststroke, in a sample of 100 patients, 21% of whom were classified as having aphasia. The findings revealed that expressive language and ADL (BI) scores were significant predictors of emotional distress at both 1 month, while expressive language and EADL predicted emotional distress at 6 months. Furthermore, baseline scores for the BI (r=0.46), expressive language (r=0.35) and receptive language (r=0.29) predicted emotional distress at 6 months.

Saxena (2008) studied the relationship between of cognitive impairment and depressive symptoms poststroke during the treatment and rehabilitation of 252 consecutive patients admitted to a large city hospital. ADL dependency (BI) (OR 5.28; CI=2.11–13.18), cognitive impairment (OR 4.78; CI=1.85–12.29) and recurrent stroke upon admission (OR 3.34; CI=1.33–8.36) were all found to be independently correlated with depression at 6 months. Thus, not only physical disability, but cognitive impairment and prior stroke appeared to predict subsequent severity of depression symptoms.

In another follow-up study of patients referred to a rehabilitation unit, Farner et al. (2010) also concluded that poor level of function and severity of stroke – along with lower pre-stroke social activity – were significantly predictive of persistent depression. In a review of the literature on PSD, meanwhile, Robinson (2010)

concluded that the strongest single correlate of depression is severity of impairment in activities of daily living.

Ayerbe et al.'s (2011) study of 3,689 first-time stroke patients yielded more evidence of the role of physical and cognitive function as predictors of outcome. Measures of physical and cognitive function were collected at baseline, and then again at 3 months, 1, 3 and 5 years poststroke. Independence in ADL predicted lower depression rates at all time points with odds ratios ranging from 0.35 (CI=0.24–0.50) to 0.50 (CI=0.36–0.88). Cognitive impairment at baseline also predicted depression at all follow-up time points with odds ratios from 1.81 (CI=1.28–2.56) to 2.30 (CI=1.36–3.87), but was only associated with severe depression only at 3 month and 1 year. Inability to work also predicted moderate and severe depression at all time points but one (2 years and 3 years respectively).

Though most studies have been concerned primarily with functional factors, some have also examined the role of prior history of depression on outcomes poststroke. A systematic review by Hackett et al. (2006) suggested that a prior history of depression was predictive of depression after stroke, with further evidence of this emerging in a study by Wulsin et al. (2012), who found that the presence of poststroke depression and a pre-existing history of depression were predictive of scores on quality of life measures 3 and 12 months later.

2.5.2 Neuroanatomy of Lesion. The aetiology of PSD is of particular interest because stroke results in diverse types of neurological damage associated with a range of functional and psychological consequences. As well as lending insight into the localisation of brain function with respect to many aspects of cognition and behaviour, the consequences of stroke also provide an opportunity to examine the

relationship between depression and brain function. However, because of the often severe psychosocial consequences of stroke, it is also an area that is of great interest to those examining the psychological and social dimensions of depression. This has lead to something of a collision between reductionist, biological interpretations of PSD, and the accounts preferred by many psychologists, in which the focus is on the concomitant psychosocial stress mechanisms (Dieguez et al., 2004).

While the studies detailed in the previous section have focussed primarily on the psychosocial mechanisms determining the onset and course of PSD, there is a growing body of evidence that the neurological damage caused by a stroke event – specifically the location and size of the resulting brain lesion – may play a significant role in PSD. Up until the 1980s, the widely prevailing view was that depression following stroke is a natural result of the disability, cognitive impairment and communication problems which frequently accompany it, and the psychosocial stresses of the resulting life changes and adjustments. The aforementioned study by Folstein et al. (1977), however, raised questions about this assumption, suggesting that there was something unique to stroke that made patients particularly susceptible to depression compared to patients with other, equally debilitating medical conditions.

A study by Robinson and Coyle (1980) suggested a neurophysiological mechanism underpinning mood disturbance following brain injury. The study examined catecholamine depletion following brain injury in rats. Robinson and his colleagues hypothesised that the lateralisation of neurotransmitters involved in emotional responses may cause mood regulation to become impaired as a result of specific types of lesion.

In subsequent studies in human stroke patients (Robinson & Szetela, 1981; Robinson et al., 1983; Robinson et al., 1984a), Robinson and others found further evidence suggesting a relationship between left hemisphere, anterior cortex or basal ganglia lesions and depression. They found a prevalence of depression in stroke patients that was three times higher than that in brain-injured patients (60% in the former versus 20% in the latter), however this difference vanished when the lesion location was controlled for, suggesting that the area of brain injury in stroke is a key determinant of PSD. They also reported that patients with lesions in the left frontal lobe had significantly greater mean depression scores than patients with lesions elsewhere, and that patients with left hemisphere lesions showed a strong relationship between severity of depression and the distance of the lesion from the frontal pole (r=-0.76). The authors also noted that while 47% of patients showed some degree of depression, 9% displayed "inappropriate cheerfulness" – an important observation that seems difficult to account for by any other explanation.

Though a number of other studies have found relationships between PSD and lesion location (Wade et al., 1987; Astrom et al., 1993; Morris et al., 1996), others have been unable to confirm these findings (Sinyor et al., 1986; Parikh et al., 1990; Stern & Bachman, 1991; Andersen, 1997; Paolucci et al., 1999). A systematic review of 13 studies examining lesion location and depression yielded mixed findings (Singh et al., 1998), while a larger review of 35 studies examining this relationship (Carson et al., 2000) found no significant differences in relative risk of depression regardless of the lesion location.

Robinson and colleagues responded with a meta-analysis of their own in which they corrected for biases they alleged in this last review, and found results that

agreed with their own earlier findings (Narushima et al., 2003). While in a further meta-analysis of 26 studies, Bhogal et al. (2004) found results which supported those of Robinson and colleagues, but suggested that the effect was highly contingent on the time since stroke, with studies within 28 days of stroke showing the clearest evidence. The authors also cautioned that the inclusion of aphasic patients in studies of this kind introduces a confounding factor that can make interpretation troublesome. Patients with aphasia almost always have left hemisphere lesions and have a higher risk of depression, which may create the illusion of a relationship between lesion lateralisation and PSD where none actually exists.

Though other studies have failed to replicate this finding (Nys et al., 2005), Robinson and his colleagues continue to demonstrate evidence for an association between lesion characteristics and depressive symptoms poststroke. In a recent review and meta-analysis of studies undertaken within 2 months of stroke, Robinson and Spalletta (2010) found a far higher prevalence of depression in those with left-anterior as compared to left-posterior lesions (OR 2.29; CI=1.5–3.4) and a similarly high prevalence difference between those with left-anterior as compared to right-anterior lesions (OR 2.18; CI=1.4–3.3). The authors maintain that time since stroke appears to be crucial in the relationship between lesion location and prevalence of depressive symptoms.

2.6 Summary

At least a quarter of patients suffer some form of depression in the first year after stroke, with the period of greatest risk being in the first months (Hackett et al., 2005). However, stroke survivors still have a higher risk of depression at least 2 years after

the initial stroke, even when controlling for functional impairment and other risk factors (Whyte et al., 2004).

Peak prevalence overall of major depression appears to be 3-6 months after stroke, after which it falls off to around 50% of initial rates at about 12 months, however prevalence can still be high 1 to 3 years after incident stroke. The incidence of PSD appears to be biphasic, being highest in the first few weeks poststroke and then subsiding again thereafter, only to rise again between one and two years poststroke (Whyte & Mulsant, 2002; Paolucci et al., 2006). There is some evidence that the time of onset may be a key predictor of outcome. As many as half of stroke patients who develop major depression in the first 2 to 3 months after stroke are still depressed 12 or 18 months later (Astrom et al., 1993; Berg et al., 2003), however those who develop depression within days of stroke are more likely to have their symptoms remiss spontaneously. Conversely, stroke patients who develop depressive symptoms at 7 weeks poststroke or later are much less likely to make a spontaneous recovery (Andersen et al., 1994). The apparent prognostic value of time of onset has prompted some to suggest subtyping PSD into early and late-onset forms, however there remains no clear consensus as to what time periods would govern this distinction.

There is a firm evidence base suggesting that PSD is significantly associated with a number of negative outcomes, including physical and cognitive impairment and higher mortality, that the presence of depression after stroke has a negative impact on recovery, and that appropriate intervention may ameliorate its impact on rehabilitative outcomes. In order to effectively treat survivors of stroke and other such neurological injuries, it is therefore very important to be able to screen for mood

disorders so that appropriate treatment can be administered. The importance of early intervention is repeatedly underlined throughout the literature (Parikh et al., 1990; Paolucci et al., 1999; Bennett et al., 2006; E. Townend et al., 2007b; Salter et al., 2007; Berg et al., 2009; Salter et al., 2009), and early screening of patients poststroke has now been adopted as a matter of national policy. The National Clinical Guidelines for Stroke now recommend routine screening for depression and anxiety at the onset of stroke and at regular intervals thereafter (Royal College of Physicians, 2012).

The evidence also suggests that the aetiology of PSD is complex. While much research has yielded findings suggesting that PSD impacts negatively on subsequent physical and cognitive function, a body of evidence also suggests that the reverse is true, with PSD being a direct consequence of the physical or cognitive impairment accompanying stroke. These are not mutually exclusive positions, however, and are consistent with PSD's multifactorial origins, where a complex interplay of biological and psychological factors may result in impairments that mutually reinforce one another, leading to a vicious cycle of low mood and impaired function (Dieguez et al., 2004). Furthermore, the differential impact of these factors throughout the time period following stroke may give rise to relationships which fluctuate substantially or even reverse, such as in the apparent role of lesion location reported by Robinson et al. (1984b). The time period following the onset of stroke is therefore a critical factor in making sense of the relationships at play in the causes and consequences of PSD. The failure of some researchers and reviewers to fully appreciate the time-sensitive nature of stroke symptomatology may account for some of the confusing and ostensibly contradictory findings, particularly in assessing the role of neurological damage on its

onset and course (Narushima et al., 2003; Bhogal et al., 2004; Robinson & Spalletta, 2010).

Though there is a growing consensus that the manifestation of depression following stroke is consistent with a biopsychosocial model of mental illness (Whyte & Mulsant, 2002), the notion that the neurological damage that is concomitant to PSD necessarily suggests – or even demands – a predominately biological explanation has been met with some resistance. In fact, the debate over the etiopathogenesis of PSD very much parallels the continuing debate over the aetiology of depression in general, where psychosocial explanations find themselves at odds with reductionist, biological accounts, such as the highly influential monoamine theory of depression. Terms such as 'endogenous' versus 'reactive' have been used to dichotomise depression into subtypes reflecting opinion about whether a particular manifestation is better accommodated by one or the other of these explanations. If depression occurs in the context of a significant environmental stressor such as an adverse life event, then the symptoms may be characterised as 'reactive'; if on the other hand depression manifests without any clear environmental cause, is chronically recurrent, or exists in the context of brain injury or a family history of mental illness, then the condition might be deemed 'endogenous'. One reason that the debate over PSD has been so contentious is that a compelling case can be made for either interpretation, rendering such classification over-simplistic. Though a proportion of PSD may comprise depressive symptoms that might best be characterised as reactive, the neurological damage accompanying stroke adds a more complex biological dimension which may prove difficult to disentangle from the impact of the profound functional impairment that frequently results from stroke.

This comorbidity of depression with the neurological damage that comprises a stroke event, and the unique complications inherent to its manifestation supports the status of PSD as a clinically distinct category. However, despite widespread acceptance of the clinical utility of this subtype, there is still widespread controversy over the role of neurological damage in subsequent depression. Findings regarding the role of lesion location and size have been, and continue to be a contentious issue, and it has proven difficult to arrive at more than a tentative conclusion about this relationship with poststroke depression.

On the one hand Robinson's (Robinson & Coyle, 1980; Robinson & Szetela, 1981) hypothesis for a physiological mechanism at play in PSD appears highly plausible, and Robinson and his colleagues have repeatedly produced evidence implicating lesions of the anterior left hemisphere in the emergence of depressive symptoms during the first few weeks following stroke. Their secondary hypothesis predicting a negative correlation between depressive symptoms and the distance of left hemisphere lesions from the frontal pole in the early stages poststroke has also found compelling support (Lipsey et al., 1983; Robinson et al., 1983; Robinson et al., 1984b; Robinson, 2003; Narushima et al., 2003), and the notion of an acute, organically based syndrome manifesting as depressive symptoms is consistent with the progression of PSD and the pattern of depressive symptomatology across time (Astrom et al., 1993; Berg et al., 2003; Whyte & Mulsant, 2002).

The presence of cases of inappropriate cheerfulness following stroke, furthermore, is particularly compelling evidence of a role of brain lesions in mood dysregulation. It is very hard to see how such a positive response could occur to such a frequently debilitating injury without the neurological damage itself being clearly

implicated as a cause. Robinson et al. (1983), reported that 9% of stroke patients in their study exhibited mood that was clearly incongruent with their circumstances. It is difficult to see how any illness that involves diverse patterns of injury to the brain would not be quite capable of disrupting mood by impacting on the neural structures and systems which regulate it

On the other hand, though, the failure of many other studies to duplicate these findings has resulted in widespread scepticism, particularly Carson et al's (2000) heavily cited meta-analysis in which the role of lesion location was roundly dismissed. However, the compelling rebuttal by Narushima et al. (2003), the findings of Bhogal et al.'s (2004) independent review and continuing evidence supporting Robinson's findings (Robinson & Spalletta, 2010) indicate that a relationship probably does exist, though it is complex and highly dependent on the time elapsed since stroke and the population under examination. In an extensive review of the literature of this area, Salter et al. (2009) concluded that lesion characteristics do not appear strongly correlated with depression, and that psychosocial risk factors and functional impairment are more influential in determining the prevalence and course of PSD.

These uncertainties regarding the aetiology of PSD have been reflected in the caution with which researchers have adopted the term 'poststroke depression'. Many seem wary of the perceived theoretical connotations of the term, and uncomfortable with drawing a line between 'ordinary' depression and that occurring in the context of stroke, so it is common to see papers using alternative, longhand descriptions such as 'depression after stroke'. Both 'PSD' and other longhand references are used here interchangeably to reflect the view that PSD simply means depression in the context

of a significant stroke event, leaving aside any theoretical presupposition vis-à-vis a common aetiology. Some of this, however, also reflects a concern with the preoccupation of much of the literature on depression alone when it comes to exploring and assessing stroke's psychosocial consequences. This preoccupation is understandable given the very high prevalence of depression in this group, however there is also reason to suspect that other types of adverse psychological reactions, such as anxiety or panic disorders are also highly problematic (Barker-Collo, 2007), and so some researchers have been careful to include measures of anxiety and use instruments that include an anxiety subscale such as the HADS, to broaden the scope of their study to encompass a more general state of negative affect, and to reflect this less restrictive construct in their terminology by using broader terms such as 'low mood' or 'emotional distress' (Bennett et al., 2006; Thomas & Lincoln, 2008).

The high correlations reported between measures of anxiety and depression, and the difficulty in distinguishing the two states (Feldman, 1995; Salter et al., 2007) may account for the tendency of researchers to focus on the latter, more prominent form of negative affect. Rather than allowing research to bifurcate into two separate streams, however, – one for poststroke depression and one for poststroke anxiety – there has been a trend towards research into the psychological consequences of stroke being more inclusive of anxiety as a related but distinct area, and for depression and anxiety to be examined side-by-side (Lincoln et al., 2012a).

This also underlines fundamental, empirical and conceptual questions regarding the nature of mood. In assessing, improving upon or constructing mood measures it is essential to first understand exactly what we mean by mood. It is therefore necessary to include in this thesis an examination of theories of affect and

the ways in which its measurement has been approached and conceptualised. If the measurement of mood is not informed by our current state of knowledge regarding the structure of affect, then it is possible that it can be improved upon by the application of a strong underlying theory. Such theories may cast light on some of the difficulties encountered by researchers in defining and measuring mood constructs.

Of particular interest, though, and the primary focus of this thesis is the issue of communication difficulties that are common sequela of stroke. A recurring theme across three decades of PSD research has been the need to exclude participants who cannot communicate by conventional means due to their aphasia. This group represents a sizable proportion of stroke survivors, and the majority of studies exclude aphasic patients, yet the evidence indicates that people with communication difficulties are particularly at risk of developing PSD. Communication difficulty is clearly a significant challenge to assessing mood, and it is important to try and develop ways to overcome this obstacle. In order to do this, however, it is necessary to look more closely at the difficulties faced in assessing depressive symptoms in the stroke population, and explore the efforts that researchers have already made to adapt the methods of mood measurement to address the complications of stroke symptomatology.

3. Measuring Mood in Stroke Patients

3.1 Problems Assessing Mood in Stroke Patients

Two main problems exist with assessing mood in stroke patients. Firstly, many of the symptoms, such as bradykinesia, dysprosody, spatial neglect and vegetative symptoms often mask or mimic the effects of depression. Symptoms such as apathy, poor concentration and psychomotor retardation – which are commonly used as criteria for depression – may be a direct symptom of stroke itself rather than an indication of the patient's mood state.

Secondly, there are the communication problems that may result from stroke. A common symptom is aphasia, where the written or verbal comprehension or expression of language is either partially impaired or, more rarely, completely absent. Aphasia affects around 20-38% of stroke patients (E. Townend et al., 2007b), and communication impairment has been found to be one of the strongest predictors of depression severity and prognosis (Benson, 1973; Astrom et al., 1993; Thomas & Lincoln, 2006, 2008; B. S. Townend et al., 2007a). Furthermore, Kauhanen et al. (2000) revealed that two-thirds of patients with aphasia met the DSM-III-R criteria for depression in the first year following stroke, a figure that was significantly higher than in those without aphasia. The presence of either expressive or receptive aphasia in stroke survivors offers particular challenges to the process of communicating internal mood states. The inability to communicate, or difficulty in communicating, frequently makes conventional assessment methods – such as diagnostic interview or

self-report instruments – impossible to use, and consequently the majority of studies simply omit patients whose communication difficulties make assessment difficult. A study of adaptations made to methods of measuring depression in stroke patients with aphasia revealed that 63% of studies examining PSD excluded patients whose aphasia was too severe for them to be amenable to standard measures (E. Townend et al., 2007b). Hackett and Anderson (2005) also reported that only 3 of 20 studies they reviewed in which predictors of PSD were examined included aphasia as a potential risk factor, underlining an overwhelming tendency for patients with communication impairments to simply be omitted from such studies.

To address communication difficulties resulting from stroke, researchers have broadly taken two approaches. On the one hand, some have focussed their efforts on developing observer-based measures that enable mood to be inferred from objective, behavioural criteria. Another solution, though, has been to develop self-report measures that do not depend on language. Some researchers have instead adopted methods of assessment that rely either partly or completely on symbolic, graphical imagery denoting particular mood states or dimensions of mood, and which are based on the use of a Visual Analogue Scale (VAS). Adaptations of mood measures for stroke patients, however, usually involve an attempt to address both the problems of communication difficulties and somatic symptomatology in the same instrument, so these problems are seldom considered in isolation from one another.

3.2 Observer-Rated Scales

With the difficulties presented in obtaining reliable self-report information from patients with stroke who also have communication difficulties, researchers have

responded by creating adapted scales which rely purely on observable behaviour indicative of a depressed state. The physical symptomatology of stroke means that these behaviours have had to be carefully selected so as to avoid those which may be a consequence of the stroke itself, rather than depressed mood.

3.2.1 The Poststroke Depression Rating Scale (PSDS). Gainotti et al. (1997a) conducted a detailed analysis of clinical symptomatology of major or minor depression after stroke based on DSM-III criteria, and constructed a scale specifically tailored to assess depression following stroke. The scale was also designed with an emphasis on differentiating endogenous depression symptomatology from that of reactive depression. The PSDS comprises 10 sections covering different aspects of depression-related symptomatology in stroke patients; depressed mood, guilt, thoughts of death, vegetative disorders, apathy, anxiety, catastrophic reactions, hyperemotionalism, anhedonia and diurnal mood variations. The PSDS, however, does not provide a global score. It was not designed to give a global assessment of PSD severity but a detailed profile of symptomatology, and is intended to be used solely by a professional examiner. The authors found satisfactory interrater reliability between ratings of researchers, a neurologist and a psychiatrist, with an average correlation of 0.83 across the 10 sections. They also found a high correlation (r=0.88) between scores on 6 of the 10 sections and corresponding items of the HDRS.

Quaranta et al. (2008) examined the accuracy of the PSDS as a diagnostic instrument in a cohort of 143 patients suffering a first time stroke in the previous 9 months. The presence of major depression-like disorder (MDL) or mood disorder with depressive manifestations (MDDM) was first assessed by a specialist using DSM IV-TR criteria. Forty-six (32.2%) were diagnosed as having MDL, and 53 (36.3%) as

having MDDM. Each patient was assessed using the HDRS and PSDS at two different time points one day apart.

PSDS showed good sensitivity and specificity in detecting MDL alone (cut-off ≥ 18; 82.6%, 81.4%) or combined with MDDM cases (cut-off ≥ 9; 84.9%, 84.1%). The PSDS had a significantly higher positive predictive value (PPV) but not negative predictive value (NPV) than the HDRS for MDL (78% vs. 59%), however no significant difference was found between the PSDS and HDRS in predicting the diagnosis of MDL and MDDM combined.

3.2.2 The Stroke Aphasic Depression Questionnaire (SADQ). Sutcliffe and Lincoln (1998) set out to address this problem by developing the Stroke Aphasic Depression Questionnaire (SADQ). The scale was initially designed for aphasic patients living in the community, and to be used by a spouse or caregiver. The SADQ was developed from depression questionnaire items relating to observable behaviour, and which excluded behaviours likely to be associated with the immediate symptoms of stroke itself, with items being scored on a scale of 0-3, based on the frequency of the observed behaviours. Beginning as a 21-item questionnaire, it was subsequently reduced to 10 items by eliminating items that that did not differentiate significantly between depressed and nondepressed patients, though the original version was retained as the SADQ-21. The condensed form of the scale – the SADQ-10 – demonstrated good internal consistency, yielding a Cronbach's α of 0.80, and split-half correlation of 0.81. The SADQ-10 showed correlations with the HADS-D of 0.32, the HADS-A of 0.63, and the Wakefield Depression Inventory (WDI) of 0.67. Test-retest reliability after a four week interval was 0.69. The authors concluded that the scale has good internal consistency and moderate validity. The original SADQ

questionnaire was intended for use in the community, but hospital versions (SADQ-H21 and SADQ-H10) were subsequently produced in which the response format was modified so as to stipulate more specific criteria for the frequency of observed behaviours.

Leeds (2004) examined the validity of the SADQ-10 against the Geriatric Depression Scale (GDS-15) in a sample of 65 stroke patients at a rehabilitation unit and found a correlation of 0.4 between the two scales. An optimal cut-off for the SADQ-10 of 14/30 was found, with a sensitivity of 70% and specificity of 77%, however the author cautioned the use of SADQ as a depression measure in patients without significant aphasia.

More positive findings emerged in Bennett et al.'s (2006) study in which a number of measures for assessing mood in stroke patients were validated against the HADS in a sample of stroke patients and healthy older adults. In the stroke patients, a significant correlation was found between measures of the SADQ-H10 with both the HADS-D (r=0.53) and the HADS-A (r=0.33), with the scale demonstrating a sensitivity of 100% and a specificity of 78% using a cut-off score of 5/6. Internal consistency for the SADQ-H10, was good, with a Cronbach's α of .68, though the SADQ-H21 fared much better, with a Cronbach's α of .84.

In a study of 125 patients from an acute, in-patient stroke unit, Hacker et al. (2010) examined the validity of the SADQ-H10 against the Brief Assessment Depression Cards Schedule (BASDEC), a screening tool for elderly in-patients (Adshead et al., 1992). The SADQ-H10 was found to discriminate between depressed and non-depressed patients and scores were significantly correlated with those of the

BASDEC (r=0.46), with a cut-off score of >6 yielding a sensitivity of 68% and a specificity of 79%.

In the only study to date in which the study sample consisted exclusively of stroke patients with aphasia, 165 patients recruited through hospital wards and community services were administered the SADQH-10 alongside the VAMS and the VASES (see 3.3). A significant correlation was found with the VAMS 'sad' item (r=0.297), but not the VASES depression item. Internal consistency was also good, with a Cronbach's α of 0.77 (Cobley et al., 2012).

3.2.3 The Aphasic Depression Rating Scale (ADRS). Before these later studies yielded stronger evidence for the suitability of the SADQ, Benaim et al. (2004), expressing concerns that the existing evidential validity of the SADQ might not be generalisable to the aphasic stroke population, set out to construct a scale of their own. The methodology for developing their scale was more elaborate than that of the SADQ, incorporating the input of many professionals involved in the clinical rehabilitation of aphasic stroke patients. Based on interviews with 18 members of a neurorehabilitation team, the most frequently reported behaviours reported in aphasic stroke patients were noted. Six experts then analysed the items on three depression scales, including the HDRS and MADRS, and selected items which corresponded to those behaviours identified. Based on consensus between these experts, a pool of items was generated which was deemed most likely to effectively detect and quantify depression in aphasic stroke patients. After a preliminary analysis, 9 items were retained to comprise their new scale, named the Aphasic Depression Rating Scale (ADRS).

The scale was validated in 50 aphasic and non-aphasic stroke patients using the HDRS for non-aphasic patients, and by a psychiatrist and the rehabilitation team using a Visual Analogue Scale (VAS) (see 3.3.1) for the aphasic patients. Both inter-rater reliabilities and test-retest reliabilities were high (r=0.89). The scale also demonstrated good criterion reliability against the HDRS, rating given by a psychiatrist, and ratings given by a rehabilitation team using a VAS (r=0.77; r=0.60; r=0.78, respectively). A cut-off ADRS score of ≥ 9/32 gave a sensitivity of 83% and a specificity of 71% compared with the diagnosis made by a psychiatrist.

Further evidence of the validity of the ADRS emerged in a study focussed more specifically on the instrument's sensitivity to change (Benaim et al., 2010). Forty-nine stroke patients admitted to two rehabilitation units were assessed twice at a one-month interval. At each assessment patients completed the ADRS and the VAMS 'sad' item. They were also examined by a trained psychologist and given a score (PSY) on scale of 0 to 10 indicating the level of apparent depressive symptoms. ADRS scores were found to significantly correlate with both PSY and VAMS measures at both the first (r=0.71; r=0.65) and second (r=0.52; r=0.64) assessments; furthermore, changes in ADRS scores were correlated significantly with changes in PSY scores (r=0.72). The authors conclude that these findings offer further evidence of the scale's concurrent and convergent validity with respect to other depression measures, and demonstrate that the instrument has good sensitivity to change. Sensitivity and specificity figures, however, were not provided.

3.2.4 The Signs of Depression Scale (SODS). The SODS (Hammond et al., 2000) was developed in response to evidence that up to one third of older patients are significantly impaired in their ability to communicate due to a variety of medical

conditions, including aphasia. The objective was to produce a scale based on observable behaviour from the criteria for depression detailed in the DSM-III-R (American Psychiatric Association, 1987), and from symptoms known to distinguish depressed people from those who are physically ill.

An initial, nine-item scale of a yes/no format was constructed and then examined in a preliminary study, from which three items with poor sensitivity, specificity or inter-rater reliability were omitted. The revised, six-item scale was then validated in a group of consecutive patients admitted to acute geriatric wards. Both of these studies used the Geriatric Mental State Schedule (GMS) for diagnosis, with the initial study also employing the HDRS.

The validation study results revealed a sensitivity of 90% and a specificity of 72% for the six-item version of the SODS based on a cut-off of ≥ 3. Re-analysis of the initial study data revealed a correlation (Spearman's rho) of 0.79 between the six-item scales score and the HDRS. The authors concluded that the SODS showed good psychometric qualities for detecting the presence of depression in geriatric inpatients.

Watkins et al. (2001) examined the utility of the SODS in 137 patients from a group of consecutive acute stroke admissions to a teaching hospital. Volunteers who survived the first week underwent a clinical interview using the MADRS, with a cut-off of >6 used to give a diagnosis of depression. On the same day, a primary nurse was asked to independently rate the patients using the SODS, and the scores were compared to the MADRS diagnosis. The results revealed a sensitivity of 81% and a specificity of 38% for a SODS score of >1, with other cut-offs demonstrating inadequate sensitivity.

Bennett et al.'s (2006) study of screening measures also examined the SODS in relationship to their criterion measures the HADS-D and HADS-A in a sample (*n=100*) of stroke patients. As in the previous study, good sensitivity (86%) was noted at a cut-off of SODS score of >1, however specificity (62%) – though better – was also lacking. The authors concluded that the SODS was less suitable than the SADQ-H10 as a screening measure because of its poor sensitivity/specificity.

3.3 Self-Report Scales

Though observer-rated scales are an obvious way of circumventing the difficulties in assessing mood where patients cannot communicate their internal state using language, depression includes complex phenomenology that is accessible only to the person experiencing it. Self-report is therefore the primary means by which much of the required, diagnostic information must be collected. Another approach to assessing depression in people with communication difficulties has therefore been to develop modified methods of self-report that reduce or entirely eliminate the need for language. Central to these methods are simplified ways of communicating the degree to which a particular mood state or symptom is felt, and using simple words or pictures to communicate what a patient is being asked about. These simplified response formats have by and large been focussed on what have now become known as *Visual Analogue Scales*, or variants thereof. These have been used – often in conjunction with pictures – as a way of addressing communication difficulties.

3.3.1 Visual Analogue Scales (VAS). The VAS was developed as a simple, scalar rating method to allow collection of detailed, self-report data unconstrained by forced-choice categories (Hayes & Paterson, 1921). Originally termed as the *Graphic*

Rating Scale, it consists of a simple, horizontal line with dichotomous descriptors at either end, often with intermediate descriptors along its length. A rater is asked to mark a point on the line which best quantifies the variable under investigation, with the distance of the mark along being treated as a scalar measure. This type of scale has benefits over other rating methods; in addition to the fine discrimination and lack of dependence on imprecise categories, it is simple and easily grasped, it is interesting and requires little motivation on the part of the rater, it is quick to fill out and easily scored (Freyd, 1923). Furthermore, it shows good test-retest reliability, and its inventors claimed an inter-rater reliability of 0.65 when using the method to assess the quality of employees' working practises (Hayes & Paterson, 1921). As different variants of this scale format were tested and adopted, the term *Visual Analogue Scale* emerged to describe scales in which descriptors are placed only at the end points of the line, with no intermediate descriptors along its length, while a *Graphic Rating Scale* (GRS) came to denote those variants in which descriptors also appear along the length of the line (Scott & Huskisson, 1976; Wewers & Lowe, 1990).

Before it was applied to mood measurement, the VAS format found many applications in other areas of psychology, including the measurement of a variety of symptoms and the perceived intensity of physical sensations (D. D. Price et al., 1983). With the more widespread adoption of the VAS response format, however, came questions about its psychometric properties compared to other response formats, and how its orientation, size and type could impact on these properties. A VAS may be horizontal or vertical, for example, and though it is by convention 100mm in length, different length versions have also been used. The VAS may also have descriptors along its length (as with the GRS), with the lengths of the words and the space

between them (or the distance between letters) varying from one scale to another, and while the VAS usually consists of a plain, solid line with anchors at each end, variants exist in which the lines are marked at regular intervals.

To shed light on these questions, Scott & Huskisson (1976) compared six visual analogue and graphic rating scales designed to measure the intensity of pain. The scale were anchored at each end with the descriptors “no pain” and “pain as bad as it could be”. Three of the scales were horizontal GRSs, each with a different distribution of words along its length. The first had the words “severe”, “moderate” and “mild” at even points along the scale, while a second had the same words, but their letters spread out so that they spanned the full length of the scale, with little room in between words. A third scale had just two descriptor words “severe” and “slight” about a sixth of the way from each end of the scale. The other three scales were vertical, the first a pure VAS with no intermediate descriptors or markings, the second with the words “severe", “moderate” and “mild” spread equidistantly along its length, and the third with 20, equally spaced, numbered markings from one end to the other.

The authors studied these scales as part of a study of an analgesic drug, with the different forms of the scale, plus a simple, descriptive pain scale being used by participants to report the degree of experienced pain at intervals after receiving a dose of the drug. Results revealed that only data collected using the vertical VAS and the horizontal GRS with words uniformly spread along its length had a uniform distribution (vertical VAS: χ^2=8.61; horizontal GRS: χ^2=2.92). The presence of numbered markings or descriptors at intervals along its length, conversely, resulted in non-uniform distributions.

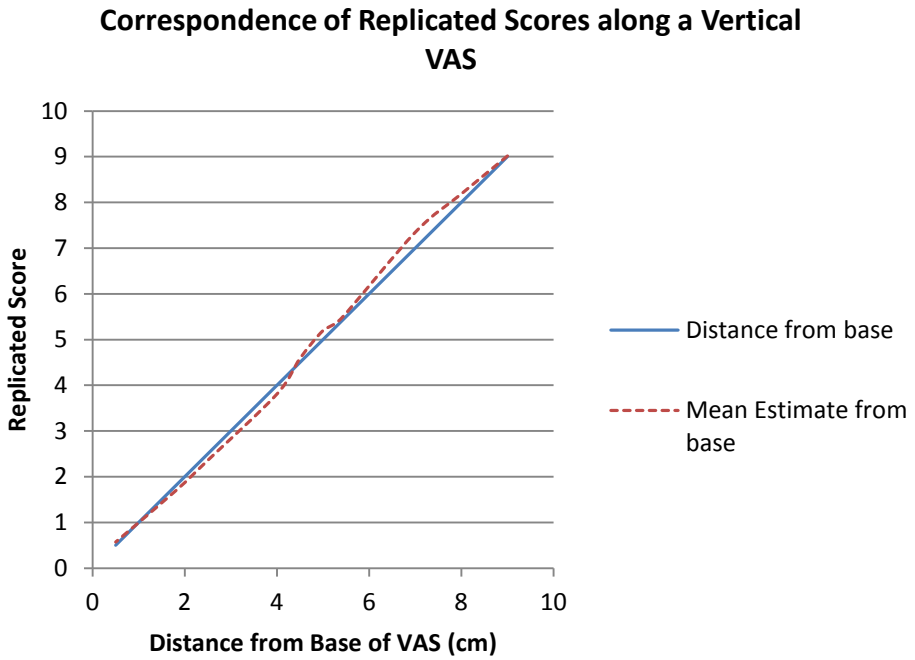
Having found evidence suggesting the superiority of VAS measures over those of the GRS format, the authors then turned to the question of whether the vertical or the horizontal VAS had better psychometric qualities. In a study of 100 rheumatoid arthritis patients attending a clinic (Scott & Huskisson, 1979), volunteers completed a vertical and a horizontal VAS pain scale in random order. The results revealed no significant difference between the distribution of scores on vertical and horizontal scales. There was also a very high correlation between the two scales (r=0.99).

A study by Dixon and Bird (1981) addressed the important question of how accurately people are able to judge and reproduce distances on a VAS. Eight volunteers were presented consecutively with 10 vertical VAS lines each with a cross marked at a random point along its length. The participant was asked to reproduce the position of the cross on an unmarked VAS. This was repeated six times, giving a total of 560 measurements. The results revealed that mean scores were accurate to within ± 1.5% of the reproduced position, however there was some variation in the accuracy along the line. The most accurate estimates were at the endpoints and the midpoint, but positions ± 2cm from the midpoint appeared most difficult to reproduce, with a marked increase in standard deviation of scores in these areas (see Fig 3.1). Furthermore, there was a tendency to underestimate scores closer to the base of the line, while overestimating those closer to the top.

Returning to the question of how psychometric qualities of different VAS formats compare, Sriwatanakul (1983) studied five different types of VAS: two horizontal, one vertical, and two curvilinear. 107 volunteers were asked to use them to rate different degrees of pain, and also to rank them in order of preference. The results showed that scores on the vertical scale had the greatest variation coefficient and the

least normal distribution, echoing the findings of Dixon and Bird (1981). As in Scott & Huskisson's (1979), study, scores on the horizontal scale also tended to be slightly lower than those on the vertical scale.

In another comparison study, Gift et al. (1989a) examined validity of a horizontal and a vertical VAS ("HVAS" and "VVAS" respectively) in self ratings of patients with dyspnea (shortness of breath). Peak expiratory flow rate (PEFR) gave an objective measure of air flow to the lungs, against which the scales' validity could be assessed. As in Scott and Huskisson's (1979) study, there was a very high correlation between the HVAS and VVAS (r=0.97), and also between the VVAS and the PEFR measures which served as a criterion measure (r=-.85). In a review of the VAS as a measurement instrument, Gift (1989b) also concludes that a vertically oriented VAS is more sensitive, yields higher scores, and is easier to use than a horizontal VAS.



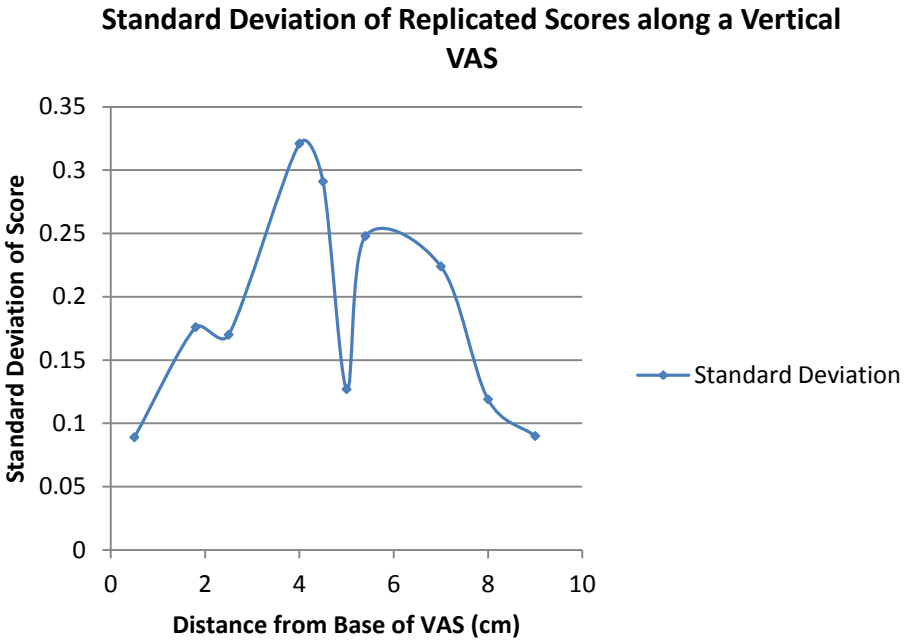


Fig 3.1 *Accuracy of a vertical Visual Analogue Scale, from data in Dixon and Bird (1981)*

The evidence is therefore strong that ordinary individuals can use a VAS response format to provide a valid and reliable measure of experienced stimuli. Some, however, have raised questions as to whether these findings can be extended to stroke patients, whose comorbid cognitive impairment may cause difficulties in making the judgements involved. If a VAS cannot easily be used by people who have had a stroke, then there will be implications for whether VAS-based instruments can be usefully adopted in this population.

C. I. Price (1999) has argued that many stroke survivors are not capable of using a VAS. In a study of 96 participants within 6 months of stroke, and 48 control subjects, participants were asked to use five different types of VAS to rate the level of tightness of a sphygmomanometer cuff at three different settings. They were also asked to compare the pressures at two different settings to confirm that they were

capable of distinguishing different pressures. The results showed that people with stroke were less likely to complete any VAS scale compared to control subjects, with significant differences in all of the scales. The findings did, however, support those of Gift (1989b), in that the results appeared to favour a vertical rather than horizontal scale, with 88% versus 81% of the control group correctly using the vertical scales. Notably, the presence of aphasia was not related to mistakes made on the scales.

In a review of research using the VAS with a criterion instrument, Wewers & Lowe (1990) concluded that though the VAS has some attractive features for measuring many subjective phenomena, there are some shortcomings and limitations: Some researchers reported participants having difficulty understanding how it is meant to be used. A VAS should be restricted to simple, unidimensional constructs that are clearly defined, and using unambiguous scale endpoints. Though test-retest measures are frequently used to assess reliability, the dynamic and subjective nature of the constructs under investigation may make reliability measures hard to interpret.

3.3.2 Visual Analogue Mood Scales (VAMS). The first documented use of the VAS response format in connection with mood measurement was in Zealley and Aitken's (1969) study of 13 hospitalised patients treated for depression and bipolar disorders. Using scores from a 100mm, horizontal VAS, along with HDRS scores and scores from a psychiatric assessment, the authors reported high correlations between both the VAS and HDRS (r=.79) and the VAS and overall assessment scores (r=.78).

Folstein and Luria (1973) subsequently adopted a modified version of Zealley and Aitken's scale, naming it the *Visual Analogue Mood Scale* (VAMS), detailing its clinical application in assessing patients for mood disorders and examining its reliability and validity. Folstein and Luria's VAMS consisted of a 100mm by 35mm

card, with a horizontal, bipolar VAS between end points signifying best and worst possible moods. This was accompanied by the instruction: "How is your mood right now? A mark on the line toward the left represents your worst mood, toward the right, your best." In their study of two groups of hospitalised patients primarily with psychiatric conditions (*n*=133 and *n*=31), the authors found the VAMS to have good criterion validity, with correlations of *r*=-.64 and *r*=-0.67 respectively between VAMS and concurrent ZDS measures. Within-patient re-test reliability was computed by correlating odd and even day scores within each patient and producing a weighted average. Within-group test-retest reliability was computed by correlating group scores between consecutive days throughout the test period and then averaging these. Within-group test-retest reliability of the VAMS was good (0.61 and 0.73 respectively), however within-patient reliability was poorer (0.32 and 0.48 respectively). Luria (1975) later conducted a replication of this study on a more representative sample of 62 hospitalised, psychiatric patients. The results again showed significant correlations between the VAMS and these two other measures, with correlations of -0.56 and -0.77 with the ZDS and Clyde Mood Scale respectively. Measures of test-retest reliability after a two hour interval were also significant, with mean correlations for groups varying between 0.56 and 0.8.

A note of caution is due here regarding the use of test-retest reliability in scales measuring aspects of mood. Assessing the validity of mood scales in this respect is problematic, as mood is by definition in a state of constant flux. Measures of this type must be approached with care, and reliability statistics must be interpreted in the context of the sample under investigation, and the length and quality of the intervals between the test and re-test phases of a study. (Wewers & Lowe, 1990). A

review of VAS based mood scales by Ahearn (1997), for example, showed that test-retest reliability was highly sensitive to the amount of time elapsed since the previous measure, with test-retest reliability measures taken within 2 hours yielding figures almost twice as high as when measures are taken 24 hours apart.

Change in mood across time, however, is something that a scale must be sensitive to in order for it to be of clinical use. It was with this in mind that Little & McPhail (1973) set out to examine the performance of Aitken's VAS over an 18 month period. Monthly measures of depression were taken using the VAS, the BDI and psychiatrist in a sample of 8 female outpatients. High correlations were found between the psychiatrist VAS ratings and patient VAS ratings (ρ=0.8), between psychiatrist VAS ratings and patient BDI scores (ρ=0.76) and between patient VAS ratings and patient BDI scores (ρ=0.76), demonstrating good concurrent and inter-rater reliability. The authors remark that it is the ability of the scale to detect shifts in mood from one month to the other which is of most value, and which this study was successful in demonstrating in the VAS, and that the speed with which the VAS can be administered makes it particularly desirable.

Though all of the mood scales used in these studies were based on a VAS, instructions and descriptors were in writing and the scale was not designed to accommodate communication difficulties. It was not until 1990 that the use of a VAS in connection with a completely nonverbal mood scale was documented. To address the communication difficulties and cognitive impairments arising as a result of stroke or other neurological damage, Stern and Bachman (1991) produced a modification of Aitken's scale specifically with this population in mind. The VAS was made vertical, so as to accommodate the left or right-side hemispatial neglect that is a common

consequence of stroke, and simple ‘smiley’ pictograms denoted bipolar, ‘happy’ and ‘sad’ endpoints, by which the corresponding words were also included. This, they called the *Visual Analogue Dysphoria Scale*, or VADS.

Fig 3.2 *Visual Analog Dysphoria Scale (VADS) - Adapted from Stern et al. (1997)*

Stern et al. (Stern et al., 1997; Stern et al., 1990) later expanded upon this format to give 7 unipolar scales each representing a single mood state. These scales, comprising the items ‘happy’, ‘sad’, ‘afraid’, ‘angry’, ‘tired’, ‘energetic’, and ‘confused’ were also named *Visual Analogue Mood Scales* (VAMS). As in the VADS, graphical faces were used to denote endpoints of a single, 100mm, vertical

VAS, however these scales were in a unipolar form, with ‘happy’ and ‘sad’ forming separate scales alongside those for the other mood states. At the top of each scale was a ‘neutral face’, and at the bottom was a face for the respective mood item, accompanied by descriptors words at both ends. Thought the initial VAMS had seven items, a ‘tense’ item was later added.

In a study of 171 university students, Stern et al. (1997) validated these scales against the Profile of Mood States (POMS) (McNair et al., 1971). The POMS is a 65-item adjective checklist in which scores are derived for six mood states: ‘tension-anxiety’, ‘depression-dejection’, ‘anger-hostility’, ‘vigour’, ‘fatigue’ and ‘confusion’. The authors examined two versions of the VAMS: the ordinary version with mood words included, and another version without them (in order to simulate somewhat how an aphasic patient might respond). One group completed the VAMS, the POMS and then the VAMS again presented in a different order, while another group completed the same protocol using the no-word version of the VAMS.

Convergent validity was supported by significant correlations with corresponding POMS scores for the ordinary ($mean_{[r]} = 0.51$; $range_{[r]} = 0.33$ to 0.66) and no-word version ($mean_{[r]} = 0.55$; $range_{[r]} = 0.33$ to 0.77) of the VAMS, while discriminant validity was supported by much lower cross-correlations between items ($mean_{[r]} = 0.16$; $range_{[r]} = 0.003$ to 0.46). Test-retest reliability was assessed by correlating the first and second sets of VAMS scores, yielding a mean of r=0.68.

In the same report, the authors described a further validation study in an independent sample of 140 university students. Participants were given the original (with-word) version of the VAMS along with the POMS, the BDI and other measures. Again, convergent validity was demonstrated by large correlations between

corresponding VAMS and POMS items ($mean_{[r]}$ = 0.64; $range_{[r]}$ = 0.51 to 0.72) and some discriminant validity was demonstrated by smaller cross-correlations between dissimilar items ($mean_{[r]}$ = 0.35; $range_{[r]}$ = 0.13 to 0.56). Correlations between the BDI and VAMS items offered further evidence of validity, with the 'sad' item (r=0.53) and the 'angry' item (r=0.51) showing the highest convergence.

As part of a study to establish normative data for the VAMS, Nyenhuis et al. (1997) attempted to replicate these findings using a similar methodology. A sample of 400 adults and 175 geriatric (age>55) participants were administered the original VAMS, along with the POMS and BDI. Strong correlations between VAMS and corresponding POMS items ($mean_{[r]}$ = 0.63; $range_{[r]}$ = 0.55 to 0.69) observed, and BDI scores correlated significantly with the VAMS 'sad' item (r=0.54). A principle components analysis revealed two factors – labelled "negative mood" and "energy" – accounting for a total of 61.8% of the variance. However while Stern et al. (1997) sensibly included both the 'with-word' and 'no-word' versions of the VAMS in their study, this study only used the version of the VAMS in which the words were included. Since the pictures are intended to communicate mood states to people with impaired comprehension of written language, it would seem inappropriate to use the accompanying descriptors in a study population who have no such impairment. If the words can be used as the primary source of affective information, then this circumvents the need to rely upon the pictures. This essentially reduces the VAMS to an adjective checklist much like the POMS against which it was validated, making the significance of these findings questionable.

In a study of 25 inpatients referred for ECT following a major depressive episode, Arruda et al. (1997) examined the VAMS' responsiveness to change from a

therapeutic intervention. Each participant was administered the 7-item VAMS and HDRS before their first ECT, and then again 1-4 days after their last ECT. All VAMS items showed significant improvement from pre-ECT to post-ECT measures, with 'sad', 'happy', 'tired' and 'energetic' items being most sensitive to change. Percentage change in both the VAMS 'sad' item and the HDRS was associated with clinical post-test improvement as measured by CGI (Clinical Global Impressions) rating provided by a psychiatrist (r=-0.57 and r=-0.57) and CES-D (r=-0.49 and r=0.52). Notably, the VAMS 'sad' item was shown to be as sensitive to change in depression symptoms as the far lengthier HDRS, demonstrating that the VAMS 'sad' item score may be useful in assessing change in depressive symptoms.

All of these studies, however, share the common limitation they did not include people with communication problems or neurological impairment, raising questions as to whether their findings extend to this population. The very problems posed in assessing mood in patients with communication difficulties, and which the VAMS was designed to address, also make validation in this population particularly challenging, as the standard, language-dependent measures against which the VAMS can be compared are ones which are of limited use in people with communication difficulties.

In a study of 41 inpatients admitted to a medical centre either for acute stroke or for poststroke rehabilitation, Arruda et al. (1999) struck a balance between these considerations by selecting stroke patients with some aphasia, but whose single word recognition remained intact. Patients meeting this criterion were recruited and tested within 28 days of stroke and administered the ordinary (with-word), eight-item version of VAMS along with a modified, yes/no version of the POMS. Good

correlations were found between the VAMS items and their POMs counterparts ($mean_{[r]}$ = 0.68) and poor correlations between non-corresponding scales ($mean_{[r]}$ = 0.14); mean reliability was 0.55.

However this study had the same fundamental flaw as that of Nyenhuis (1997). The sample studied necessarily needed language intact in order to rate themselves on the language based POMS, yet the standard (with-word) version of the VAMS was used, rather than the no-word version originally used as validation. Again, since the pictures are intended to communicate mood states to those whose communication difficulties may preclude the comprehension of written language, it seems inappropriate to use the accompanying descriptors in a study population who have no such impairment. People who can read do not need to interpret the picture at all; they do not need to recognise that the 'sad' face means sad, all they have to do is read the word "SAD", which is written right there on the card below the picture. So the only thing that these studies may be demonstrating is the correlation between participants responses to the VAMS mood words ('sad', 'afraid', 'angry', 'tired', 'energetic', and 'confused') and synonyms of these items on the POMS ('depression-dejection', 'tension-anxiety', 'anger-hostility', 'vigour', 'fatigue' and 'confusion').

A later study by Temple et al. (2004) also sought to address the dearth of evidence for the validity of the VAMS in people with cognitive impairment. Co-authored by Stern and Arruda, this study employed an almost identical design to its predecessors, except this time the study sample was comprised of 31 patients with clinical dementia. The eight-item VAMS was validated against a simplified, yes/no version of the POMS subscales 'depression-dejection', 'anger-hostility', 'vigor-activity', 'fatigue-inertia', and 'confusion-bewilderment'. Again, convergent validity

was demonstrated by fair correlations overall between the VAMS items and their POMs counterparts ($mean_{[r]} = 0.54$) and relatively poor correlations between non-corresponding scales ($mean_{[r]} = 0.14$) though their levels of significance were not reported. The authors concluded that the VAMS is of clinical use in patients with dementia, with the 'confused' item proving particularly useful in their sample.

Outside of these papers, validation evidence has been sparse, however there have been two notable studies that examined the utility of the VAMS to assess mood in stroke patients. In both of these studies, the VAMS scales have been used more loosely as a unitary measure of valence rather than as metrics of the individual mood states comprising them, with composite scores or measures of internal consistency (Cronbach's α) being assigned to VAMS scores overall, by either dropping the positively valenced items 'happy' and 'energetic' (Bennett et al., 2006) or reversing them (Kontou et al., 2012).

Bennett et al. (2006) assessed the validity of a number of potential screening measures for depression in stroke patients, examining the VAMS alongside other instruments including the HADS and the SADQ. When the 'happy' and 'energetic' items were omitted, the authors reported highly significant correlations between VAMS ('VAMS -HE') scores and the HADS and both of its subscales (HADS: r=0.42; HADS-D: r=0.35; HADS-A: r=0.40). Though a cut-off score of 22/23 was identified for the VAMS 'sad' item, the sensitivity and specificity of this item (0.88 and 0.62 respectively) was deemed too poor for the VAMS to be recommended as a screening measure. The authors concluded that though the VAMS offers a broad measure of severity of depression, it is of little use as a screening instrument.

Kontou et al. (2012), observing that the reversed valence of the 'happy' and 'energetic' items may be confusing to people using the VAMS, produced a version of the VAMS (the 'VAMS-R') with the scales for these two items reversed, such that that the symbols at the top of the scales are the ones which are most positively valenced. So while the 'neutral' VAMS faces are usually at the top of the scale, this was reversed for the happy and energetic items, such that the graphic at the top of a scale is always the most positively valenced of the two.

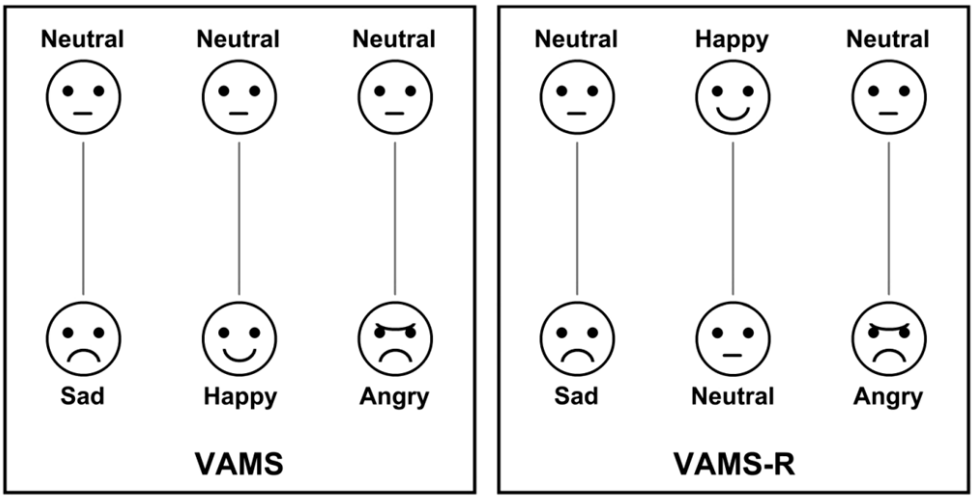


Fig 3.3 *Visual Analogue Mood Scale VAMS & VAMS-R, with example of reversal of positively valenced items in modified version - Adapted from Kontou et al. (2012)*

This revised version of the VAMS was tested alongside the HADS, SADQ-H-21 and VASES in a study of 50 healthy older people and 71 aphasic stroke patients. The results revealed the VAMS-R correlated significantly with the HADS in healthy adults (HADS Total: r=0.62; HADS-D: r=0.49; HADS-A: r=0.59), and with the VASES and SADQ-H21 in aphasic stroke patients (VASES: r=-0.69; SADQ-H21:

r=0.43). Internal consistency of the VAMS-R was high in both healthy adults (0.74) and aphasic stroke patients (0.80).

Since Benaim et al.'s (2010) study (see 3.2.3) used the VAMS 'sad' item as a criterion measure for the ADRS, the results are also worth noting here. Correlations between these two measures in stroke patients assessed at two time points one month apart were r=0.65 and r=0.64 respectively.

In other studies which have utilised the VADS or the VAMS, these tools have generally been used as a supplementary measure to aid in assessing those with aphasia or other cognitive impairment (Gainotti et al., 1997a; Paolucci et al., 1999; Paolucci et al., 2006), in line with the creators' cautionary advice that such tools are not intended for use in isolation (Stern et al., 1997). However where the VAMS has been used, it has almost always involved employing only the 'sad' item, rendering it effectively as a unipolar version of the original VADS.

In studies which adopted the VAMS or VADS as a measure of depression, however, it has become clear that a significant proportion of people have problems understanding these scales and how they are supposed to be used. House et al. (1989), for example, describe patients presented with the Folstein's original VAMS as "bewildered by it", while Gainotti et al. (1997a) reported that only 9 of 23 participants could use the VADS, and that participants were unable to understand nonverbal gestures used in an attempt to communicate how the scale is to be used. Authors of the aforementioned Benaim et al. (2010) study also reported difficulties, with 18% of patients having trouble understanding directions to complete the VAMS. Difficulty in understanding the VAMS is also evidenced in the way that it has sometimes been

simplified as a binary instrument, whereby patients simply point to the 'sad' face if that is how they feel, rather than using the VAS itself (Paolucci et al., 2006).

These communication issues highlight a fundamental problem of such nonverbal instruments: Though the instruments may offer a nonverbal response format, only verbal instructions are provided to describe how to use the scale. Though Stern (1997) instructs researchers using the VAMS to explain "through both word and gesture" how the scales are to be used, this is not very helpful as it is hard to see how such a concept can be translated into simple gestures by the assessor.

In a systematic review of 60 papers in which depression or low mood was studied in aphasic stroke patients, Townend et al. (2007b) examined the way in which techniques for diagnosing depression were adapted to accommodate the difficulties of people who have aphasia. The findings again highlighted the difficulties that many people with communication problems had using the VAMS; they also echoed the concerns of C. I. Price (1999), observing that the concept behind the use of the VAMS is quite abstract and might therefore prove difficult for people who are cognitively impaired. They concluded that "on balance, the visual analogue scales reviewed provide neither a suitably comprehensive nor a particularly successful method of diagnosing depression in aphasia" (E. Townend et al., 2007b, p. 3081).

Berg et al. (2009) tested a variety of screening instruments for depression after stroke in a hospital based study of 100 consecutive stroke patients at 2 weeks, and 2, 6, 12, and 18 months following stroke. The presence of aphasia was assessed with the Western Aphasia Battery (Shewan & Kertesz, 1980) (score $\leq$ 93.7), with 31% of patients being classified as aphasic. The findings revealed that VAMS was not significantly correlated with the BDI at any time point in those with aphasia, and only

correlated significantly at 18 months in patients without aphasia. Sensitivity and specificity characteristics were also poor, with a VAMS cut-off of ≥ 50 offering figures of 60% and 87% respectively. The authors concluded that: "the use of Visual Analogue Mood Scales amongst patients with aphasia and other cognitive impairments cannot be recommended" (Berg et al., 2009, p. 523).

3.3.3 Visual Analogue Self-Esteem Scales (VASES). The VASES emerged from a study by Brumfitt and Sheeran (1999) in which they aimed to produce a measure of self-esteem that did not require the use of complex language. The scale was based on a semantic differential measure of self-esteem (SDMSE) constructed for the purposes of validation, which consisted of 24, 7-point bipolar scales comprising a range of item related to self-esteem, (such as "confident/unconfident", "talkative/quiet" etc.). Pairs of pictures corresponding approximately to the SDMSE items were created (one for each scale endpoint) and tested in a sample of 243 student participants, who also completed the SDMSE, the Rosenberg Self-Esteem Scale (RSE) (Rosenberg, 1965), the GHQ and the HADS. The ten highest SDMSE-correlating items where then selected to create the final score of the VASES.

The internal consistency of the VASES was good (Cronbach's α = 0.86), as was test-retest reliability of ratings taken one month apart (0.73). Correlations with the SDMSE and the RSE were also high (r=0.76 and r=0.6). Notably, correlations with the HADS and GHQ depression and anxiety subscales were also significant (HADS-A: -0.51, HADS-D: -0.51, GHQ-A: -0.45, GHQ-D: -0.39).

The VASES was then tested in two small samples of people (n=14, n=20) with some degree of aphasia, but with language largely intact. The first of these groups completed a version of the VASES which had the corresponding SDMSE-item word

labels printed above each of the pictures (such as “Optimistic”/“Pessimistic”), along with the RSE and GHQ. The second of these groups completed a no-word version of the VASES, along with the RSE and the HADS. As would be expected, the former group whose pictures were labelled showed high correlations with the GHQ criterion measures (-0.85 for both GHQ-A and GHQ-D), whereas the latter group with the no-word versions of the pictures fared worse against the HADS, though it still correlated well with the HADS-D (HADS-A: -0.35, HADS-D: -0.64). Both versions of the VASES had good internal consistency with a Cronbach’s α of 0.78 and 0.87 respectively.

Though the VASES, was designed as a measure of self-esteem and not depression, the overlapping nature of the two constructs and the fair-to-good criterion validity shown against HADS-D mean that it has sometimes been used as an indirect measure of depression (Thomas & Lincoln, 2008; Cobley et al., 2012). Elsewhere in the literature, though the VASES has shown good correlations with HADS-A/HADS-D, sensitivity and specificity have been too poor for it to be recommended as a screening instrument (Bennett et al., 2006), and unlike the VAMS ‘sad’ item, it did not significantly correlate with the SADQ (Cobley et al., 2012).

3.3.4 Disc Intensity Scale Circles (DISCs). The DISCs (Turner-Stokes et al., 2005) was designed for patients with language and cognitive impairment following acquired brain injury (ABI). Noting the problems that some people have using alternative instruments such as those using a numbered graphic rating scale (NGRS) (such as a Likert scale) or VAS response format, the authors developed a six point, vertical scale comprising six circles, each with a progressively larger area of dark grey shading. The fully shaded circle at the top represented the most depressed

state, while the empty circle at the bottom represented as state of no depression. A pictorial variant of the scale was also created, in which a simple, line-drawn sad face accompanied the top circle, while a 'smiley' accompanied the circle at the bottom.

It was hoped that as well as providing support for people with communication problems, this tool might be of particular assistance to people with visuo-spatial impairments that often accompany ABI, and which might impact upon a person's ability to scale abstract quantities into numeric scores.

The DISCs were validated in a cohort of 114 patients recruited at a service for young adults with ABI. Participants able to respond to questions were assessed using a semi-structured interview, based on which a gold standard, DSM-IV diagnosis was made and a score assigned using the Beck Depression Inventory (BDI-II). They were also asked The Yale question, "Do you often feel sad or depressed"? Those unable to respond to interview questions were assessed using the SADQ. Participants then completed DISCs alongside a conventional NGRS.

The results revealed that DISCs scores correlated significantly with the NGRS measure (r=0.87) and with the BDI-II (r=0.66). Using a cut-off of DISCs ≥ 2, gave a sensitivity and specificity of 60% and 87% respectively against DSM positive cases, compared to 68%/73% for the Yale question and 74%/80% for BDI-II cases based on a cut-off of BDI-II ≥ 2.

Responsiveness to change of DISCs was also tested by comparing pre- and post-treatment measures, with results indicating that this measure was sensitive to change, comparing favourably with the NGRS, BDI-II and DSM case categorisation.

Outside this initial validation study, however, DISCs does not appear to have been independently tested or reported in reviews of screening instruments for

depression following stroke (Bennett et al., 2006; E. Townend et al., 2007b; Berg et al., 2009; van Dijk et al., 2016).

3.3.5 Other Adapted Scales. The instruments discussed thus far predominate in the literature regarding the assessment of mood in those with aphasia, but there are some others that deserve mention.

The Face Scale (Lorish & Maisiak, 1986) is a brief, pictorial mood scale using a sequence of 20 drawings of faces along a bipolar continuum between happy and sad. In a study of 174 rheumatoid arthritis patients, the authors found significant correlations with the BDI (r=0.49), Bradburn Positive Affect (r=-0.37) and Negative Affect (r=0.37) scales, and a self-reported pain rating scale (r=0.44); they also reported that its responsiveness to change was comparable to these other measures.

The distress thermometer (DT) (Roth et al., 1998) was developed to provide a fast screening for psychological distress in the context of cancer. It comprises a labelled VAS in a range of 0 to 10, where 0 is "no distress" and 10 is "extreme distress", and is accompanied by the question "How distressed have you been during the past week on a scale of 0–10?". A comprehensive review of the accuracy of the instrument (Mitchell et al., 2010), showed that the DT has good negative predictive value (93% for depression, 80% for anxiety), however positive predictive value is quite poor (33% for depression, 55% for anxiety). Furthermore, in a systematic review of 33 papers in which it was validated (Stewart-Knight et al., 2012), the authors argued that though the DT is widely used in cancer support and palliative care, there is little evidence to support its validity, and doubts about what it is actually measuring.

Wong & Baker (1988) examined a scale of six, simple, line-drawn faces as part of an assessment of pain scales suitable for use in children aged 3 to 18 years. The faces scale was compared to five others, including a simple graphic rating scale and a numbered rating scale. The findings showed that there was little difference in validity or reliability of the scales, but that children preferred the faces scales. A more recent study of the Wong-Baker FACES scale in children admitted to an emergency department (Garra et al., 2010) found that the scores on the scale had a high correlation with scores on a simple, VAS measure for pain. However, while the former offers a nonverbal means of self-rating distress, it is intended to rate physical distress rather than mood *per se*.

Returning to the assessment of depression in the context of stroke, a study of 253 patients one month after stroke used three smileys, – 'happy', 'sad', and 'flat' (neutral) – as a self-report scale, which was then validated against nurses' diagnosis based on interview using DSM IV criteria. Though the 'happy' and 'neutral' faces lacked suitable sensitivity/specificity cut-offs, the 'sad' face fared better, with a sensitivity of 75.9% and specificity of 77.4% (Lee et al., 2008).

3.4 Suitability of Instruments for Use as Screening Measures

3.4.1 Observer-rated measures. In assessing the suitability of instruments for measuring mood in stroke patients, there are a number of important considerations to bear in mind. Obviously, the instrument should demonstrate good psychometric qualities when assessed against other criterion measures, however it is not sufficient to just demonstrate fair or good correlations with other measures of depression. While an instrument of this sort may be useful as part of a battery of outcome measures used

for research purposes, to be of use in clinical practise it should also demonstrate good sensitivity and specificity. It should be successful in correctly identifying those classified as 'depressed' by a criterion measure (either by 'gold standard' diagnosis, or as designated by a cut-off value on a continuum of possible scores), yet also successful at identifying those who are classified as 'not depressed'. These latter psychometric qualities are key to deciding whether an instrument is suitable for use as a screening measure by which patients are assessed and treated in a clinical setting.

In addition to an instrument's psychometric qualities, however, there are also some important practical considerations. The instrument should be accessible and easy to use even for people with a range of disabilities associated with stroke, such as ataxia, dyspraxia, hemianopsia and visual field neglect. It should be reasonably quick to administer and capable of being used by staff with little or no training.

Because of the difficulties that people with communication problems have in expressing themselves, the focus of research into methods to screen for depression or low mood following stroke has been primarily on observer-rated measures. Of the observer-based instruments covered here only one – the SADQ (with its brief H-10, and its more comprehensive H-21 forms) – stands out as having the characteristics and evidence base attesting to its usefulness as a screening measure in a typical hospital or rehabilitation environment. The PDRS, which is intended for use solely by a professional examiner, is unsuitable as a screening measure, and though the ADRS appears to be at least equal to SADQ in terms of its psychometric qualities, its usefulness is severely limited because it requires a team of rehabilitation specialists to administer it. The SADQ, however, can be administered by just a nurse, so is much more useful in a real-world, stroke aftercare setting.

Differences in factor structure between the ADRS and the SADQ also suggest caution in the way that the ADRS should be used. Both the SADQ and the ADRS share a common primary factor of negative mood, and an axis representing sleep-related symptoms, however, while the ADRS lists two further factors relating to anxiety, and factors for retardation and somatic symptoms, the SADQ identified social interaction, loss of interest and apathy (Sutcliffe & Lincoln, 1998). The face validity of the SADQ factor structure seems better as it clearly includes the most important aspects on non-somatic depression symptomatology, whereas the ADRS appears to include factors representing physical symptomatology that are confounding factors in measuring depression in stroke patients, and which the SADQ expressly sought to exclude (Benaim et al., 2004).

A recent systematic review of instruments to assess PSD in patients with aphasia, (van Dijk et al., 2016) was roundly critical of the quality of validation studies for these adapted instruments in general, but concluded that "the SADQ-10, the SADQ-H10 and the SODS show acceptable feasibility" for clinical practise (van Dijk et al., 2016, p. 14). However, in the paper from which the authors draw their conclusions about the SODS, the positive predictive value for SODS was quite poor. Even with a sensitivity of 86% at an optimal cut-off against the HADS-D, the specificity equates to a false positive rate of almost 40%, though this is not surprising given the rather narrow range of SODS scores (an integer on a scale of 0 to 6). The SADQ – with a cut-off at 100% specificity, 78% sensitivity – seemed to fare much better. This is consistent with the findings of Watkins et al. (2001), in which the SODS performed poorly as a screening instrument. Only the original validation study

by the authors of SODS yielded reported sensitivity and specificity figures adequate for a screening measure (Hammond et al., 2000).

E. Townend et al. (2007b) only included adaptations of conventional measures of depression assessment in their review, and therefore offered no assessment of the clinical utility of either the SADQ or the SODS, while Salter et al.'s (2007) review on the assessment of poststroke depression noted a shortcoming common to many studies, which is that the available evidence came from studies which did not adopt the gold standard for depression or include aphasic stroke patients.

In the assessment of mood, observer-rated measures are more suitable as an adjunct to self-reported measures than a replacement. It is only the absence of a satisfactory self-report measure that necessitated recourse to relying entirely on the purely behavioural component of depression. Depression comprises both phenomenology and behavioural components, and any assessment that omits a person's self-reported state can never be fully adequate. It is, after all, the experience of depression that is central, and that can only truly be accessed by communication of a person's affective state.

In self-reported measures, disentangling somatic symptoms of depression from those that may be due to physical illness is of paramount importance in identifying the presence and level of depression in a person with stroke, an issue which was underlined in Lincoln et al.'s (2003) examination of the utility of standard, self-report depression rating scales in the stroke population. In a hospital based sample of 143 stroke patients, the performance of self-report scales compared to 'gold standard' diagnosis using psychiatric interview and DSM-III-R or ICD-10 criteria was examined. The BDI, WDS, and the GHQ were tested, and sensitivity and specificity

values were computed for a variety of cut-offs. The results revealed that the psychometric instruments performed quite poorly in this population, demonstrating the problems with their suitability for people who have had a stroke. While such measures are well validated in healthy adults, symptoms due to physical illness confound these measures by inflating or confusing responses relating to somatic symptoms. In this study, this was evidenced by the superior performance of the GHQ – which focussed on psychological distress – above the BDI and WDS, which incorporated questions relating to physical symptoms. E. Townend et al. (2007b), in their review of adapted depression measures for use in those with aphasia, likewise concluded that conventional language-based methods are only suitable for individuals with mild aphasia.

It is these problems of confounding physical symptomatology that prompted the creation of adapted measures such as the HADS, and which emphasise the importance of focussing on the phenomenology of depression rather than physical symptoms when assessing people with physical illnesses. This underlines the importance of mood as a key defining characteristic when dealing with stroke patients as well as people with other medical conditions, and the focus on 'low mood' and 'emotional distress' in some studies (Bennett et al., 2006; Thomas & Lincoln, 2008; Cobley et al., 2012). It also suggests an inherent limitation to diagnosing depression *per se* in this population, in that physical symptoms considered to be defining characteristics of depression cannot be properly assessed and must therefore be removed from the assessment of depression in stroke survivors as a necessary adaptation.

Though observer-rated measures such as the SADQ are significant, positive developments in assessing mood in people with aphasia, any assessment that does not included self-reported mood is necessarily limited in the conclusions that can be drawn. This is not to say that self-report measures are not without limitation as accurate conveyance of a person's affective state, but it is only through language of one form or another that one can fully open a window onto a person's experience and have a proper understanding of how they are feeling. Observer-rated and self-report measures are therefore complementary facets of the assessment of depression and low mood, and both should ideally be examined together to give a broader picture.

Using both self and observer-rated measures is especially important because of the poor correlations often seen between the two. Williams et al. (2006), for example, found that carer's proxy ratings of patient mood differed significantly from patients' own ratings, and were affected by their own perception of caregiver burden. Likewise, in Berg et al.'s (2009) study of depression assessment following stroke, notable differences between caregiver and patient self-rating were also found: Caregivers using the BDI rated depressive symptoms in patients as consistently higher than the patients themselves. They also found significant correlations between caregivers' ratings of patients and the caregivers' own, self-rated BDI scores (0.60 to 0.61) – figures that were even higher than correlation between caregiver ratings of patients, and patient self-ratings (0.37 to 0.43). Robinson (1981) also noted extremely high correlations (r=0.83) between observer-rated measures and carers' own self-ratings using the ZDS, while Sayer et al. (1993) noted only modest correlations between HDRS (observer-rated) measures and self-reported BDI ratings in a group of 114 depressed inpatients.

Though this mainly applies to proxy ratings using conventional depression rating scales such as the HDRS, BDI, or ZDS, even adapted scales such as the SADQ, which focus primarily on overt behaviour, have shown signs of similar problems. In Cobley et al.'s (2012) study of 154 stroke patients with aphasia, only a weak correlation was found (r=0.297). Therefore even when ratings are based on what should be purely objective, behavioural observation, interpretation of that behaviour is inevitably subject to personal bias, and even the most exacting observational criteria are not completely immune to the projective or priming elements that may come into play in evaluating another person's inner state.

Assessments of depression based purely on behavioural observation are therefore necessarily incomplete, and given the frequent disparity of these two modes of assessment, it is particularly important that convergence is sought through a combination of these methods. Both observer-rated and self-rated methods cast imperfect yet complementary reflections of the symptoms of depression, and where discrepancies exist, it is important to carefully weigh and consider the strengths and weakness of each approach in its specific context, identifying sources of bias and weighing up the extent to which one source of information should be trusted over the other. Differences between these sources would promote closer examination of the complexities of a case, and help avoid being mislead by measures that were obscured or obfuscated in some way by other factors.

3.4.2 Self-report measures. Any comprehensive assessment of depression should therefore include a self-reported measure, however limited. Of the instruments covered, the VAMS has been the most widely used, though the evidence base for its utility in stroke survivors is quite slim. Early studies revealed some encouraging

though mixed results regarding construct and criterion validity and test retest reliability (Stern et al., 1997; Arruda et al., 1997; Arruda et al., 1999; Nyenhuis et al., 1997; Temple et al., 2004), and the 'sad' item seemed to correlate well with the HDRS (Arruda et al., 1997). However none of these latter studies used the 'no-word' version of the VAMS and therefore did not offer any evidence of their utility in people too language-impaired to understand the words accompanying the actual scales. Neither did these studies – typically employing a Multitrait-Multimethod design – offer figures for sensitivity and specificity that would enable its use as a diagnostic or screening measure to be properly evaluated.

E. Townend et al. (2007b) concluded that the VAMS is inadequate as a means of diagnosing depression in people with aphasia, noting that although it is often described as a "validated diagnostic instrument", the evidence base for this is lacking. The validation study of VAMS in stroke patients (Arruda et al., 1999), though yielding some good correlations, was flawed due to the use of the standard version of VAMS (with mood words included) in a sample who had intact single-word recognition (Arruda et al., 1999, p. 677). This enabled participants to circumvent the use of the cartoon face altogether as a means of identifying the mood represented by the card. This means that the findings cannot be extended to those with more severe communication problems. It would have been better to use the no-word version of the VAMS either in place of or in addition to the version with words included.

Berg et al.'s (2009) study of the assessment of depression after stroke likewise judged the VAMS unsuitable for people with aphasia and cognitive impairments, citing poor sensitivity and/or specificity. This is consistent with the findings of Bennett et al. (2006), who concluded that the VAMS was more useful as an indication

of severity of low mood than as a screening measure. The problems encountered by Benaim et al. (2010) where both aphasic and non-aphasic patients frequently failed to understand the directions to complete the VAMS further underlines shortcomings of using this instrument within people with cognitive or language impairments.

The evidence base for VASES as a possible screening measure is limited. Bennett et al. (2006) report that though the VASES (with and without the 'depression' item) correlated moderately well with the HADS-A and HAD-D, sensitivity and specificity were poor, while Cobley et al. (2012) found no significant correlation between the VASES 'depression' item and the SADQ despite a highly significant, though weak correlation with the VAMS 'sad' item. The DISCs is also lacking in this respect; it has not been independently assessed or validated outside the original validation study (Turner-Stokes et al., 2005), so cannot be recommended due to its poor evidence base. It is possible that this reflects the more widespread adoption of the VAMS (or its 'sad' item) and the VASES as substitute measures in studies where participants with communication problems need to report their mood.

In a recent review, Van Dijk et al. (2016), critical of the methodology of almost all of the research into the psychometric qualities of VAMS and the VASES, concluded that the evidence base was insufficient to determine the utility of VAMS or the VASES as a measure of depression. This lack of proven, adapted screening measures for people with aphasia is underlined in *Psychological Management of Stroke,* (Lincoln et al., 2012b), where the authors conclude that though screening measures for low mood after stroke are satisfactory, "measures suitable for those with communication problems are less robust" (Lincoln et al., 2012b, p. 328).

This may, in part, be due to the inherent difficulties in validating such instruments in the target population. It is impossible to apply the gold standard for diagnosis of depression to people who cannot communicate verbally, and the value of including people with limited aphasia as a compromise position is questionable. The more profound a person's aphasia is, the more difficult it will be for them to complete a criterion measure, leaving us with a methodological paradox: It is logically impossible for us to distinguish between poor correlations due to poorer performance on language-based criterion measure, and those due to shortcomings of the instrument under investigation. It is difficult to see how this impasse can be resolved.

3.4.3 The need for a nonverbal self-report mood scale. There is clearly a significant lack of provision for well validated instruments that rely on nonverbal means to assess an individual's risk of depression following stroke. Depression after stroke is a significant problem, and it is important to monitor mood following stroke as per clinical guidelines (Royal College of Physicians, 2012). However around 20–38% of stroke patients (E. Townend et al., 2007b) have a significant degree of aphasia, and so cannot be easily assessed using conventional self-report methods which rely upon language. However the evidence suggests that it is people with communication problems after stroke that are most at risk of depression.

The limitations of existing nonverbal mood measures not only impacts on the ability of doctors, carers and rehabilitation specialists to screen for depression following stroke in order to effect appropriate treatments, but it impedes important research into the effectiveness of interventions by hampering the ability of scientists to track changes in mood over time in a substantial proportion of stroke survivors. The lack of suitable instruments has left researchers who wish to accommodate

patients with severe communication problems with few choices. Typically the VAMS 'sad' item is used to give an approximate indicator of depression phenomenology, or VASES is used as an indirect measure, but both are limited

Research into depression after stroke and assessment of interventions into PSD have therefore been significant hampered by the inability to adequately quantify mood in people with aphasia. In fact research of any kind in which the psychological wellbeing of stroke survivors is examined is notably impacted by problems in assessing self-reported mood in this group. Nearly two thirds of studies examining PSD excluded patients whose aphasia was too severe for them to be amenable to standard measures (E. Townend et al., 2007b), while Hackett & Anderson (2005) reported that only 3 out 20 studies reviewed in which predictors of depression were examined after stroke included aphasia as a potential risk factor. This underlines an overwhelming tendency for patients with communication impairments to simply be omitted from such studies.

There are over 1.2 million people living with stroke in the UK alone (The Stroke Association, 2015). With aphasia affecting around 20-38% of stroke patients, this means that some 300,000 to 450,000 people in the UK will have some degree of communication impairment as a result of their brain injury. There is therefore a sizeable population whose care needs may necessarily be incomplete because of the lack of accessible mood assessment methods that can alert medical and care support services to depression or other mood problems following stroke. As things stand, many people with communication problems cannot communicate their feelings because no suitable and accessible method to do so exists.

The existing evidence suggests that while progress has been made in developing adapted, observer-rated measures, this has yet to be matched by similar progress in adapted self-report measures, specifically, measures which are completely language independent. With such a large population of people with communication problems as a result of stroke (as well as other types of acquired brain injury) there is a pressing need to develop better ways of allowing mood to be assessed nonverbally and to explore novel strategies to bypass verbal communication.

A good, fully nonverbal mood measure would be useful not only for stroke patients, but for people with cognitive impairment due to a disability, neurological disease or acquired brain injury. It would facilitate a more informed triangulation of a person's mood by allowing concurrent observer and self-reported measures to complement one another, and alert researchers to any disagreements between the two. Such a measure would also be a great help in regions with very diverse languages or poor rates of literacy.

Instruments like VAMS and VASES have offered a starting point in this respect, but they are limited as mood assessment tools and technologically outmoded. In order to advance mood measurement outside language, it is time to explore more modern and innovative ways of communicating affect states.

Earlier discussion has established that the use of a VAS has proven to be a valid and reliable way of quantifying self reported mood in the general population, and that using pictures in the place of words is a logical way to overcome the obstacle of language in people with communication problems. However, instruments like the VAMS have fundamental limitations which must be addressed in envisaging a new mood scale design. By examining these limitation, three key improvements were

identified that allowed mood scales to be reconceptualised in the light of recent theoretical and technological developments. These were: (1) the use of photographs of faces; 2) explicit interpolation of images corresponding to VAS values; and (3) the use of an underlying structural model unifying the scales.

3.5 Addressing the Limitations of VAMS

Though the VAMS has proved to be limited, the essential principle behind them is sound: In the absence of a person's ability to use language following a stroke, we must fall back on a method of communicating mood that relies on other modes of communication that remain largely intact. Vision and comprehension of imagery is usually unimpaired, or – as in the case of visual neglect – only partially impaired, and therefore it makes sense to use affect-bearing imagery in place of words and phrases normally used to denote emotions. VAMS is a rudimentary attempt to do this, but it has some key, identifiable shortcomings that need to be addressed in the design of an improved set of scales.

3.5.1 Poor realism of faces. Firstly, there is the issue of the crude and simplistic nature of the images used in the VAMS. There is good reason to believe that the use of such simple and stylised graphics denoting mood states is inherently problematic. Some of the graphics denoting mood states on the VAMS are far from clear, as is attested to by differences noted between results for the VAMS which included the word for the mood state and those which did not. Though correlations between the word and no-word versions of the VAMS were generally good, the 'afraid' and 'confused' items performed particularly poorly in comparison to the other items (Stern et al., 1997).

A large body of evidence would indicate that facial expressions are the most effective way to convey detailed non-verbal information about a person's mood state, with studies demonstrating consistently high recognition rates of posed facial expressions (Ekman & Friesen, 1971; Ekman, 1993; Izard, 1994). Ekman (1994), for example, reported recognition rates of posed facial expression of between 78% and 94.7% in western cultures, and cross-cultural recognition rates of between 59% and 87.8%. Even when taking into account the weaker recognition rates of some cross-cultural studies, it is generally accepted that Ekman's 'basic' emotions are cross-culturely universal, and that facial expressions offer a universal language for key affective states (Ekman et al., 1987; Ekman, 1992, 1993, 1994; Russell, 1995; Elfenbein & Ambady, 2002; Ekman, 2005).

The use of photographs of actual facial expressions, therefore, may be one way to improve accuracy. Actual facial expressions are far more complex and nuanced, and carry a wealth of detail which such simple graphics cannot do justice to. Given the cognitive impairment that frequently accompanies stroke, it would seem prudent to adopt this well-understood "common currency" of visually represented affective states without too much modification, using validated images of mood states posed by actors to form a basis for the scales. This concern over the poor realism of VAMS images was echoed in the conclusions of E. Townend et al. (2007b), in which the authors remark that "use of realistic looking pictures [...] may usefully support communication about mood with people with aphasia" (p.3081).

Furthermore, there are important neurological aspects of aphasia that have implications for whether or not a person with aphasia can necessarily read the VAMS faces. It is possible that disruption caused to a person's ability to recognise symbols in

written language may also impact on their ability to recognise the kind of simplified graphic 'smiley' upon which the VAMS depends. Neuropsychological evidence, however, suggests that that recognition of emotion in facial expressions is primarily mediated by right hemisphere processes, with unilateral brain injuries resulting in corresponding impairment (Borod et al., 1986; Moreno et al., 1990; Adolphs et al., 1996; Borod et al., 1998; Kucharska-Pietura et al., 2003; Philippi et al., 2009). Since most aphasic patients have left hemisphere lesions, recognition of facial expression is unlikely to be impaired in this group. Conversely, it is more likely that stroke patients with significant alexia (impaired comprehension of written language) may also be unable to understand the meaning of the VAMS faces, as brain structures implicated in decoding and comprehending written language (pathways joining the lateral posterior temporal lobe to the visual cortex) may serve both of these functions.

The fact that stroke survivors are generally older, though, may have implications for the use of faces for communicating affective states. It is generally accepted that older adults have more difficulty recognising key emotions (Ruffman et al., 2008; Mill et al., 2009), however there is also the question of whether this impairment is affected by the characteristics of the face expressing the emotion. If the population in which a face based instrument is to be used consists mainly of older people, for example, might it benefit the recognition of facial expression to have older actors posing expressions for the photos used by the instrument? Evidence addressing recognition of facial expression across age groups, however, indicates that both young and older groups are better at recognising expressions in young compared to older faces, suggesting that younger faces may be a better choice (Ebner & Johnson, 2009).

There are also implications for the presence of depression in the interpretation of facial expressions. A major review in this area demonstrated differences in the way that they are interpreted by people with depression (Bourke et al., 2010). A general response bias was observer, where neutral or ambiguous expressions are interpreted as more sad, and a selective attention towards sadder expressions; however there was also a general, reduced accuracy in the ability to recognise both sad and happy expressions. A study of perception of emotional facial expressions in depressed and non-depressed stroke patients yielded results consistent with this latter finding, concluding that depressed stroke patients were less sensitive to sad, angry and happy expressions (Montagne et al., 2007). It is not clear how these effects, combined, would impact on the use of scales based on images of facial expressions.

3.5.2 Lack of underlying theory. Perhaps the greatest shortcoming of the VAMS is that the separate scales and the moods they represent are not connected in any meaningful way. The VAMS are not separate subscales of a single scale, but a collection of discrete and disparate scales that measure a small number of qualitatively distinct mood states, and which are not joined by any underlying theory. It is therefore difficult to arrive at a meaningful VAMS 'total score'. In studies using the VAMS in its entirety, the scales have therefore been used to give a broad measure of valence, by either dropping the positively valenced items 'happy' and 'energetic' (Bennett et al., 2006) or reversing them (Kontou et al., 2012), thus allowing a composite score or measure of internal consistency (Cronbach's α) to be used.

However, because the mood state most relevant to the measurement of depression comprises just a single scale – the 'sad item', most studies using the VAMS for this purpose have simply used this one scale and dismissed all the others

leaving a single-item scale not unlike the original VADS. This use of a single scale yielding only a single measure greatly weakens its utility as a psychometric instrument. On more conventional measures a scale may be comprised of many items designed to tap the construct in question, and it is this repeated measures which allows a total score to be accurate enough to be useful. An underlying theory of affect in which these different moods could be related empirically to one another through more basic underlying variables could allow these separate scales to act instead as separate items which can meaningfully be combined into a single score.

A major key to developing better mood measures, therefore, is to understand what mood is and the factors which comprise it. This will be covered in detail in section 4.4.

3.5.3 Limitations of the VAS format. Finally, there have been concerns about the use of VAS measures with cognitively impaired patients. There is evidence to suggest that many people with cognitive impairments that accompany stroke are simply not capable of using a conventional VAS (C. I. Price et al., 1999). Scaling notional concepts into numerical quantities or proportions of physical distance is something that most of us take for granted, but when examined in detail it is revealed to be a complex and cognitively intensive process. It is reasonable to surmise that cognitive impairments resulting from stroke may result in a diminished ability to use these types of scales, by way of impairment of a person's ability to translate notional measures into scaled physical distances. Indeed it was this very concern about the impairment of a person's ability to scale and quantify abstract notion that gave rise to the DISCs (Turner-Stokes et al., 2005).

These processes by which we scale abstract concepts into quantities which we can report as numeric scores, or positions along a Likert scale or VAS are necessarily highly individual, varied and idiosyncratic. This *cognitive interpolation*, is likely subject to substantial variation between even cognitively high-functioning individuals, so it is reasonable to suppose that this may be a substantial obstacle to assessing mood in people after a significant brain injury. But there are now other options open to us than these relatively crude psychometric tools designed within the constraints of pen and paper. The growing ubiquity of high-definition, tablet-based technology, with its substantial computing power, has brought with it the opportunity for far more sophisticated and interactive interfaces, and redefine the way that we can visualise, scale and modulate abstract constructs.

What if, instead of relying on a respondent to mentally scale constructs relating to their mood (cognitive interpolation), we instead used a process of *explicit interpolation*, whereby a particular mood or emotion type would be embodied as an image of a corresponding facial expression?

This approach would both address the limitations of a traditional VAS and enable the medium of facial expression to its fullest effect. A software-based design was therefore adopted utilising a new, dynamic form of VAS (or DVAS). In the DVAS a reference image dynamically changes in response to the position of a slider control on a touch-screen interface. Under this design, any position along the length of a VAS would be *explicitly interpolated* and displayed in the form of an image corresponding to a particular scale value. This interface, along with suitable images of facial expressions would form the basis of a new set of mood scales *Dynamic Visual*

Analogue Mood Scales (D-VAMS) designed to run on modern, multi-function devices.

3.6 How Should We Measure Mood?

Having established a basic blueprint for a new kind of scale, we come to the question of what we should be using it to measure. How many moods are there, and which ones are important? Though our mood assessment instrument should be capable of providing an overall picture of the most important aspects of mood, we also need to focus on aspects most relevant to the construct of depression. What is the relationship between mood and depression, and what elements of mood are most central to depression? Though the concept of mood or emotion is something that we are tacitly familiar with in the course of everyday life, analysing it raises many questions regarding how it should best be conceptualised. In order to assess something we first need a clear idea of what it is that we are trying to measure, what it consists of, and how best to model it. A mood assessment instrument needs to be underpinned by a solid theory.

These questions are central to the academic study of the area broadly known as *affect*. Addressing these questions requires recourse to a substantial body of research, and it is to this literature that discussion will now turn.

4. What is Mood? Structural Theory and Mood Measurement

4.1 Structural Theories of Mood

4.11 Introduction: background and early research. The study of emotion was, in the early days of psychology, almost inseparable from the study of facial expression, and early researchers into the structure of what we now call 'affect' tried to understand it by examining how people describe or classify photographs of facial expressions. The main objective of these early studies was to establish a taxonomy of 'basic' emotions, which could act as a structure for further study, however this was to prove problematic. Free response naming of photographs of facial expression in studies typically yielded over a hundred adjectives (Feleky, 1914; Ruckmick, 1921; Frois-Wittman, 1930) and discerning a pattern from such data was difficult, but eventually persistence paid off as broad clusters were identified.

Frois-Wittman (1930) asked participants to label 46 photographs of his own posed facial expressions and examined the adjectives returned by them, organising them into six basic groups into which these labels seemed to cluster. Of particular interest was the emergence of a pattern of "characteristic confusions" (p.114), – a blurring between these categories which enabled them to be assembled along something of a continuum. The American psychologist Woodworth (1938), re-examining the data of these earlier studies, also arrived at six basic groups. He found that by ordering these groups alongside those with which there was most overlap, it was possible to assemble them into a rudimentary dimension. Woodworth named

these categories, *Love-Happiness-Mirth*, S*urprise*, *Fear-Suffering*, *Anger-Determination*, *Disgust* and *Contempt*. When using these as basic categories of emotions, he calculated a correlation between pose and judgment of 0.92.

Schlosberg (1941) examined Woodworth's scale further in a sort task using the Frois-Wittman photos (Frois-Wittman, 1930; Hulin & Katz, 1935). As before, the categorisation of photos seemed to follow a single dimension underlying Woodworth's categories, which he likened to a spectrum along which separate colours are distributed. Of particular interest, though, was that the scale did not appear to be purely linear, but seemed to wrap around on itself in a circular fashion. Photographs from the final 'step 6' of the scale, *Contempt*, were also frequently categorised under the first step, *Love-Happiness-Mirth*. Charting his photographs along this continuum and noting the changes in facial expression along the length of scale, he began to hypothesise what types of dimensions might be inferred from this.

One clear dimension 'Pleasantness/Unpleasantness' was easily discernible, yet his continuum was not just a simple blend from pleasantness to unpleasantness through a series of intermediate stages, so he reasoned that at least one other dimension must exist that would explain this circularity. He therefore hypothesised the existence of a second, less pronounced dimension, which he provisionally labelled 'Attention-Rejection' (AR). On examining the data from Frois-Wittman's (1930) earlier study, and that of another researcher, Kanner (1931), he found that their results also supported a similar arrangement.

This rudimentary circumplex model, however, had some problems. Schlosberg (1952) reported difficulty explaining this AR dimension to study participants, and doubts about its validity led him to explore alternative dimensions. Heavily influenced

by the James-Lange and Cannon-Bard theories of emotion, he recognised the vital role that physiological processes played in emotional behaviour, and proposed a new dimension denoting biological arousal, which he called 'Sleep-Tension' (ST) (Schlosberg, 1954). All three of these dimensions were examined in scaling experiments using a new set of facial expression photographs (Engen et al., 1957, 1958; Triandis & Lambert, 1958), with the results favouring the PU and ST dimensions over AR in terms of reliability.

However only with the arrival of multidimensional scaling (MDS) methods (Torgerson, 1958) did it become possible to examine factors directly. In a crucial test of Schlosberg's model, Abelson and Sermat (1962) applied MDS to similarity judgements of paired facial expressions. The results yielded two interpretable dimensions. The first, accounting for 44.8% of the variance, correlated with Schlosberg's PU dimension (0.95), while the second, accounting for 28.2% of the variance correlated with his AR and ST dimensions (0.88 and 0.92, respectively).

This finding of a two-factor solution, the poorer reliability of AR, its lower stability across scaling methods (Engen & Levy, 1956) and the better performance of the ST dimension in predicting dissimilarity data, led the authors to conclude that ST was a better contender for this second factor. Once the AR factor was removed, there remained a two-factor, PU-ST solution accounting for 73.2% of the observed variance.

4.1.2 A two-factor solution: growing convergence in studies of affect. At about this time, studies of affect using judgements of facial expressions gave way to studies in which adjective checklists (ACLs) were used to examine people's self-reported mood. Extensive studies by Osgood et al. (1957) into the structure of

meaning inherent in language (the so-called ‘semantic differential’) yielded three key factors ‘evaluation’, ‘activity’ and ‘potency’, which, they concluded, were the major components of the meaning of natural language. The parallels between the first two of these factors, and Schlosberg’s PU and ST dimensions of affect were noted by Mehrabian and Russell (1974), who later proposed a parallel, three-factor theory of affect composed of pleasure-displeasure, degree of arousal and dominance-submissiveness (Russell & Mehrabian, 1977). Bush (1973), in an MDS study of 264 affect terms, also arrived at a three factor solution. Again, two primary factors, pleasantness and level of activation emerged, plus a third that was harder to interpret, but which seemed related to aggressiveness or potency.

This convergence of evidence left little doubt that pleasure-displeasure and degree of activation comprise the main components of what we call mood. On the matter of a third factor, however, consensus has been less clear-cut, with researchers offering varying conceptualisations. While Russell and Mehrabian’s (1977) dominance-submissiveness factor does parallel Osgood’s potency factor, it has also been characterised elsewhere as locus of control (Nowlis & Nowlis, 1956) or ‘depth of experience’ (Averill, 1975) both of which appear to be subtle, but qualitatively different variations on this theme. This uncertainty over the validity of this third factor within the domain of affect was underlined in Russell’s (1979) examination of 11 ACL scales and their factor structure, with the results showing little evidence of the dominance-submissiveness factor in this larger pool of affect words. Noting that this third factor accounted for only a relatively small proportion of the variance (Russell & Mehrabian, 1977), that it was a component of only some, and not all affect words (Russell, 1978) and that dimensions beyond the first and second “have consistently

been found to differ from pleasure and arousal in several important ways" (p.354), he proposed a simpler, two-factor scheme that echoed Schlosberg's (1952) earlier model.

This two-factor model was formalised in Russell's classic paper: "A circumplex model of affect" (Russell, 1980). In an elegant series of four experiments examining 28 commonly used emotion words, he demonstrated that they could be represented by two dimensions: 'valence' (pleasantness-unpleasantness) and 'activation'. Each mood word, he proposed, could be represented by some combination of these two factors, and when his data were plotted into the space delineated by two corresponding axes, he found that they fell into a roughly circular arrangement (see Figure 4.1).

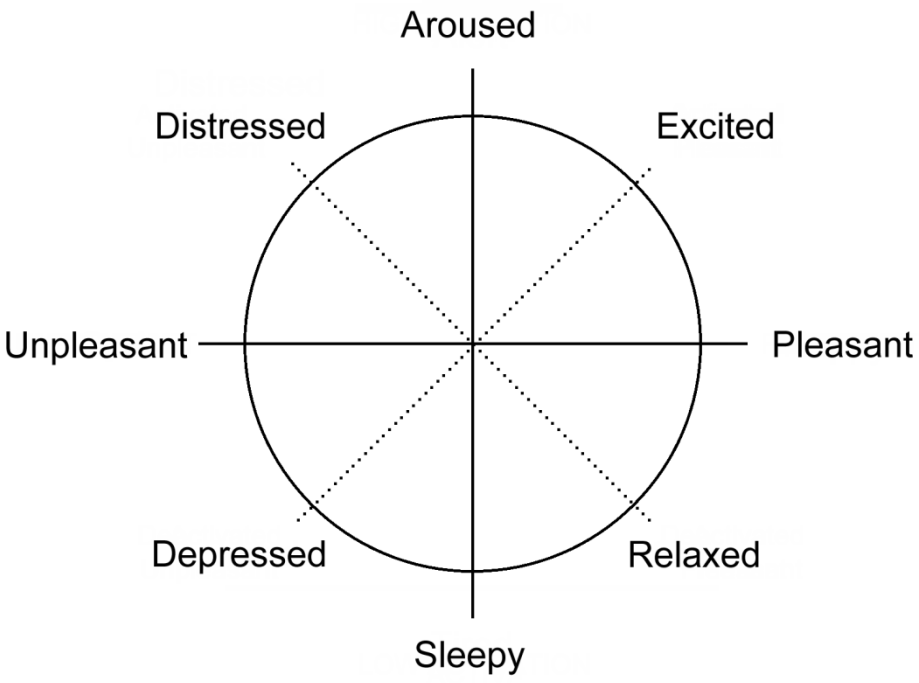


Figure 4.1. *The Affect Circumplex – adapted from Russell (1980)*

Within this scheme, Russell clarified his conception of affect by marking off this two-factor domain as being of special significance. He viewed this as representing the core element of mood before it enters into cognition, and akin to our concept of mood rather than emotion *per se*. He would later term this 'core affect', which he defined as "the most elementary consciously accessible affective feelings" (Russell & Barrett, 1999, p. 806). Factors beyond this (such as 'dominance-submissiveness') he viewed as reflecting the distinction between mood and emotion, in what he termed 'prototypical emotional episodes'. Whereas core affect is viewed as detached from a specific event or stimulus, prototypical emotional episodes are viewed as relating to specific stimuli and behavioural responses, and the antecedents and consequences of affect rather than affect *per se*, such as with the 'flight or fight' response of behavioural accounts of emotion.

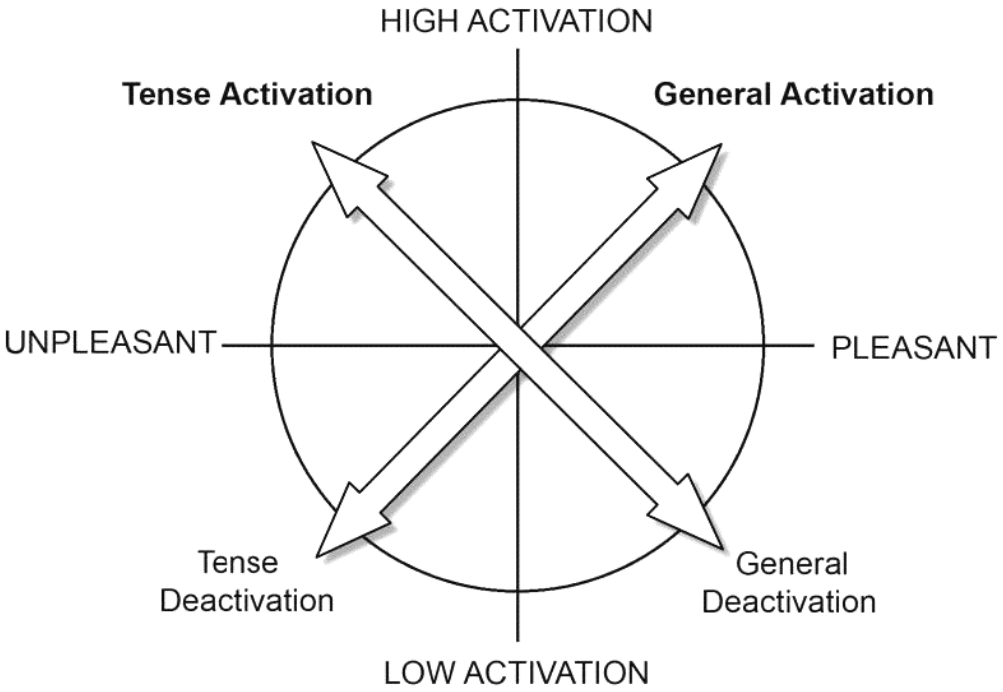


Figure 4.2. *Thayer's (1967) four factors, subsequently reduced to two: Energetic and Tense Arousal*

Crucially, Russell also consolidated earlier research in which emotion was characterised primarily in terms of physiological activation. Thayer (1967) in a study of data from his activation-deactivation adjective check-list (AD-ACL) proposed four unipolar activation factors that could account for self-reported mood. Russell (1979) demonstrated that these four apparent unipolar factors could be reduced to two bipolar ones which could be represented as dimensions crossing this same affective space, only at different angles. Observing how a 45° rotation of the valence-activation axes could yield two other bipolar dimensions 'excitement-depression' and 'distress-relaxation' (Russell, 1979, p. 355), he suggested how Thayer's factors could be incorporated into the same scheme. Thayer (1989), acknowledging this interpretation, subsequently renamed his factors 'energetic arousal' and 'tense arousal' (see Figure 4.2).

This "45° rotation hypothesis" was to prove particularly prescient, as other researchers were also arriving at two-factor, circular models along the same lines. Most prominent amongst these was the model proposed by Watson et al. (1985), which comprised two dimensions they named *Positive Affect* (PA) and *Negative Affect* (NA), and which subsequently formed the basis of their mood measurement instrument the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988). These two apparent, unipolar dimensions were treated as orthogonal dimensions much like those of Russell's valence and activation, but the authors themselves acknowledged that they could also be represented within an affective space much like the one denoted by Russell's circumplex, illustrating their model with a circumplex of their own (p.221). Once the axes of their circumplex were rotated and the x-axis reversed, their PA and NA constructs could be seen to be compatible with Russell's

own circumplex, with the concept of activation or arousal being replaced with one of 'engagement'.

As with Thayer's proposed dimensions, Watson and Tellegen's (1985) PA and NA dimensions could also be rotated approximately 45° to fit into essentially the same structure. PA and NA, however, comprised both a valence and activation component, and their continued use of the names Positive and Negative *Affect* (Watson et al., 1988; Watson & Clark, 1994, 1997) caused some confusion in subsequent debate. Eventually the authors changed the names of their PANAS constructs from Positive/Negative *Affect* to Positive/Negative *Activation* to reflect the role of the activation dimension (Feldman Barrett & Russell, 1998; Watson & Tellegen, 1999; Watson et al., 1999).

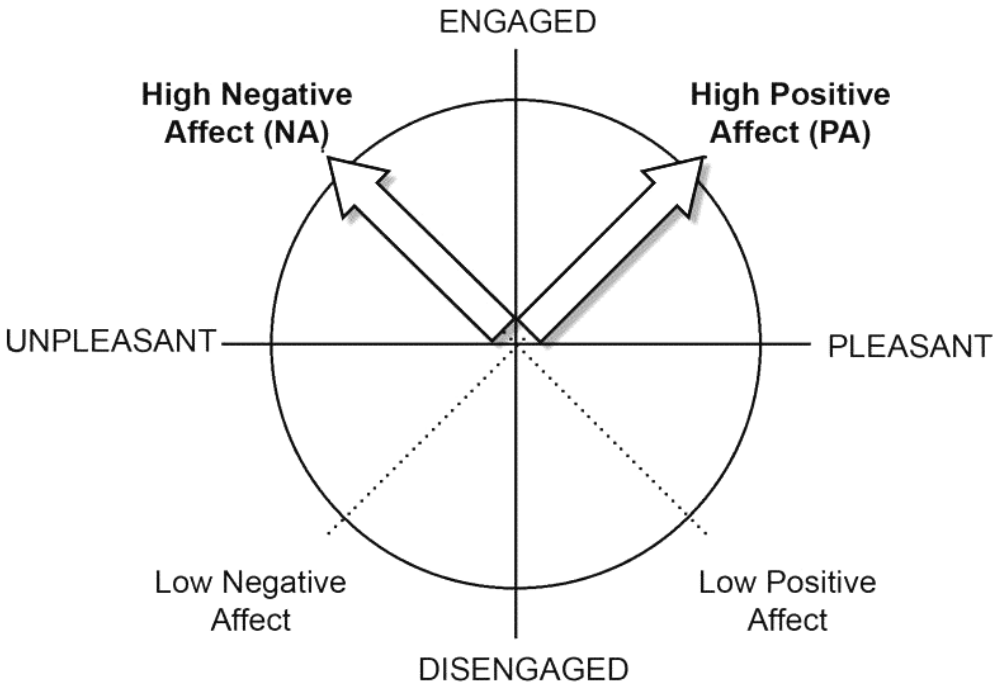


Figure 4.3. *Positive Affect (PA) and Negative Affect (NA) adapted from Watson et al. (1985)*

Larsen & Diener (1992), in a review of the growing body of research in structural theories of affect, proposed their own nomenclature for a structure within which all of these models could be encompassed. In this model, two dimensions 'Pleasant-Unpleasant' and 'Activation' divided the factor space into four quadrants: *Activated Unpleasant*, *Activated Pleasant*, *Deactivated Unpleasant* and *Activated Pleasant* (see Figure 4.4). This basic structure has remained essentially unchanged ever since, with much subsequent research devoted to locating these alternative constructs within the same space (Feldman Barrett & Russell, 1998; Yik et al., 2011).

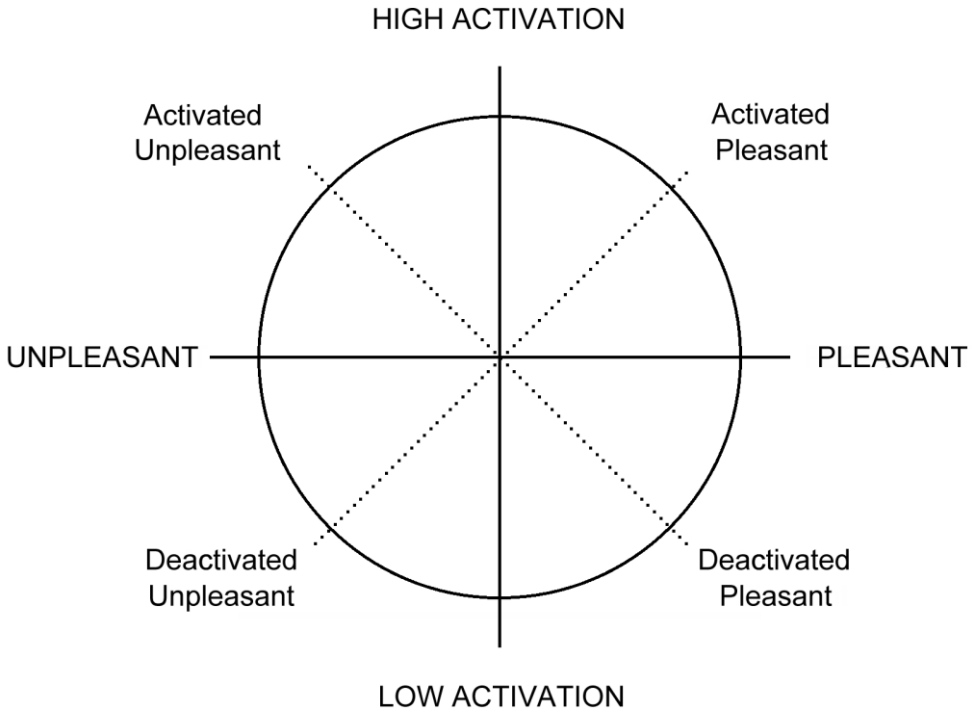


Figure 4.4. *The Self-Report Affect Circumplex (Larsen & Diener, 1992)*

From a theoretical standpoint, there are good reasons for supporting valence and activation as the latent constructs comprising affect. The notion of physiological

arousal and its various markers (i.e: heart rate, blood pressure, etc.) has been fundamental to psychology and the conceptualisation of the emotion since the earliest theories of James Lange and Cannon Bard. The concept of 'arousal', with its roots in the activity of the autonomic nervous system and structures of the limbic system is now widely accepted as a fundamental dimension of physiology which is central to our understanding of human psychology and emotional behaviour, as is attested to by its influence on Thayer's efforts to conceptualise the structure of affect in these terms (Thayer, 1967, 1978, 1989). Likewise, 'hedonic tone' or valence reflects the basic reinforcement system underpinning human behaviour, with its focus on avoiding painful and seeking pleasurable experiences. This concept is at the root of our ideas of good and bad, or pleasant versus unpleasant. It would make sense that these distinct but powerful dimensions of physiological and psychological function would be embodied in the terrain of our affective experience, and that our experience of affect fundamentally reflects the neurological subsystems that comprise these processes; indeed, there is a growing body of neurophysiological evidence suggesting that this is the case (Posner et al., 2005; Posner et al., 2009; Colibazzi et al., 2010; Baucom et al., 2012; Sieger et al., 2015). As observed earlier, the critical convergence of affect theory with the structure inherent within the meaning of language (Osgood et al., 1957; Osgood, 1966) is also highly suggestive of dimensions that are fundamental to our felt, inner state.

4.1.3 PA, NA and the bipolarity debate. The Circumplex Model of Affect and its variants where to prove highly influential in the development of self-report, mood assessment measures. Mehrabian & Russell's (1974) scales of, Pleasure Arousal, and Dominance, Watson, Clark and Tellegen's Positive and Negative Affect

Schedule (PANAS) (1988), and Russell's (1989) Affect Grid were all underpinned by the essentially the same structural model, as was a subsequent, Swedish language instrument (Vastfjall et al., 2002). However there remained an ongoing debate about the true orientation of this structure with respect to the underlying constructs.

Typically, Principal Components Analysis (PCA) was used to extract factors as part of an exploratory analysis of data from self-reported affect. These analyses frequently incorporated rotational techniques such as Varimax, designed to reveal simple structure, where the variables under examination load on one and only one factor, and very little or not at all on the other extracted factors. While Varimax rotation tended to favour the rotational scheme underpinning the PANAS (Watson & Tellegen, 1985; 1988), unrotated solutions yielded the valence-activation orientation of Russell (1980). Circumplex models, however, do not have simple structure (Guttman, 1954; Acton & Revelle, 2004), and so the use of Varimax on data with circular solutions is not appropriate.

More importantly, the use of only high activation mood words in the original version of the PANAS (1988) and the use of the terms Positive Affect (PA) and Negative Affect (NA) was to create an artefact that caused a certain amount of confusion and conflated the issue of factor structure and rotational scheme with a more fundamental question relating to the bipolarity of affect. As shown in Figure 4.3, PA and NA are made up not just of valence, but of a combination of valence and activation. Were PA and NA to represent only valence, and align themselves accordingly along the x-axis, they would be clearly negatively correlated. Yet in this schema the attenuating effect of the activation component of these vectors yield two

constructs that are apparently uncorrelated. There is instead the illusion of two independent, unipolar factors lying orthogonally to one another.

As mentioned before, the authors duly renamed their factors *Positive Activation* and *Negative Activation*, so as to clarify this, and a revised version of the PANAS, the PANAS-X was produced, which reflected the activation dimension and included low activation as well as high activation mood words (Watson & Clark, 1994). However, this confusion over the independence of PA and NA was to lend credence to a much more controversial conception, – the idea that pleasantness and unpleasantness are not in fact polar opposites, but separate factors acting independently of one another.

Rather than being taken as a self-evident *non sequitar*, the idea that positive and negative mood states may be uncorrelated was instead greeted by many as a tantalising paradox. Indeed, even the authors of the PANAS continued to maintain that positive and negative valence were not complementary ends of a single dimension, but that, as a fundamental psychometric principle, "oppositely valenced affects … tend to be only weakly negatively correlated with one another" (Watson & Clark, 1997, p. 282).

Despite the counterintuitive nature of such a notion, a debate about the bipolarity of affect had been ongoing for many years. Following their proposal of a four-factor structure of affect, for example, Nowlis & Nowlis (1956) examined intercorrelations of self-report data on ACLs and came to the surprising conclusion that there was very little evidence of bipolarity in the factors obtained; distinct unipolar factors emerged from factor analysis rather than single bipolar ones. Similar unexpected results were also found by other early researchers (Borgatta, 1961; Clyde,

1963; Lorr et al., 1967; Thayer, 1967; McNair et al., 1971). Though unipolar scales were used for the ACLs comprising their self-rated measures, it was generally assumed that the underlying constructs were bipolar in nature; however factor analyses of their data revealed two independent unipolar factors where one bipolar one was expected, with the anticipated negative correlations between positive and negative affect being either very weak or entirely absent. Debate was further ignited by the publication of Bradburn's (1969) book "The structure of psychological well being", in which the author detailed studies of scales he designed to examine positive and negative affect; he concluded that positive and negative affect were largely independent of one another, as evidenced by the extremely low negative correlations that he found between the two.

Other researchers, however, were unconvinced by these findings, and focussed instead on detailed examinations of the ways that sources of systematic and random error can mask bipolarity in ratings given on such scales. Bentler (1969) demonstrated the striking effect that acquiescence bias – the tendency to agree with a statement regardless of its content – had on correlations between responses to negative and positive words; correcting for this artefact saw correlations rise markedly, with an unadjusted correlation of 0.03 rising to -0.76 when this factor was compensated for. Russell (1979) also examined the role of artefacts in a study of self-reported mood using 11 sets of adjectives on unipolar scales of varying formats, including Thayer's (1967) scales. In addition to acquiescence bias, Russell identified a number of other factors that appeared to mask bipolarity. These included response formats that did not yield ordinal data, inadequate sampling of affect terms, and proximity error – the tendency to respond similarly to items close together in time or space, all of which

tended to inflate positive correlations amongst items which one would expect to be negatively correlated.

Significant amongst these proposed artefacts, was the time period over which a respondent was asked to report their mood. The longer the time period, the more likely it is that markedly opposite feelings may have occurred, and therefore the lower the negative correlations that could be expected in a person's negative and positive reported mood. The importance of the role of time frame in the apparent independence of negative and positive affect was subsequently vindicated in a study by Diener and Emmons (1984), who examined self-reported mood over one year, 70 days, 30 days, 3 weeks, daily, and present-moment. They found, as Russell predicted, that the relation between positive and negative affect was strongly affected by the time frame, and that the correlation decreased in a linear fashion as the time frame increased.

Correcting for these sources of measurement error, however, appeared to profoundly influence correlations between negative and positive affect. In an influential paper "Measurement error masks bipolarity", Green et al. (1993) used a technique to find correlations between latent constructs. Using Confirmatory Factor Analysis (CFA), the authors estimated random and systematic sources of error, and compensated for them. Doing this strongly favoured bipolarity, with a happy/sad correlation – in one dramatic example – from rising from -.25 to -.84. Researchers who initially found independent factors later found that controlling for measurement error yields single, bipolar factors in place of two unipolar ones; Thayer's (1967) four unipolar factors, for example, became two bipolar ones when adjusting for this (Thayer, 1989, 1986).

As mentioned earlier, however, the authors of PANAS and its successor continued to maintain the independence of PA and NA (Watson & Clark, 1997). Though adjusting for random error and acquiescence bias (Tellegen, Watson & Clark, in press) favoured bipolarity in their measures – yielding a correlation of -0.43 – the authors maintained that these remain "largely independent" (Watson & Clark, 1997, p. 267). It was, however, increasingly evident that the bipolarity debate was based on misconceptions on both sides about what was meant by 'bipolarity' in measures of affect, as well as the misleading names of the PA and NA constructs of the PANAS.

In a paper exploring the matter of independence and bipolarity in some depth, Feldman-Barrett & Russell (1998) clarified a prevailing misunderstanding, distinguishing the concept of bipolarity in its most basic form – that is the *a priori* negative relationship between 'good' and 'bad', or 'pleasant' and 'unpleasant' – with a looser form, in which positive affect terms generally correlate negatively with negative affect terms. While this former position is necessarily correct by definition, the latter, looser form is confounded by the presence of a second, activation dimension in commonly chosen mood words. In this second case, it does not necessarily follow that 'excited' and 'calm' are positively correlated because they are both positively valenced, as they are also semantically distanced from one another by a negative correlation in a second, underlying activation dimension. Likewise, 'bored' and 'calm', though differing in their valence, are judged semantically closer due to presence of a common, low activation component. It is only in this latter sense that 'independence' is evident. In the same paper the authors also performed a series of experiments applying the CFA techniques of Green et al. (1993) to examine the PA and NA constructs of the PANAS. Using these and other methods to remove random

and systematic sources of error, latent correlations were computed that revealed the clear presence of underlying bipolar constructs equivalent to two independent dimensions of valence and activation, with "positive affect the bipolar opposite of negative affect, and deactivation the bipolar opposite of activation" (Feldman Barrett & Russell, 1998, p. 967).

In a further paper, "On the Bipolarity of Positive and Negative Affect" (Russell & Carroll, 1999a), the authors present some important and illuminating examples of the pitfalls of using unipolar scales to measure bipolar constructs. Key amongst them were the assumptions that respondents may bring to bear on a scale presented in a simple unipolar format. In a scale on a continuum whose endpoints are "not happy" and "happy", for example, respondents in a small study had a tendency to treat "not happy" as the opposite of happy, rather than as a neutral point that would exist in the middle of a bipolar, "sad-happy" scale, and labelled points along the scale as if it were bipolar rather than unipolar (with points towards the "not happy" end being described as "glum" or "sad"). This would indicate that people tend to automatically assume bipolarity in responding to scales asking about mood.

Another key problem is that even when it is made completely explicit that a unipolar scale is conceptually defined as one half of a bipolar scale spanning "sad" to "neutral" to "happy", thereby creating "neutral-sad" or "neutral-happy" scales, the response format automatically introduces a "dead zone" in which half the possible responses cannot be properly encompassed by a purely unipolar scale. A "sad" response cannot be charted on a "neutral-happy" scale and *vice versa*. This leads to a mathematical artefact which substantially impacts on the theoretical correlations of items opposite to one another in a circumplex structure. The authors calculated that

whereas with a strictly bipolar format, data from scales had – as would be expected – a maximum theoretical correlation of -1.00, data employing responses from scales of a strictly unipolar format had a maximum theoretical correlation of -0.467 (Russell & Carroll, 1999a, p. 10). Adding the systematic and random error that would occur in real life, it is easy to see how expected, strong negative correlations between opposite mood states could diminish to levels that would imply there is very little, or no correlation at all.

In an ingenious weather analogy, Russell also demonstrates an example whereby temperature is split into two separate unipolar measures, 'hotness' and 'coldness'. In two fictional scenarios, two cities have their measures of 'hotness' and 'coldness' gauged from a hypothetical world mean temperature. In the first scenario, all cities have mean temperatures that are exactly the same, but have a different variance. In one, for example, bitterly cold winters are followed by very hot summers, while in the another city (with the same mean temperature) slightly cool winters are followed by only moderately warm summers. In such a world, mean hot and mean cold would correlate by +1.00, as seasonal changes of cold and hot would be in perfect lockstep.

In the second scenario, however, cities have mean temperatures that vary across the globe, but have similar variance. In this case, measures of hot and cold may be negatively correlated simply because one city happens to have a very cold mean temperature and another has a very hot mean temperature.

By this example, mean hot and mean cold temperatures, measured on purely unipolar scales can thereby have any correlation between -1.00 and +1.00. When this transposed to positive and negative affect, the same thing applies: *The correlation*

between mean positive affect and mean negative affect can be anything at all, depending on the characteristics of the sample under examination. Underlining the analogy between temperature and mood, Russell says, "As demanded by bipolarity, in any one place at any one time, when it is hot, it is not cold, and when it is cold, it is not hot. Indeed, any specific temperature rating precludes any other." (Russell & Carroll, 1999a, p. 22).

In a critical exchange covering bipolarity and a range of other technical issues on the measurement and conceptualisation of affect (Russell & Carroll, 1999a; Watson & Tellegen, 1999; Russell & Carroll, 1999b), it was largely agreed that the strong, paradoxical formulation of bipolarity was a result of misconceptions and misunderstandings in the language used to describe affect. It was at this point that the authors of PANAS – as mentioned earlier – agreed that their terms *Positive Affect* (PA) and *Negative Affect* (NA) should be renamed *Positive Activation* and *Negative Activation*, to clarify their positions in relation to the CMA valence and activation dimensions (see Figure 4.3), and to help dispel the myth of paradoxical self-contradiction, whereby negative and positive affect (as in negative and positive *valence*) are uncorrelated despite being polar opposites of one another.

4.2 Applying the CMA to Nonverbal Mood Measurement

The preceding overview of the history and theory of affect and its measurement has been fairly thorough and detailed. However this detail is necessary to properly illustrate and understand important technical and theoretical issues that are key to informing the design of an instrument to measure mood. To select an appropriate framework for understanding, measuring and quantifying mood, it is essential to

understand and resolve key disputes that have endured, and to elucidate and unravel the misunderstandings and confusions that have emerged over the conceptualisation and technical language of affect and its measurement. From this examination of affect theory, a number of key conclusions can be drawn which serve to form the basis of this approach to devising a nonverbal mood assessment device.

Firstly, the CMA has proved to be a theoretically sound and well supported model that has proven of use in existing mood scales, including the PANAS and its revised version PANAS-X (Watson et al., 1988; Watson & Clark, 1994), Russell and Mehrabian's Affect Grid (Russell, 1989), as well as a Swedish, Circumplex-based mood measurement instrument (Vastfjall et al., 2002). The CMA therefore offers a sound theoretical way of creating, conceptualising and organising scales for the measurement of mood. In contrast, the POMS – upon which the original VAMS appears to be based – is outdated, and based on an unconnected, categorical taxonomy of mood that predates more complex dimensional accounts of mood that were to supersede it. This is attested to by the fact that VAMS scores are usually summed by simply treating each as an indicator of valence, or else by omitting all scales except for the 'sad' item.

Secondly, it is also clear is that affective space is bipolar and should be treated as such (Bentler, 1969; Russell, 1979; Feldman Barrett & Russell, 1998; Russell & Carroll, 1999a, 1999b). Happy really is the opposite of sad, and the two represent (more or less) endpoints of a single valence dimension, and not some peculiar combination of "happy" and "sad" systems that are independent and uncorrelated. The use of unipolar scales in the measurement of mood has caused a lot of unforeseen problems, and both practical and theoretical considerations support the use of scales

in a bipolar format. Though dividing bipolar dimensions into unipolar measures may be tempting because they double the amount of measurements taken (Stern et al., 1997), this comes at the cost of introducing troublesome issues in analysis and interpretation of the resulting data.

Thirdly, the evidence indicates that a VAS is a valid and reliable way of reporting and scaling notional quantities (Scott & Huskisson, 1976; Dixon, 1986; Dixon & Bird, 1981) with little difference in accuracy between horizontal and vertical formats (Scott & Huskisson, 1979).

In section 3.5, the limitations of the existing VAMS were discussed with some emerging solutions suggesting a central design concept for a new mood assessment instrument. The design concept essentially comprises a number of animated scales (D-VAMS) (see 3.5.3) in which images of a facial expressions are modulated using a slider. Each scale would consist of a total of 101 images each representing 100-point intervals along its respective length. This level of granularity would be achieved by photographing a number of facial expressions representing intervals along the scale, identifying their empirical placement, and then morphing them into sub-transitions to create a smooth and seamless blend through all of the images.

This novel approach – *explicit interpolation* – should allow for a more concrete way to select a quantity of a particular affect type, as rather than relying on an abstract cognitive process to report one's mood as a position on a scale, an image will be presented dynamically which expresses a quantity of an affect type through a corresponding image of facial expression. This approach releases the respondent from the burden of trying to imagine how a particular affective quantity would scale as a numeric value or measurable length of visual space. It also means that if – as Price

(1999) contends – stroke patients are generally too cognitively impaired to use a VAS, then this approach may offer a solution by circumventing the need for any cognitive interpolation at all.

An important and innovative aspect of this new design was the use of the CMA as a an underlying theory. The Circumplex can be used as a geometric space across which a number of bipolar scales cross at different angles, thereby representing a kind of repeated measures, with each scale tapping into different proportions of valence and activation.

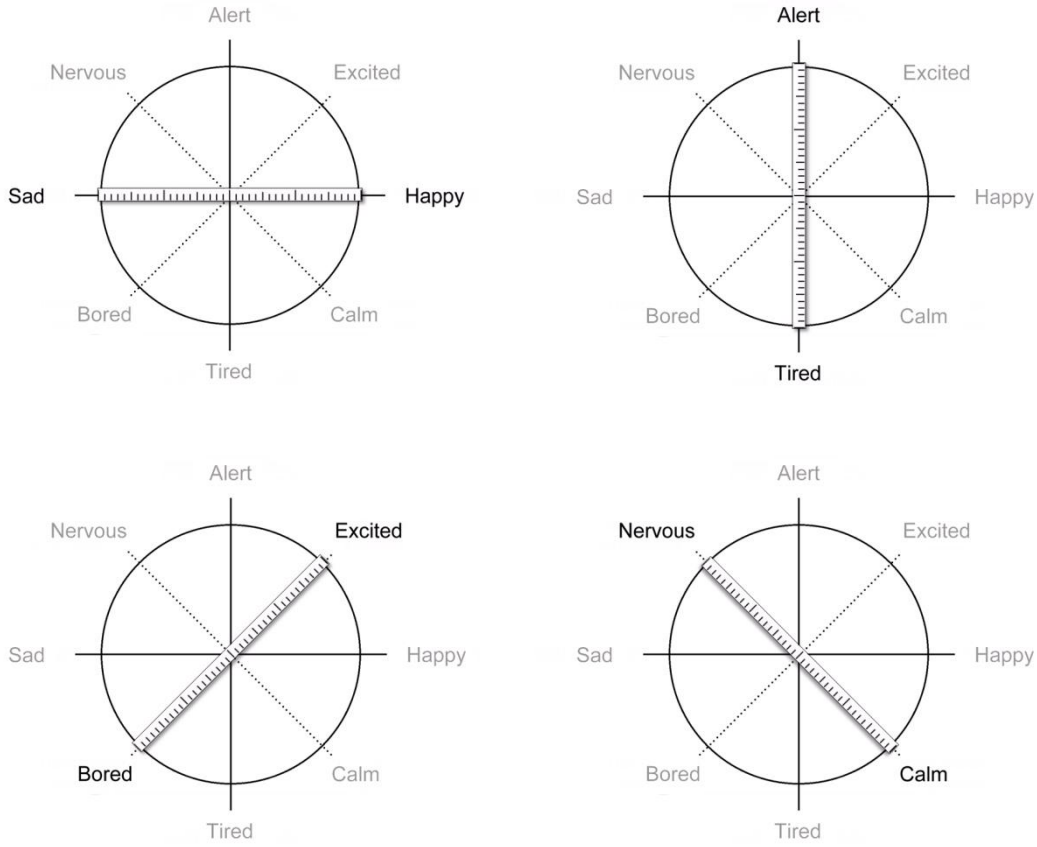


Figure 4.5. *Examples of Prospective Scales across CMA Affect Space*

Examples of prospective scales that could be derived in this manner are illustrated in Figure 4.5. Scales crossing 2-factor space at different angles may allow

for an instrument with separate, individual scales denoting different mood types, while also enabling their scores to be unified using the CMA as an underlying coordinate system.

In order to accomplish this, however, we need to explore a number of questions. Which mood descriptors are most clearly associated with their corresponding facial expressions? Which bipolar scales should be chosen from these various mood words, and which scales provide the most comprehensive coverage of affect and emotion states? We also need to examine judgements of facial expression in a number of actors, selecting the strongest candidates to pose facial expressions for the final scales. These actors need to be selected based on their ability to convey recognisable facial expressions for the mood words we select.

4.3 Study Plan and Protocol Summary

In order to develop a new, nonverbal mood assessment instrument, a basic three-tier protocol was outlined consisting of three studies (see Figure 4.6). In Study 1, the CMA structure was tested in judgements of facial expressions, suitable mood words were identified to form bipolar scales across affect space, and actors were ranked according to their overall ability to pose identifiable facial expressions. In Study 2, two high-scoring actors were recalled to pose facial expressions representing intervals along each of the scales. A series of these images was then selected to form continua for the individual scales, and a further judgement study provided data to create a coordinate system to unify scale scores into a single metric. Next, a further judgement study established scaling data to empirically place the position of each of the images upon their respective continua. In Study 3, consecutive images were morphed to

generate further images so that a finely graded continuum of images for 100-interval scales could be generated, and the scales were implemented as a software-based interface.

Finally, a validation study took place in which the scales were validated against a suitable criterion measure in a sample of stroke survivors, and the psychometric properties of the instrument were examined.

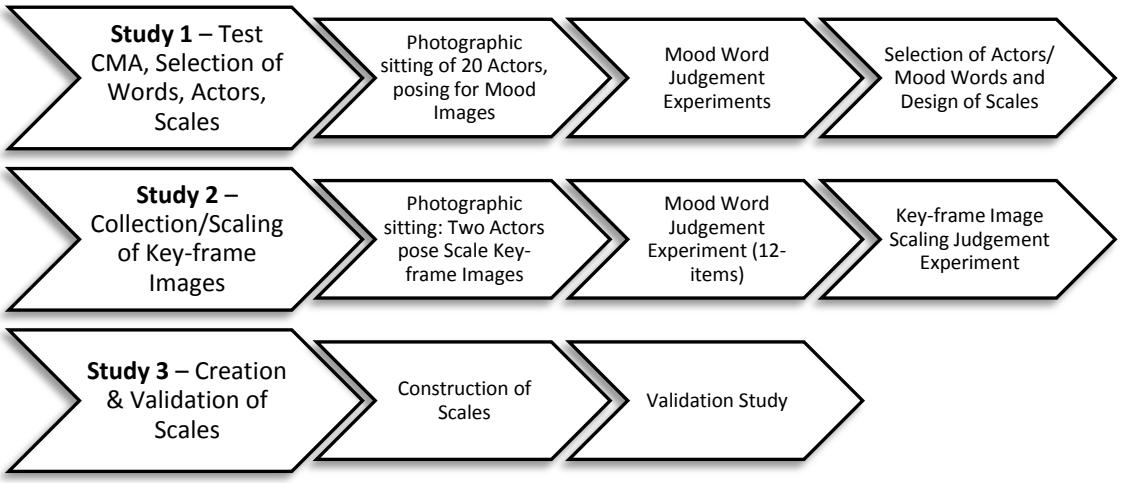


Figure 4.6. *Study Plan*

5. Study 1: Experimentally Identifying Candidate Scales

5.1 Introduction

5.1.1 Objectives. The first objective was to select and test a variety of mood words for their ability to evoke recognisable facial expressions. The weaker words (i.e.: those less able to evoke a recognisable expression) would then be eliminated, based on which a smaller pool of candidate descriptors would be selected for further study. The second objective was to confirm that judgments of facial expression are consistent with a two-factor, Circumplex Model of Affect of the type that has emerged from previous research, and upon which it was hoped that the new scales could be based. If a circumplex structure were confirmed, then the smaller collection of descriptors whose corresponding facial expression could most reliably be identified would act as anchors within a coordinate system within which a number of bipolar scales would be created.

Thirdly, having established a smaller pool of candidate items, the experiment would then be duplicated with these items and corresponding images only. The objective of this would be to refine the coordinate system and offer independent validation of it in an independent sample, but also to offer detailed enough data to identify actors who are particularly skilled at posing recognisable facial expressions, and whose images would later be used as part of the prototype scales.

5.1.2 Selection of mood words. The first stage was to select a suitable pool of mood words which are commonly used to describe states of core affect. The

criteria for choosing the words were threefold: Firstly, the selection as a whole needed to be comprised of descriptors that are spread reasonably evenly around the circumference of the circumplex, so as to increase the likelihood of strong candidate words being identified across all the quadrants and poles of the circumplex. Secondly, the words should represent persistent mood states rather than the more complex, transient states better described as emotions. Thirdly, the words should also include those used by the VAMS, so that direct comparisons are possible between the two scales. This latter criterion did to some degree conflict with the second, as the VAMS includes the items 'anger' and 'confusion', whose status as states of core affect is debatable.

The choice of these mood words was guided by previous research in the area, and focussed upon descriptor words that had already been studied to some degree. Studies of the circumplex structure of mood (Russell, 1980; Russell & Bullock, 1985; Vastfjall et al., 2002; Yik et al., 2011), the influential Positive and Negative Affect Schedule (PANAS/PANAS-X) (Watson et al., 1988; Watson & Clark, 1994; Crawford & Henry, 2004) and other mood measurement studies involving the use of adjective checklists (McNair et al., 1971; Bush, 1973; Thayer, 1967, 1986) offered a wide-ranging source of mood adjectives to choose from.

Since the purpose of the scales to be developed was to accurately measure and discriminate between mood states, it was also felt prudent to address the issue of problematic correlations that are frequently reported between measures of anxiety and depression, and which make the two constructs difficult to discriminate empirically. Feldman (1995) examined the circumplex model in relation to this issue, and identified a scale that has proven useful in distinguishing 'anxious' versus 'depressed'

mood states. The Emotion Questionnaire (EQ) is a version of the Multiple Affect Adjective Check List (MAACL) designed to maximise differentiation between anxiety and depression (Higgins et al., 1985), and adjectives from the EQ anxiety and depression subscales were therefore incorporated into the initial pool.

After careful consideration an initial selection of 24 mood words was made (see Figure 5.0). These are charted at their theoretical locations on the circumplex based on the findings of Russell (1980) and Yik & Russell (2011). Items included on the EQ anxiety subscale are marked with an *A*, and those appearing on the EQ depression subscale are marked with a *D*. Items appearing on the PANAS are marked *p*, and items included in Russell's (1980) study are marked with an *r*.

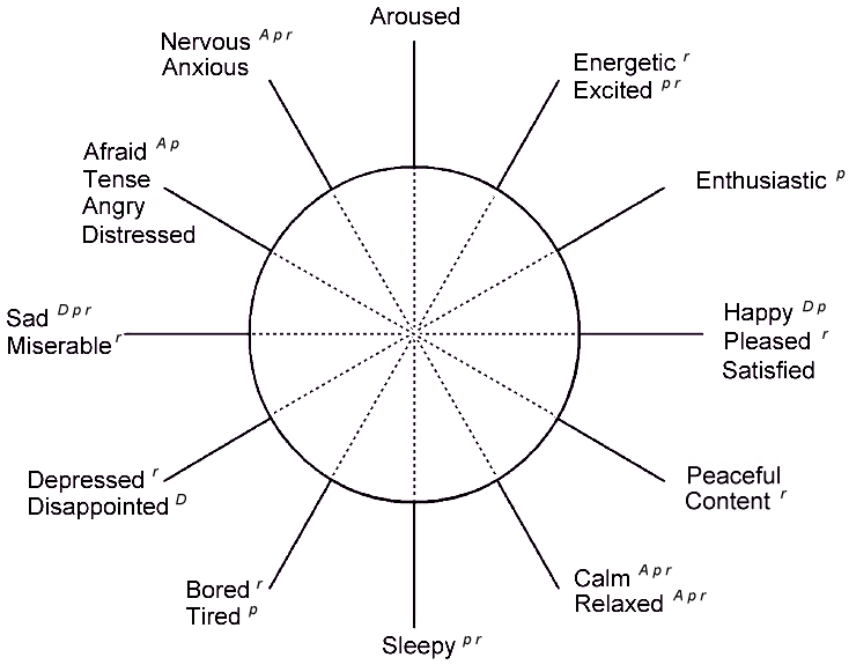


Figure 5.0. *Adjectives selected for the initial pool.*

In addition to these items, ‘confused’ and ‘neutral’ were also added. ‘Confused’ was included because it one of the scales used by the VAMS, and ‘neutral’ was included in order to identify a midpoint or origin of the circumplex. Once these were included, the complete pool comprised a total of 26 items:

Energetic	*Disappointed*	*Calm*	*Confused*
Pleased	*Tense*	*Bored*	*Sleepy*
Excited	*Angry*	*Tired*	*Content*
Happy	*Nervous*	*Miserable*	*Peaceful*
Enthusiastic	*Distressed*	*Depressed*	*Neutral*
Sad	*Anxious*	*Satisfied*	
Afraid	*Aroused*	*Relaxed*	

A Study in three parts was then designed to test these words. For the first part of the study, sets of photographs were produced of participant ‘actors’ posing mood states in response to the stimulus words. For the second part, the photographs were rated by participant on scales for each of the corresponding mood descriptors. Finally, based on the results of this judgement experiment, a smaller pool of words would be selected on which a further judgement study would be performed. Approval for this study was granted by the Institute of Work, Health & Organisations (iWHO) Ethics Committee, University of Nottingham.

5.2 Part 1: Producing Photographs of Posed Facial Expressions

5.2.1 Recruitment and participants. Advertisements were posted around the University of Nottingham and on local, community centre notice boards asking for

volunteer actors. The advertisement stipulated that candidates should be "spontaneous and animated, and capable of authentically posing facial expressions". Selection criteria were that participants 1) were at least 18 years of age, 2) were fluent in English, 3) had no facial hair, tattoos or significant facial disfigurement and 4) were not blind or had any medical condition that impaired their ability to control their facial expression. A small monetary incentive was offered for participation. In total, 20 participants, 10 male and 10 female, aged 18 to 26 years, (mean 20.6 years; S.D. = 1.7 years) each provided a satisfactory set of photographs. Of these participants, 12 (60%) were ethnically European, 6 (30%) were Asian or East Asian and 2 (10%) were of African descent. 12 (60%) spoke English as a first language, while the remaining 8 (40%) spoke English fluently as a second language. The photographs of one additional participant were omitted from the study because the posed expressions were deemed too similar to one another.

5.2.2 Methods and equipment. For the photographic sittings, a portable studio was assembled consisting of a DSLR camera, adjustable flash unit, stands, softbox, and backdrop. The studio could be set up at any indoor location, and was assembled in rooms that were booked specifically for the purpose of the study. All participant 'actors' underwent a single sitting lasting 25-45 minutes.

Upon arrival, participants were first screened to ensure that they met the criteria for participation in the experiment. If necessary, participants were asked to remove any makeup or facial jewellery (rings, studs etc.), and to tie back long hair. A black T-shirt was provided for them to wear for the duration of the sitting. They were also given a copy of the XVAMS Information Sheet (see

Appendix I) and Instruction Sheet (see Appendix II) to read in order to understand the purpose of the study and what they were required to do. The instruction sheet included two suggested methods for helping to summon up a suitable facial expression: *Remembered Moments* and *The Mime*. 'Remembered Moments' is a technique often used by researchers to help evoke a particular emotion or mood state in a participant, and has been found to be useful in facilitating the associated facial expression (Yik & Russell, 2003). 'The Mime' is an alternative strategy invented to emphasise the role of the facial expression in communicating a particular mood state. In this approach, the participant was asked to imagine that they were trying to communicate their mood state to an imaginary person behind a window pane through which the participant could be seen but not heard. Participants were provided with these methods as suggestions, but were not obliged to use either. For some participants, the photographic sitting was arranged in advance, in which case details of these requirements and copies of the instruction and information sheets were provided beforehand. Once the participant had familiarised themselves with the information sheets, a consent form was provided (see Appendix III), with statements detailing the conditions of participation, each with a corresponding checkbox. Participants were required to tick the boxes to confirm that they understood the conditions, and asked to sign and date the form and give their age in a space provided.

When the participant was ready to begin, they were seated at a desk in front of a white backdrop. The desk contained a small mirror which they could use if they wished to practise a pose first, and a pile of laminated, A5 size cue cards containing the descriptor words for the expressions to be posed. The order of the cards for each sitting was pre-arranged according to a randomly ordered, computer-generated list.

The participant was asked to face forward and avoid sideways tilts or turns of the head throughout the session, but told that raising or lowering the head slightly was permitted if it naturally accompanied the facial expression.

For each pose, the participant was asked to take, in turn, a card from the top of the pile and hold it up to the camera such that the mood word fell within the camera's field of view. One photograph was taken of this for reference purposes, after which the participant was directed to put the card aside. They were then asked to pose the mood state as the word on the card was repeated to them verbally. At least 3 photographs was taken of the pose (mean 3.8; S.D=0.54), with more being taken if the participant had trouble producing an expression that was recognisable. This process was repeated until the mood words on all 26 cards had been posed.

After the sitting, the participant was asked to indicate which of the two methods described in the instructions ('Remembered Moments' or 'The Mime') had proven most useful in facilitating their poses, but told that they could respond with 'neither' or 'both' if a preference could not be stated.

5.2.3 Results: preferred method. Of the methods suggested, 12 (60%) of the participants stated a preference for the 'remembered moments' method, while 5 (25%) said that they preferred the 'mime' method. Three (15%) participants expressed no preference, reporting that they used neither or both of the methods. The results suggest that there may be some utility in including the 'mime' method as an option in future studies involving posing facial expression.

5.2.4 Screening and selection of images. Once the photographs had been collected, the next stage was to choose one of the photographs taken for each mood state posed by each actor. These selected photos – 26 per actor – would comprise the

image sets to be used for the next part of the study. In order to select the best candidate photographs, a screening procedure was created to eliminate photographs of poses deemed to be weakest. First, any candidate photographs surplus to the 3 required for each posed expression were eliminated based on the experimenter's judgement. Next, the best of 3 remaining posed expressions was identified by a scoring procedure. For this procedure, three judges independently assigned ranks to each of the facial expressions on a 'best-of-three' score sheet, with the rank of each being based on the judges' assessment of the degree to which each expression was deemed to concur with the mood word to which the actor was responding. Tied ranks were permitted where no preference existed between particular images. The results for the three judges were summed in order to identify the highest scoring images, and where a tie existed a random number generator was used to make a selection. The 26 images per actor emerging from this selection process comprised the image sets which were to be judged in the following stage (See Appendix IV to VII).

Finally, the images were cropped and scaled such that the faces were uniformly centred and the frame of the image showed an inch or two to either side of and above the top of the head, with the base of the image terminating near the neckline of the actor's T-shirt.

5.3 Part 2: Judging the Photographs

5.3.1 Sample size. It was decided to use the variable reduction technique Principal Components Analysis (PCA) to extract and examine the factor structure in data from this study. Guidance on sample size for PCA recommends that not less than 100 cases should be selected regardless of the number of variables under examination

(Gorsuch, 1983). The practical constraints of this study called for participants to be allowed to respond to more than one of the 20 images sets, and so a more flexible interpretation of caseness was adopted by which it applied to the number of datasets for participant/image-set combinations, rather than number of participants *per se*.

The software package G*Power (Faul et al., 2007) was also used to provide some guidance in this respect. Though a PCA does not entail an examination of statistical significance based on a specific, hypothesis-driven test, examining recommended sample sizes across a range of prospective statistical tests was a good way of corroborating this advice. Though a paper could not be found which reported effect sizes for a study of this type, qualitative assessment of the likely magnitude in judgements of facial expressions suggest an effect size of somewhere between 'medium' and 'large', as defined by Cohen (1992). Though G*Power does not provide a PCA analysis option for sample size calculation, a number of figures were examined which generally offer sufficient power (0.8 or more) for studies of a medium to large effect size to detect statistically significant effects (α of 0.05 or less) using a range of standard parametric and non-parametric tests. Sample size figures returned based on these were generally in the region of 100. However, as a necessary compromise in order to keep the experiment within manageable proportions, this figure was taken as a guide to the number of unique participant/actor, datasets, rather than a specific sample size for participants.

5.3.2 Recruitment and participants. Recruitment took place by means of advertisements posted about University of Nottingham, offering a small payment for participation in an online task. Selection criteria were that participants 1) were at least 18 years of age, and 2) were fluent in English. The data were provided by a total of 44

participants, 21 male and 23 female, aged 18 to 44 years (mean 24.6 years; S.D. = 6.2 years). Of these participants, 14 (32%) were ethnically European, 25 (56%) were Asian or East Asian, 2 (4.5%) were of African descent and 3 (6.8%) described themselves as mixed race. 24 (55%) spoke English as a first language, while the remaining 20 (44%) spoke English fluently as a second language.

5.3.3 Method. A website and back-end database was purpose-built to administer the judgement tasks for part two of the study (Experiment 1), and to automate the collection of the judgement data and participants' demographic information (xvams.com).

Before being allowed access to the website experiment portal, participants first completed a brief sign-up process in which they set up a username and password and provided an email address where they could be contacted. They were then asked for some personal details: their gender, year of birth, ethnicity, and whether or not they speak English as a first language. Participants not speaking English as a first language were also asked to select their country of origin from a dropdown list. Finally, participants were asked to indicate consent by checking five clauses relating to the terms and conditions of participation (see Appendix V), confirming that they had read the study information sheet (see Appendix X) and were aged 18 years or above.

The main page for the experiment portal consisted of a list of open experiments (tasks). Each experiment contained a complete set of images for the 26 mood states posed by a single actor. Upon starting an experiment, these images were presented, in turn, on a separate page, accompanied by 26, 7-point Likert scales corresponding to the mood words used as stimuli for the posed expressions (See Figure 5.1). For each image, the participant rated each of the mood words according

to how well they deemed it to agree with the image. Both the order of presentation of the images, and the order in which the mood words were listed were randomised, with each page containing the list of mood words in a newly randomised sequence. Participants were free to save their data at any point in the session and come back to it later, allowing them to complete tasks at their own rate as time permitted. Participants were invited to complete at least one but no more than three separate experiments, which they were free to select at random from the list.

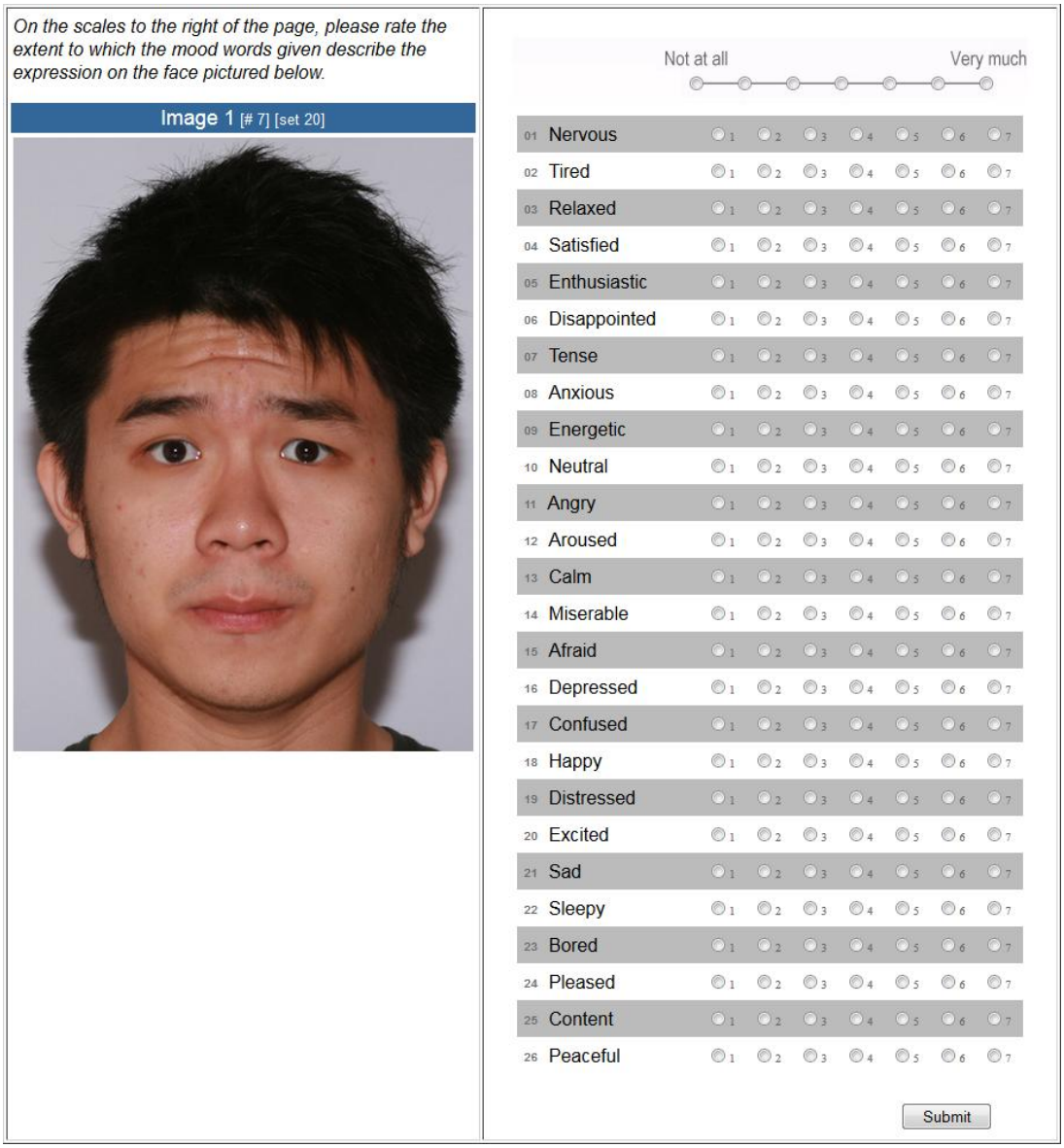


Figure 5.1. *Experiment 1 - Response Page for the Judgement Task from the Project Website*

5.3.3 Analysis. A total of 100 datasets were completed, with each participant completing a mean of 2.3 experiments each (S.D.= 1.04). Systematic response biases in participants returning Likert scale evaluations of this kind are known to cause problems in factor extraction, resulting in distorted correlations and anomalous factor structure (Larsen & Diener, 1992). In order to compensate for this, the data were centred such that an individual's score was represented as a function of its departure from the mean response for a given stimulus image. This centring of data compensates for idiosyncratic biases in the use of the Likert scale and allows for a more accurate representation of the data, however the un-centred data was also retained and analysed for comparison purposes.

The datasets were summed to form a summary table comprising a matrix of totals (see Table 5.0). This table gives a detailed profile of the way in which participants judged facial expressions evoked by these words. Mood words were ordered according to their positions on the circumplex as has been established in prior research. The circumplexical structure of the data is evidenced by a clear diagonal trend highlighted in the grey-scaled results, whereby correspondences tend to occur most strongly where scored mood words match the one that evoked a particular facial expression, and falling off in proportion to the displacement of these words from one another as predicted by their position on the circumplex.

For the first part of the analysis, a principle component analysis (PCA) was conducted on the table of the summed data for all 100 datasets. A PCA analysis can be conducted on either rows or columns (in what are often known as *r-type* and *p-type* analyses). If columns are used, then the variables are the mood words, and the resulting plot becomes one of significances of mood words. If rows are used, then the

expressions (or faces) become the variables, and the resulting plot becomes one of significance of facial expressions. Since both row-wise and column-wise analyses are equally legitimate projections of the underlying factor structure, both analyses were performed to give as clear a picture as possible of the data's composition.

However, performing a PCA on summary data means that intra-individual variations are collapsed out of existence, inflating the variance accounted for and yielding a factor structure that may be over-simplified. This type of analysis is known as an *indirect PCA*, and though useful for this purpose, it is also desirable to run an analysis on the complete dataset. Therefore a second, more detailed analysis was also conducted on the data at the cell-level. Instead of using data comprising 26 rows of summed totals of all responses to each facial expression, all 2600 rows of data were included and a PCA performed on this also. Furthermore, this analysis was performed both row-wise and column-wise – as before – in order to examine the factor structure from both of these perspectives. In all of these analyses, the factor solutions were left unrotated. Though rotations such as varimax are frequently applied to arrive at simple structure (i.e: where variables load strongly on one factor, but weakly or not at all on the other factors), circumplexical solutions do not have simple structure and so rotation is not recommended (Acton & Revelle, 2002).

The data were then plotted onto charts showing loadings of variables on the extracted components. These plots were examined to understand the locations of the mood words/expressions in factor space and establish a basic coordinate system within which the prospective scales will be anchored. Finally, these plots, accompanied with a qualitative examination of the summary table and the S.D. metrics therein were used to identify and eliminate 'weaker' items which either

performed poorly as stimuli for recognisable facial expressions, or which were shown to be essentially indistinguishable from other, similar items.

5.3.4 Results.

5.3.4.1 Indirect principle component analyses. The correlation matrix was first subjected to a column-wise principle components analysis, which yielded three components with eigenvalues larger than 1.0 accounting for 94.3% of the variance (Table 5.1a). Factor loadings of mood words are shown in Table 5.1b. Examination of factor loadings identifies factors 1 and 2 as corresponding respectively to the predicted valence and activation dimensions. Factor 1, which accounts for 70.3% of the variance, loads proportionally to the negative valence of mood words, while factor 2, which accounts for 19.1% of the variance, corresponds to the level of activation or arousal associated with a given mood word. In total these two factors account for 89.4% of the variance. A third factor, accounting for 4.9% of the variance, was not as easy to interpret. Though the 'neutral' item had a particularly high positive loading on this factor in comparison to other mood words, it did not seem to follow a readily discernible pattern in relation to the loadings of other mood words. The alternative *p-type* (expressions) analysis reveals very similar results. To examine whether the data corresponded to a circumplex structure, a plot was made of loadings of the mood words on *x* and *y* axes representing the valence and activation factors respectively. The results show clear evidence of circumplexity, with the positions of mood words in factor space closely following those of other studies in which adjectives denoting affect were examined (Bush, 1973; Feldman, 1995; Russell, 1980; Russell & Bullock, 1985; Stanley & Meyer, 2009; Vastfjall et al., 2002; Yik et al., 2011).

Table 5.0. *Summary of Word Scores(centred) from 100 datasets (mean Likert score = x/100)*

			Total Word Score (by ID)																								
Posed Mood	*ID*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*
Happy	1	180.2	351	296	233	237	220	5.08	-123	-128	-121	-153	-143	-145	-162	-164	-149	-160	-158	-146	-125	-137	121	90.1	62.1	268	210	-80.9
Pleased	2	158.6	297	277	133	153	107	9.04	-109	-112	-122	-143	-147	-132	-153	-149	-133	-144	-144	-130	-105	-119	166	151	109	247	206	-14
Excited	3	175.1	312	279	301	271	265	15.5	-99.5	-88.5	-110	-129	-139	-132	-153	-156	-144	-143	-149	-139	-142	-150	69.5	41.5	8.54	226	186	-92.5
Enthusiastic	4	160.9	294	244	278	230	223	22.5	-98.5	-83.5	-101	-132	-112	-132	-153	-150	-138	-141	-146	-138	-119	-137	83.5	52.5	16.5	222	170	-57.5
Energetic	5	159.7	290	254	279	243	231	44	-84	-75	-84	-118	-119	-120	-150	-150	-136	-143	-141	-135	-131	-141	63	49	-10	201	154	-70
Aroused	6	63.55	94.4	105	50.4	51.4	57.4	8.38	-30.6	-19.6	-13.6	-31.6	-6.62	-51.6	-80.6	-96.6	-79.6	-76.6	-69.6	-56.6	-45.6	-72.6	56.4	49.4	74.4	90.4	51.4	42.4
Nervous	7	80.14	-92.6	-78.6	-76.6	-74.6	-91.6	-56.6	92.4	120	109	88.4	180	101	-31.6	33.4	62.4	49.4	45.4	5.38	21.4	-41.6	-66.6	-60.6	-49.6	-86.6	-74.6	-27.6
Anxious	8	86.23	-100	-92.1	-66.1	-78.1	-80.1	-43.1	138	151	146	98.9	140	126	-1.08	29.9	47.9	44.9	21.9	-7.08	-0.08	-53.1	-83.1	-69.1	-66.1	-90.1	-80.1	-35.1
Tense	9	98.19	-120	-106	-82.2	-90.2	-82.2	-41.2	113	123	188	70.8	171	142	52.8	65.8	61.8	80.8	45.8	-10.2	13.8	-41.2	-101	-100	-80.2	-115	-97.2	-60.2
Afraid	10	102.3	-93.2	-96.2	-4.19	-53.2	-10.2	-15.2	133	146	165	195	207	144	-17.2	3.81	42.8	22.8	1.81	-81.2	-28.2	-88.2	-109	-101	-102	-95.2	-89.2	-76.2
Confused	11	95.89	-100	-90.3	-70.3	-73.3	-66.3	-37.3	65.7	141	168	1.65	228	134	70.7	31.7	59.7	91.7	34.7	-17.3	21.7	-29.3	-96.3	-101	-90.3	-99.3	-87.3	-87.3
Distressed	12	106.9	-125	-114	-97	-104	-88	-68	92	127	160	87	173	169	22	90	117	132	75	-15	9	-41	-113	-103	-89	-114	-102	-80
Angry	13	119.3	-121	-116	-74.4	-86.4	-37.4	-16.4	44.6	116	243	3.58	42.6	169	321	39.6	82.6	125	53.6	-46.4	-17.4	-64.4	-114	-120	-109	-111	-105	-96.4
Sad	14	116.1	-133	-129	-124	-121	-128	-100	6	65	75	48	82	138	-21	213	170	181	173	66	88	31	-95	-89	-72	-127	-110	-87
Miserable	15	111.4	-137	-124	-131	-108	-119	-87.3	-2.31	61.7	98.7	-6.31	72.7	128	15.7	190	165	217	130	77.7	86.7	2.69	-89.3	-89.3	-60.3	-127	-97.3	-64.3
Disappointed	16	110.8	-135	-130	-135	-122	-135	-93.6	8.42	33.4	63.4	-21.6	61.4	105	38.4	157	160	201	116	133	135	53.4	-82.6	-88.6	-49.6	-126	-111	-41.6
Depressed	17	107.3	-130	-116	-132	-112	-129	-99.2	-18.2	31.8	54.8	-19.2	29.8	104	-4.15	184	163	163	172	116	114	63.8	-81.2	-63.2	-33.2	-115	-105	-37.2
Bored	18	94	-115	-101	-112	-108	-105	-84.8	-32.8	5.23	42.2	-57.8	12.2	50.2	19.2	113	101	148	114	191	139	91.2	-45.8	-62.8	-4.77	-98.8	-87.8	-12.8
Tired	19	117.6	-95.3	-80.3	-116	-116	-129	-96.3	-72.3	-50.3	-29.3	-69.3	-24.3	-24.3	-71.3	13.7	22.7	6.73	29.7	180	295	342	69.7	39.7	81.7	-63.3	-51.3	8.73
Sleepy	20	125.8	-103	-93.4	-133	-112	-130	-83.4	-84.4	-60.4	-50.4	-85.4	-51.4	-15.4	-70.4	8.62	0.62	12.6	9.62	162	318	363	120	77.6	80.6	-68.4	-43.4	34.6
Relaxed	21	93.38	63.5	86.5	-47.5	-31.5	-43.5	-40.5	-76.5	-80.5	-60.5	-94.5	-75.5	-83.5	-92.5	-74.5	-68.5	-85.5	-76.5	5.46	16.5	-2.54	165	142	179	110	96.5	168
Peaceful	22	97.51	66.7	85.7	-61.3	-18.3	-63.3	-38.3	-93.3	-73.3	-66.3	-96.3	-80.3	-83.3	-106	-70.3	-79.3	-67.3	-66.3	-10.3	12.7	-1.31	185	161	193	120	116	137
Calm	23	92.81	56	89	-45	-24	-41	-59	-74	-68	-57	-90	-62	-88	-87	-71	-79	-73	-77	4.96	-2.04	-14	173	159	180	96	82	172
Satisfied	24	131.6	215	244	75.9	81.9	58.9	-11.1	-106	-107	-97.1	-127	-98.1	-113	-134	-121	-130	-114	-125	-86.1	-73.1	-81.1	153	137	146	212	186	14.9
Content	25	120.2	191	196	26.5	44.5	11.5	-38.5	-109	-94.5	-85.5	-117	-97.5	-97.5	-121	-109	-111	-107	-110	-61.5	-46.5	-73.5	171	151	167	195	156	75.5
Neutral	26	74.18	-81.3	-69.3	-109	-87.3	-93.3	-83.3	-28.3	-14.3	13.7	-28.3	-9.27	-8.27	-28.3	14.7	26.7	24.7	23.7	106	65.7	15.7	44.7	50.7	132	-58.3	-29.3	210

Table 5.1a. *Indirect PCA - Words*
Total Variance Explained

Comp-onent	Extraction Sums of Squared Loadings		
	Total	% Variance	Cumulative %
1	18.279	70.303	70.303
2	4.963	19.088	89.391
3	1.278	4.917	94.307

Table 5.2a. *Indirect PCA - Expressions*
Total Variance Explained

Comp-onent	Extraction Sums of Squared Loadings		
	Total	% Variance	Cumulative %
1	17.690	68.038	68.038
2	4.873	18.743	86.781
3	1.467	5.642	92.423

Table 5.1b. *Indirect PCA - Words*
Factor Loadings

	Component		
	1	2	3
Content	-.993	.012	.016
Satisfied	-.992	.063	-.003
Pleased	-.982	.143	-.051
Happy	-.971	.196	-.109
Enthusiastic	-.896	.380	-.202
Relaxed	-.853	-.480	.161
Peaceful	-.844	-.474	.223
Excited	-.841	.469	-.214
Energetic	-.839	.483	-.195
Aroused	-.732	.608	.089
Calm	-.672	-.661	.312
Neutral	-.259	-.712	.588
Sleepy	.428	-.763	-.229
Tired	.620	-.696	-.208
Bored	.667	-.693	-.181
Angry	.784	.207	.056
Nervous	.794	.477	.312
Afraid	.803	.378	.320
Confused	.861	.342	.250
Anxious	.865	.443	.201
Tense	.883	.390	.185
Sad	.929	-.133	-.210
Depressed	.948	-.143	-.187
Disappointed	.951	-.041	-.192
Distressed	.961	.256	.043
Miserable	.966	-.043	-.161

Table 5.2b. *Indirect PCA - Expressions*
Factor Loadings

	Component		
	1	2	3
Pleased	-.971	-.074	.044
Satisfied	-.962	.051	.122
Happy	-.942	-.229	-.074
Content	-.929	.201	.218
Aroused	-.921	-.136	.314
Enthusiastic	-.917	-.313	-.093
Excited	-.898	-.349	-.131
Energetic	-.897	-.365	-.111
Relaxed	-.733	.577	.349
Peaceful	-.730	.573	.336
Calm	-.721	.562	.390
Sleepy	.212	.876	-.174
Neutral	.256	.812	.396
Tired	.281	.849	-.189
Afraid	.708	-.558	.339
Angry	.783	-.325	.012
Bored	.785	.522	-.192
Anxious	.855	-.296	.380
Nervous	.870	-.179	.393
Confused	.871	-.295	.225
Depressed	.873	.349	-.134
Disappointed	.909	.304	-.115
Tense	.918	-.264	.260
Miserable	.923	.161	-.056
Sad	.924	.146	-.074
Distressed	.936	-.222	.205

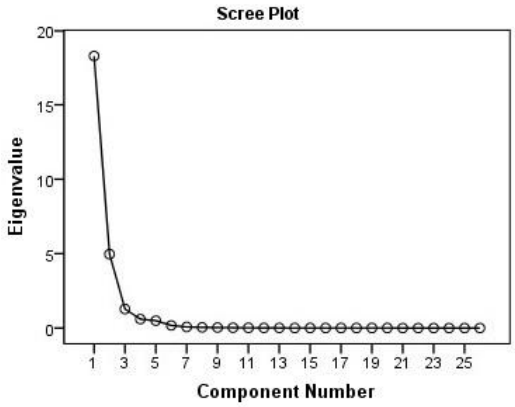


Figure 5.1a. *Indirect PCA - Words. Scree Plot*

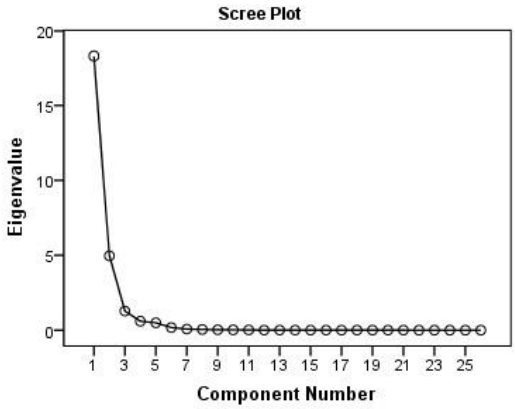


Figure 5.2a. *Indirect PCA - Expressions. Scree Plot*

The plots for both 'words' and 'expressions' (see Figure 5.1b & Figure 5.2b) are fairly similar, with many of the items appearing in approximately the same position on both plots, reflecting similar factor loadings on both of these dimensions. Two-tailed Spearman correlations performed on factor loadings for the two analyses revealed a rho of 0.96 (p<0.001) for factor 1, -0.88 (p<0.001) for factor 2, and 0.92 for factor 3 (p<0.001). The reversal of signs on the loadings for factor 2 relates to the transformation performed on the data in which rows and columns were reversed. The PCA extracts factors which delineate the space of the data examined, however the polarity of axes is arbitrary and a matter of personal interpretation. For convenience of comparison, the y-axis was reversed in the second plot to correct for this.

Inspecting both the summary table and the plots together, it was evident that those items showing the most marked differences between the plots (such as the 'energetic' item) were those which appeared weaker in terms of their ability to evoke recognisable expressions. This offers additional support for decisions regarding the elimination of weaker items, as discrepancies between these plots reflect the variance in results pertaining to each of the items.

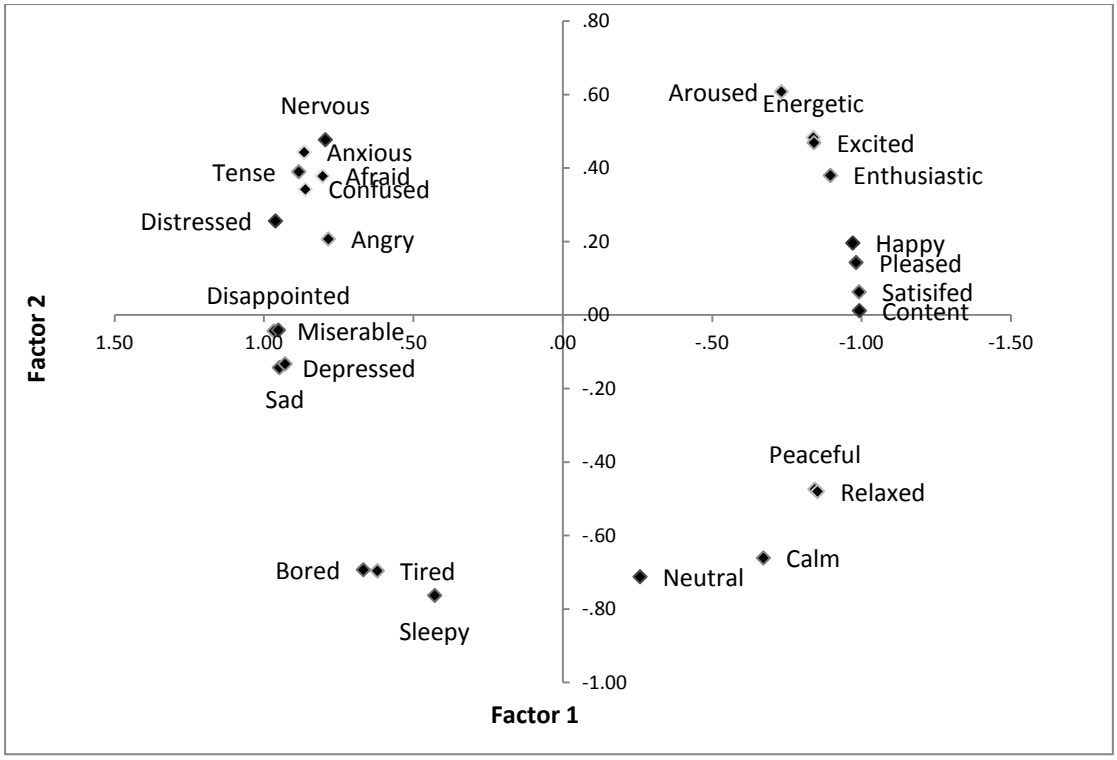


Figure 5.1b. *Indirect PCA - Words. Plot of factor loadings.*

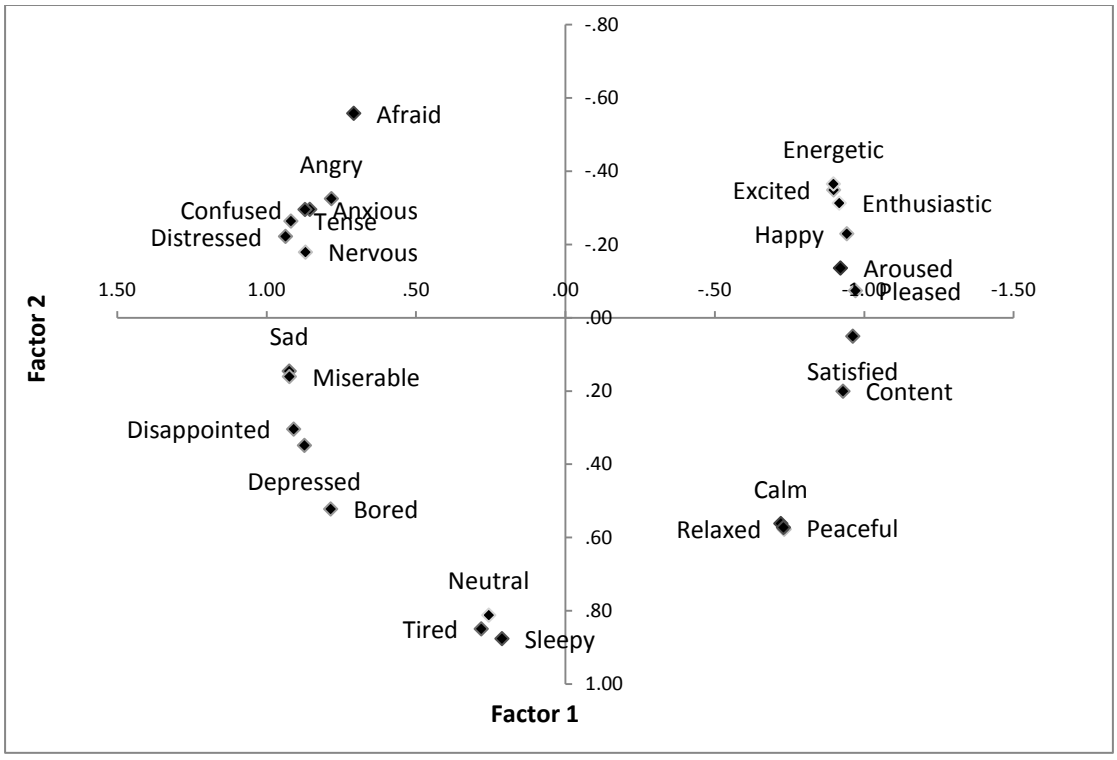


Figure 5.2b. *Indirect PCA - Expressions. Plot of factor loadings.*

5.3.4.2 Cell-level principle component analyses. The data were then subjected to a cell-level analysis whereby the datasets were rendered to 2600 rows of scores analysed as 26 variables. As with the indirect PCA, these data were analysed both column-wise and row-wise to yield plots from both 'word' and 'expression' perspectives respectively. Given that the design of the study was based upon multiple judgements of words being made in response to a single expression, rather than vice-versa (i.e: as scores being returned for a number of facial expressions in response to a single word) it was anticipated that the 'words' (column-wise) PCA would offer a truer reflection of the underlying factor structure, however this latter (row-wise) analysis was also included to offer an alternative perspective that yields further confirmatory evidence.

In this first, column-wise, principle components analysis, four components were extracted with eigenvalues greater than 1.0 accounting for a total of 71.8% of the variance (Table 5.3a). Factor loadings of mood words are shown in Table 5.3b. Examination of factor loadings identifies factors 1 and 2 as corresponding respectively to the predicted valence and activation dimensions. Factor 1, which accounts for 46.6% of the variance, loads proportionally to the negative valence of mood words, while factor 2, which accounts for 14.1% of the variance, corresponds to the level of activation or arousal associated with a given mood word. These two factors account for 61.5% of the variance. Third and fourth factors, accounting for 6.2% and 4.2% of the variance respectively, were not as easy to interpret, though, as in the indirect PCA analysis of words, the 'neutral' item was loaded particularly negatively on the third factor. Neither of these additional factors, however, appear to follow a readily discernible pattern in relation to the loadings of mood words.

Table 5.3a. *Cell-level PCA - Words*
Total Variance Explained

Comp-onent	Extraction Sums of Squared Loadings		
	Total	% Variance	Cumulative %
1	12.129	46.649	46.649
2	3.850	14.808	61.457
3	1.614	6.206	67.663
4	1.088	4.183	71.846

Table 5.3b. *Cell-level PCA - Words*
Factor Loadings

	Component			
	1	2	3	4
Pleased	-.902	.153	-.143	-.046
Happy	-.893	.201	-.180	-.014
Satisfied	-.890	.071	-.102	-.088
Content	-.843	.000	-.057	-.098
Enthusiastic	-.755	.399	-.239	.144
Relaxed	-.724	-.438	.170	-.123
Excited	-.719	.455	-.185	.180
Energetic	-.708	.459	-.180	.127
Peaceful	-.695	-.447	.261	-.171
Calm	-.567	-.561	.329	-.211
Aroused	-.318	.361	.104	.258
Neutral	-.270	-.491	.497	-.236
Sleepy	.346	-.674	-.063	.498
Bored	.454	-.594	-.092	.289
Tired	.483	-.653	-.130	.399
Nervous	.559	.439	.406	.148
Angry	.591	.173	-.041	-.197
Afraid	.620	.361	.332	.066
Anxious	.633	.466	.303	.013
Confused	.668	.233	.233	.151
Tense	.669	.414	.258	-.040
Sad	.768	-.124	-.342	-.239
Miserable	.768	-.080	-.337	-.223
Distressed	.769	.244	.026	-.142
Depressed	.785	-.097	-.342	-.208
Disappointed	.795	-.083	-.305	-.193

Table 5.4a. *Cell-level PCA - Expressions*
Total Variance Explained

Comp-onent	Extraction Sums of Squared Loadings		
	Total	% Variance	Cumulative %
1	9.580	36.847	36.847
2	3.291	12.656	49.503
3	2.194	8.440	57.943
4	1.184	4.554	62.497

Table 5.4b. *Cell-level PCA - Expressions*
Factor Loadings

	Component			
	1	2	3	4
Happy	-.848	-.053	.284	.225
Pleased	-.846	.079	.243	.124
Enthusiastic	-.798	-.097	.299	.260
Excited	-.798	-.156	.298	.277
Energetic	-.761	-.164	.307	.262
Satisfied	-.739	.247	.267	.099
Content	-.696	.335	.211	.034
Peaceful	-.411	.611	.141	-.241
Relaxed	-.401	.625	.179	-.259
Calm	-.384	.618	.199	-.219
Aroused	-.362	.190	.431	-.082
Neutral	.215	.642	.107	-.267
Sleepy	.218	.656	-.283	.115
Tired	.292	.649	-.190	.156
Afraid	.481	-.265	.511	-.219
Nervous	.531	.079	.493	-.128
Anxious	.536	-.070	.432	-.159
Bored	.564	.400	-.034	.329
Confused	.574	-.131	.367	-.003
Angry	.602	-.196	.232	-.043
Depressed	.632	.328	.104	.341
Tense	.634	-.077	.396	-.066
Distressed	.671	-.141	.358	-.023
Disappointed	.677	.305	.092	.307
Miserable	.681	.204	.207	.281
Sad	.701	.186	.179	.334

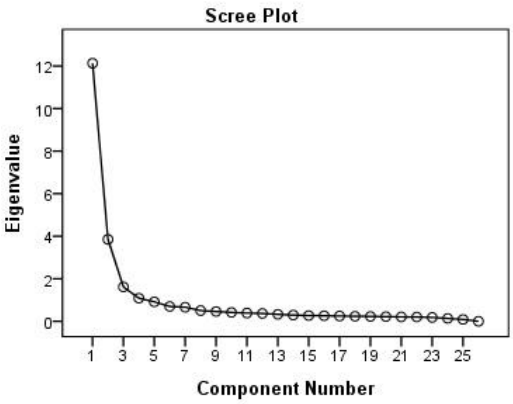


Figure 5.3a. *Cell-level PCA - Words. Scree Plot*

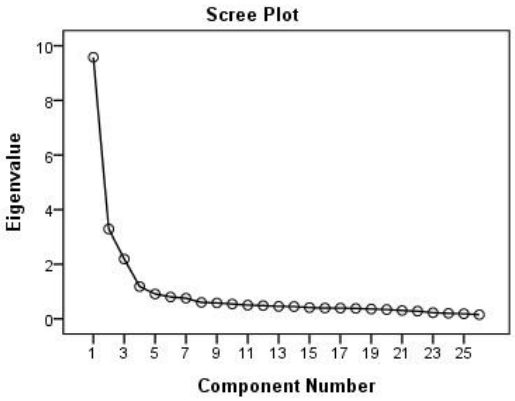


Figure 5.4a. *Cell-level PCA - Expressions. Scree Plot*

The row-wise, 'expressions' analysis (see Table 5.4a & Table 5.4b) yields a similar but less distinct factor structure. Again, four components were extracted with eigenvalues greater than 1.0 but this time accounting in total for only of 62.5% of the variance (Table 5.4a). Factor loadings of mood words (Table 5.3a) again identifies factors 1 and 2 as corresponding respectively to the predicted valence and activation dimensions, with factor 1 accounting for 36.8% of the variance and factor 2 accounting for 12.7% of the variance, together totalling 49.5% of the variance. Again, third and fourth factors – accounting for 8.4% and 4.6% of the variance respectively – were difficult to interpret.

Turning to the plots of the cell-level factor loadings (Figure 5.3b & Figure 5.4b), the 'words' plot is similar the one yielded by the Indirect PCA, with a comparable spread of items throughout the four quadrants. The 'expressions' plot, likewise, has a similar layout to its Indirect PCA counterpart, however in this projection the plot as a whole appears notably out of alignment with the axes, and seem distorted by comparison to the former. The marked difference in variance accounted for and the irregular plot yielded in this latter ('expressions') analysis, however, is understood when we take into consideration the aforementioned

incompatibility of this analysis with the methodology employed. The experiment involved rating 26 mood words *in response to images presented one-by-one*, it did not involve rating 26 expressions *in response to presentation of words presented one-by-one*. The standardisation, likewise, was applied row-wise so as to be appropriate for the design employed. These plots represent two legitimate but distinct projections of the data. In an ideal scenario, the facial expressions would be judged to correspond precisely to their evoking mood word, and the plots would be equivalent. However, though similar, the two plots are telling us different things: the 'words' plot demonstrates how people map the *mood words* used to judge a sample of expressions in semantic space, while the 'expressions' plot tells us how *these particular images* were located in space relative to one another according to ratings on a sample of mood words. Both analyses, however, prove a good test of convergent validity, and demonstrate the essential veracity of the factor structure. However, though, both plots should be used to guide the creation of the bipolar scales within, it is the 'expressions' plots of the faces finally used for the scales that should form the ultimate basis for the coordinate systems in which the scales will be based.

As with the results of indirect PCA, signs on the loadings for factor 2 were reversed in the second projection due to the transformation performed on the data in which rows and columns were reversed. Again, for convenience of comparison, the y-axis was reversed in the second plot to correct for this.

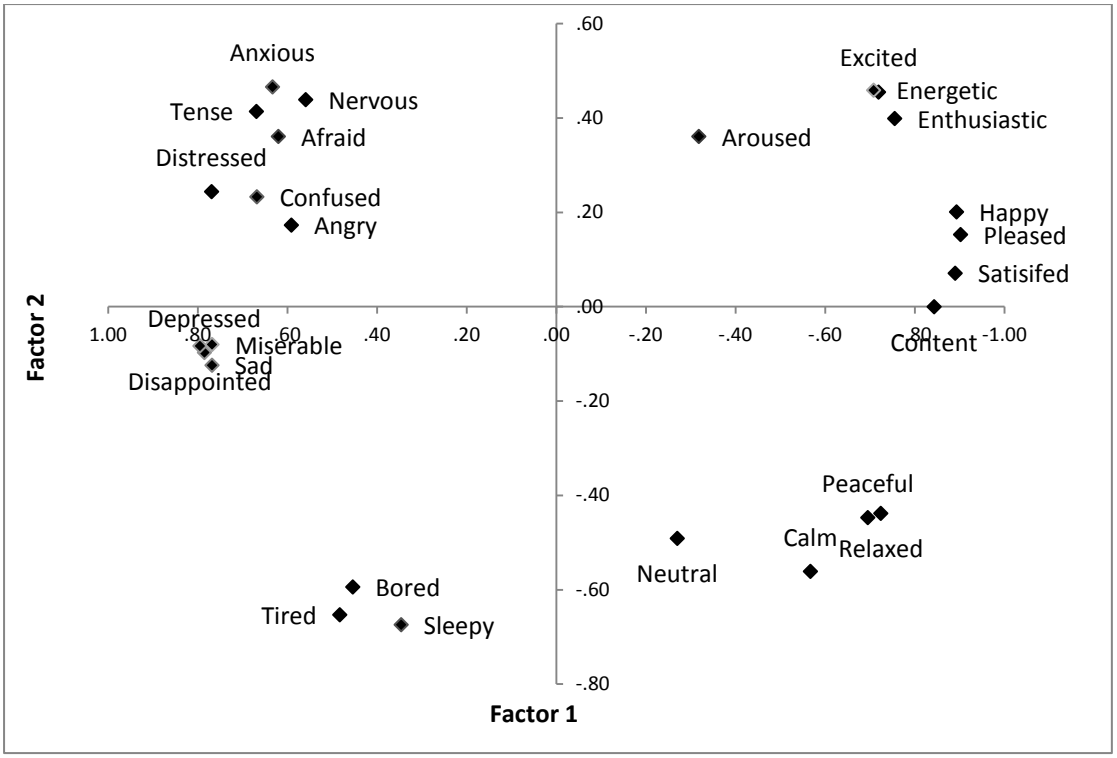


Figure 5.3b. *Cell-level PCA - Words. Plot of factor loadings.*

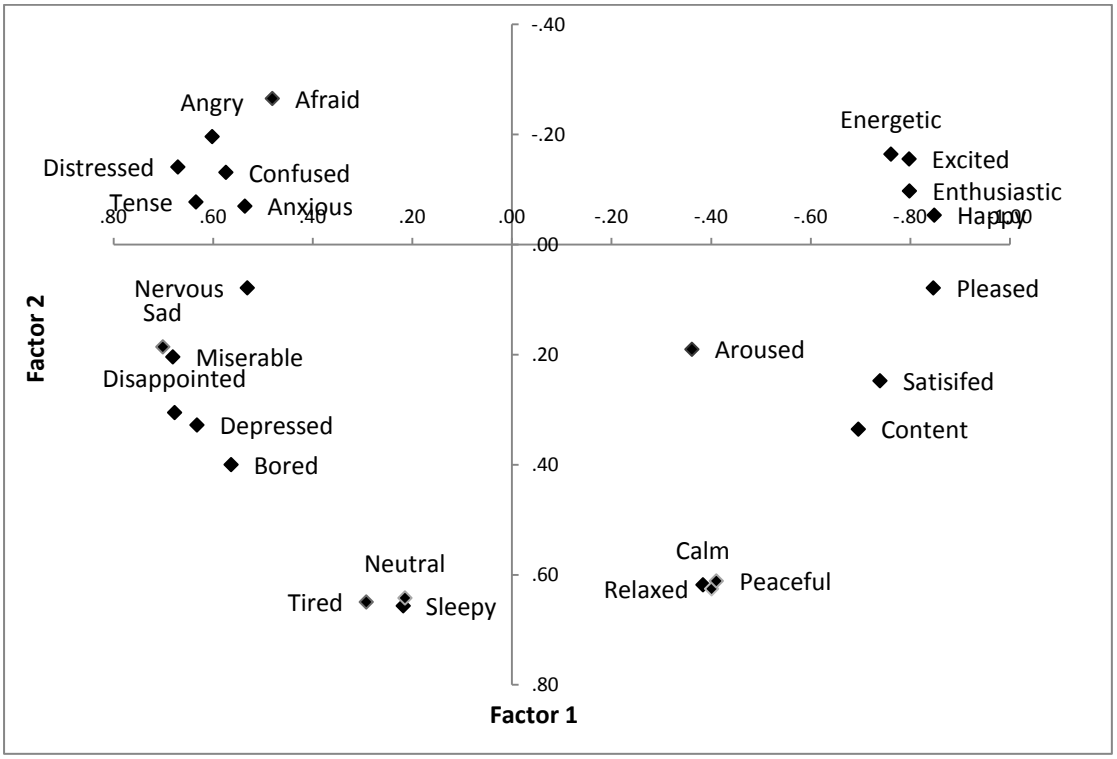


Figure 5.4b. *Cell-level PCA - Expressions. Plot of factor loadings.*

5.3.4.3 Assessing the number of factors. To establish the number of factors that should be deemed an optimal solution, the Kaiser criterion (eigenvalue of 1.0 or above) is often used in conjunction with a qualitative examination of the scree plot, with the amount of variance accounted for by each consecutive factor also being used to guide interpretation. However neither the Kaiser criterion, nor – counterintuitively – the proportions of variance accounted for by successive factors are regarded by experts as being reliable methods of assessing the cut-off point beyond which additional factors should be regarded as superfluous. Velicer et al. (2000) offer a detailed explication of this difficult and controversial area, and evaluate the merits of methods used to determine the number of factors that should be deemed as comprising an optimal solution. The authors deem the Kaiser criterion to be arbitrary, and find that it frequently results in over-extraction of components, while examination of the scree plot is recommended only as an adjunct to other procedures. They recommend the use of a parallel analysis procedure to empirically derive an eigenvalue cut-off point for a given analysis. Sets of random data are generated based on the number of variables under examination and the sample size of the data to be analysed. These random data are subjected to a PCA analysis and a scree plot is derived which can act as a reference line for comparison with a plot for the genuine data. This, in addition to a qualitative examination of the scree plot, can offer a more sound and empirically defensible figure for a suitable eigenvalue cut-off.

An online, parallel analysis engine (Patil et al., 2007) was therefore used to produce a baseline plot to establish a suitable eigenvalue cut-off for this analysis. For the sample size of this study, there are two possible interpretations. Since each dataset comprises a unique combination of judge and stimulus images, one could adopt the number of datasets completed (n=100) as the sample size. However, because

participants usually completed more than one dataset each (mean = 2.3), the datasets can therefore not be considered fully independent, so one might instead adopt the stricter criterion of number of participants. For the purposes of the parallel analysis, both figures were used in order to give a fuller picture of the impact of this figure on the resulting cut-offs.

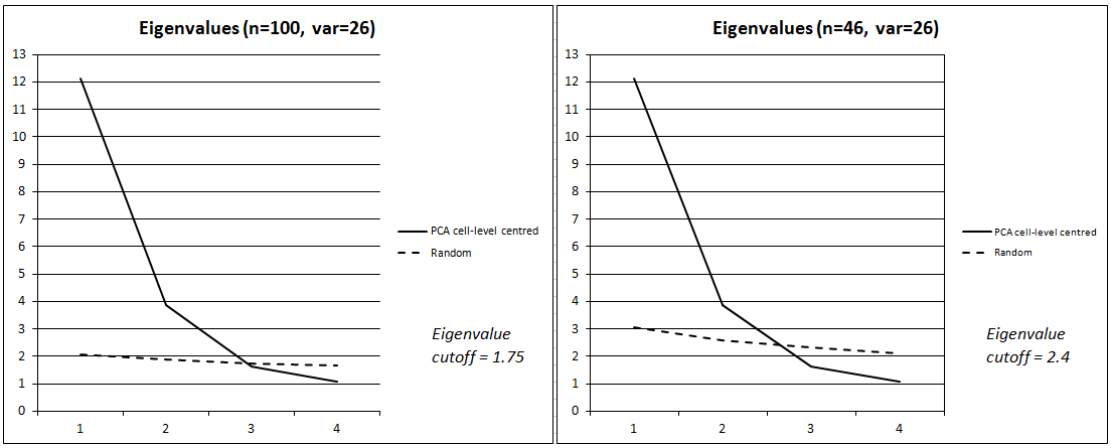


Figure 5.5. *Eigenvalue cut-offs for n=100 and n=44 respectively*

The results yielded a recommended cut-off point of 1.75 for the *n=100* analysis, and 2.4 for the *n=44* analysis (see Figure 5.5). In both of these analyses the cut-off points yielded recommend retaining factors 1 and 2 only, with further factors deemed to be unnecessary and most likely due to over-extraction. This is consistent with the difficulty in interpreting either of these factors and their loading on individual mood words.

5.3.4.4 Factor loadings correlations. The question of how alike the varying projections were was addressed by running Pearson's *r* correlations between the factor loadings for the different analyses. The results revealed extremely high correlations between these different projections for both the valence and activation factors.

Table 5.5. *Factor Loading Correlations (Pearson's* r*), Indirect & Cell-Level (**factor 1**/factor 2)*

	Words (C-l)	Expressions (Ind.)
Words (Ind.)	**.995****/.989**	**.985****/.948**
Expressions (C-l)	**.976****/.910**	**.985****/.974**

**Correlation is significant at the 0.01 level (1-tailed).

The correlation between the factor loadings for the indirect and cell-level analyses were very high, .995 for the valence factor, and .989 for activation for the 'words' projection, and .985 for the valence factor, and .974 for activation for the 'expressions' projection. The correlations between factor loadings for the words and expressions plots were marginally lower, .985 for the valence factor, and .948 for activation for the indirect analysis, and .976 for the valence factor, and .910 for activation for the cell-level analysis.

5.3.4.5 Circular model fit. Inherent to the idea of an affect circumplex is the notion that the data forms a circular pattern, and that consistent with other research examining this structure (Russell, 1980; Watson & Tellegen, 1985; Larsen & Diener, 1992; Remington et al., 2000; Yik et al., 2011) items or externally correlated constructs can be assigned an empirical angle where they can be charted within this factor space. However, as has been noted, when factor loadings are plotted in this way there seems to be a mild discrepancy between the overall loadings for valence and activation respectively, giving a line of best fit that is mildly elliptical rather than perfectly circular. This discrepancy means that a suitable scaling factor must first be introduced to enable a circular best-fit solution where angles can be calculated that are free from this error.

In order to find a suitable scaling ratio and generate a circle of best fit, an approximate scaling factor was first calculated based on a circle of best fit being

plotted by eye onto different scalings of plots (see Figure 5.1b and Figure 5.3b) generated by Excel. The 'by-eye' examination gave a solution of approximately 1.2. Next, the scaling ratio was raised and lowered to examine best fit statistics as they applied to items within the coordinate system given by the factor loadings. For each scaling ratio examined, a circle was plotted centred on the origin of the axes upon which the factor loadings were plottted, with a diameter of the mean vector size. This gave a circle of best fit for a particular scaling factor. From these figures, error, mean square of error, and values for root mean square of error of approximation (RMSEA) were derived to give a statistic of model fit.

Table 5.5a. *Indirect PCA (Words) – Best Fit Circle for Scaling Ratios*

Scaling Ratio	Best-fit Radius	Mean Sq.	RMSEA
1.1	0.965	0.002359	0.04858
1.12	0.970	0.002327	0.04824
1.13	0.972	0.002324	0.04821
1.14	0.974	0.002330	0.04828
1.15	0.976	0.002345	0.04843
1.16	0.978	0.002369	0.04867
1.2	0.988	0.002559	0.05058

Table 5.5b. *Cell-level PCA (Words) – Best Fit Circle for Scaling Ratios*

Scaling Ratio	Best-fit Radius	Mean Sq.	RMSEA
1.1	0.798	0.009242	0.09613
1.12	0.802	0.009189	0.09586
1.14	0.806	0.009164	0.09572
1.15	0.808	0.009161	0.09571
1.16	0.810	0.009165	0.09573
1.18	0.815	0.009192	0.09588
1.2	0.81988	0.009247	0.09620

Of the two levels of analysis, the plot derived from the indirect PCA yielded a better fit, with an RMSEA of <0.0483 indicating a "good" model fit (MacCallum et

al., 1996). The best fit was found using a scaling ratio of 1.13. For the cell-level analysis, the RMSEA was <0.958 indicating a “mediocre” fit with the factor loadings derived by this method. The optimal scaling ratio for this solution was 1.15, a figure that was very close to the one yielded by the better fitting plot.

5.35 Discussion. The findings represented a good validation of the two-factor, valence/activation model of mood in the domain of recognition of facial expression. Furthermore, the use of indirect and cell-level PCA analysis represented a good test of convergent validity of the resulting factor structure. Both methods have their merits and shortcomings, however. The former method – simply running a PCA on the summary table – is one way of examining the form of the data, but collapsing the data in this way yields artificially high measures of variance accounted for by the extracted components. The latter method, on the other hand, enables all of the data to be analysed, allowing individual differences and their contribution to the variance to be more adequately accounted for. However the fact that rows are not fully independent in this latter analysis does mildly compromise this approach because of inequalities in the number of datasets provided by participants. While more advanced techniques such as singular value decomposition (SVD) might be deemed more suitable, such techniques are elaborate and not currently supported by most statistical packages. It was decided that these two approaches combined, and applied to a centred rendition of the data were more than adequate for the purposes of assessing the component structure of the data and establishing a rudimentary coordinate system for the variables under examination. It also provided highly detailed information based on which each mood word could be assessed for its ability to invoke recognisable facial expressions with suitable convergent and discriminant validity.

This study, however, did have some shortcomings. Like many studies, the sample used in this experiment was comprised primarily of university undergraduates, with a relatively low mean age and particularly high proportion of international students who were not native English speakers. This was of some concern, as it was reasoned that people speaking English as a foreign language may have a less nuanced understanding of the mood words on which the faces were being judged. The experiment also fell short of its target of 63 participants, meaning that it may have been underpowered. It was agreed that a larger and more representative sample of English speakers should be sought for the second judgment study.

The main results table (Table 5.0) comprises a correlation matrix displaying sum totals of all responses to all actors posing a particular mood state. With the exception of the 'neutral' term, the rows and columns of the table have been ordered according to the word's theoretical location on the circumplex as reported in other studies (Russell, 1980; Yik et al., 2011). Gradient-shading of the table values enables an approximate diagonal band to be observed which broadly reflects the clarity of correspondence between the mood expression and the corresponding mood words. A circumplex structure of affect predicts a broadly sinusoidal pattern in the responses scores as one progresses across the columns (or down the rows) of such a table.

'Strong' words can therefore be characterised as those which show the clearest pattern, with high values for an expression's corresponding mood word (e.g.: the 'happy' word score for a 'happy' expression), and a steady decline across neighbouring cells until it reaches a pronounced low-point in the cell representing the opposite response. (e.g.: the 'sad' word score for a 'happy' expression). Conversely, items whose facial expressions result in an ambiguous response pattern without such clear peaks and troughs can be considered 'weak' for the purposes of this study,

The resulting patterns of scores could be analysed in a purely qualitative way, however the nature of the predicted response pattern suggests at least one firm metric that can assist in identifying the 'strength' of particular words. The variation in the scores for a particular row offers a useful metric for quantifying the degree of spread of scores given in response to a particular mood expression: Where only a mild fluctuation occurs across the row, it is safe to say that participants were less sure about the affective significance of the expression than if a very sharp and pronounced change in values occurs. Therefore the inclusion of S.D. values offers a basic, quantified measure of this spread that can assist in selecting the most useful terms for further study.

This table therefore offers a detailed assessment reflecting both how those mood words are judged in response to a particular mood expression and how judgements of a particular mood word are made in response to different mood expressions. Examination of row-wise values in the summary table gives an idea of the sensitivity of each mood word in response to a particular facial expression, whereas a column-wise examination shows the specificity of a particular mood word in relation to facial expressions. Together these considerations can allow expressions to be identified which most clearly identify the region of the circumplex in which the corresponding mood word is located.

5.3.6 Selecting a Subset of Stronger Items.

5.3.6.1 Criteria for selection of items. The purpose of this study was to identify weaker mood words whose corresponding expressions were deemed to be least recognisable by the participants as a whole, or which evoked very similar judgements to other, better performing words. These words would be candidates to be

dropped from the current pool of 26 adjectives in order to form a smaller, but representative pool of the better performing items.

The sensitivity and specificity of a particular mood word with respect to its corresponding mood expression, are not the only criteria, however. The plots of factor loadings offer insight into the location and spread of these items about the circumplex that can also be used to guide decisions about which items to retain. To get the most accurate triangulation from scales within the circumplex model, it would be desirable to have candidate mood words which are spread fairly evenly about the circumference of the circumplex. It would also be desirable to have scales that are reasonably symmetrical and complementary, such that one scale would, for example, represent valence, another would represent activation, while the remainders would represent combinations comprising scales running diagonally across the circumplex at various angles.

This of course also raises the question of what form the scales based in this circumplex should take: Should they be a larger number of independent, unipolar scales as in the VAMS, or a smaller number of bipolar ones? Though some evidence suggests that unipolar scales have proven more accurate for a conventional VAS, there is no telling whether a similar advantage exists for the kind of D-VAS that I propose, and the superiority of one format over the other would have to be established by experiment. Stern (1997) argued that bipolar scales may be prone to ambiguity because two constructs – rather than just one – are included. A score of 0 on a bipolar VAS with 'happy' at the top end and 'sad' at the other could therefore mean *either* a lack of happiness, *or* a high degree of sadness. This argument does not seem especially compelling as the meaning of a bipolar scale is so self-evident from the clear juxtaposition of opposites, but it may have some merit in the context of the

cognitive dysfunction that frequently accompanies stroke. He also offers a more clear-cut argument, though, which is that splitting bipolar scales in complementary, unipolar scales doubles the amount of measurements taken, which should therefore improve the accuracy of the score overall.

With so many questions remaining as to the merits of each type of scale, it would be prudent, therefore, to also select items that have a good counterpart item on the opposite side of the circumplex – as close to 180° as possible – such that both unipolar and bipolar scales can be constructed. With these considerations in mind, a target was set to reduce the existing 26 mood words to a pool of 12 items, 3 for each quadrant of the circumplex. This basic framework appeared like a good basis for seeking out a solution that would satisfy these criteria, or at least strike a good compromise between their requirements.

The selection of items was therefore based on three criteria: 1) The variance as reflected in the SD values of the rows of the summary table, 2) qualitative examination of the pattern of correlations in the summary table, and 3) the position of items in relation to other items on the circumplex, as derived from plots of factor-loadings.

5.362 Selection of items. Examining the summary table, two items immediately stand out as having a pattern of correspondences that reflect particularly poor recognition rates. Both the ‘aroused’ and ‘neutral’ items have very low S.D. values, and a pattern of scores indicating that the corresponding images were particularly difficult to recognise. These items were therefore eliminated from the pool.

Examining the charts, the first notable thing is the close clustering of many of the items, particularly in the high activation quadrants of the circumplex. In the

positively valenced quadrant, the items ‘excited’, ‘enthusiastic’, and ‘energetic’ were fairly close to one another, indicating that the corresponding facial expressions were judged in a similar way. Out of these, however, ‘excited’ stood out as a better scoring word, as its S.D. value was higher and it was more often chosen as the word most appropriate for the corresponding expression (mean score of 5.69), though all of these items also scored highly on ‘happy’. The ‘enthusiastic’ and ‘energetic’ items were therefore dropped in favour of ‘excited’. ‘Happy’, was by far the most identifiable expression, with a mean score of 6.22, the highest S.D. score, and a very clear, circumplexical pattern of correspondences. This item was therefore also retained for the next stage.

In the negative valenced, high activation quadrant, a number of items appeared very tightly clustered in a similar region. ‘Tense’, ‘anxious’, ‘nervous, ‘confused’ and ‘distressed’ were clustered close together in both projections of the indirect PCA factor loadings plots. Of these, ‘distressed’ appeared to be the best performer, with the highest S.D. and the clearest pattern of correspondences, and so this was retained while the other four words were dropped. ‘Afraid’ was retained because it appeared – in the primary ‘word’ projection – to be located in a slightly different region to these other terms, – at a steeper angle more in the direction of the activation dimension. ‘Angry’ is unique in that out of the ‘primary emotions’ it does not fit well into a two-factor framework, with some arguing that a third factor (‘dominance’ or ‘potency’) is required to account for it. However the results showed that the item had a fairly high S.D., high sensitivity and a particularly high specificity, with ‘angry’ being rated much more highly than other words in response to an ‘angry’ face, and non-‘angry’ faces scoring much lower than ‘angry’ faces when rated on the ‘angry’ word. Its inclusion in the VAMS and its unique quality also supported including it in the final

pool. It was therefore retained along with ‘distressed’ and ‘afraid’ as the sub-group shortlisted for this quadrant of the circumplex.

On the negative valenced, low activation side, two items stand out as particularly high scorers with respect to S.D. values. Though ‘sleepy’ and ‘tired’ were charted as very close together in the ‘words’ projection, ‘sleepy’ was the second most identifiable expression out of all of them, with a mean score of 6.06. Though the ‘sad’ item did not score a great deal higher than the ‘miserable’, ‘disappointed’ or ‘depressed’ items that were clustered close by, ‘sad’ was the most recognisable of these, with a mean score of 4.68. Since it is the direct antonym of ‘happy’ it would also be a good candidate for a bipolar scale running along the valence dimension. Though ‘bored’ did not have a particularly high recognition score in comparison to the other items in this quadrant, it was located at approximately 45° to the valence and activation axes, and almost opposite its antonym ‘excited’, thereby rendering it of strategic importance in terms of supporting a bipolar diagonal scale. The ‘sad’, ‘miserable’, ‘sleepy’ and ‘bored’ items were therefore retained, and the others discarded. Though it essentially duplicated the position of ‘sad’, ‘miserable’ was also retained to comply with the target of 3 items per quadrant.

Of the items in the high valence, low arousal quadrant ‘calm’, ‘peaceful’ and ‘relaxed’ appeared at almost identical positions at about a 45° angle. Of these items, ‘relaxed’ was the weakest, with the lowest S.D. and a fairly low recognition score. It was also the most frequently misapplied of the three words, with people often rating ‘pleased’ faces than ‘relaxed’ faces as ‘relaxed’. ‘Satisfied and ‘content’ were located quite close to the valence axis at a similar position. Out of these two, ‘satisfied’ had both the higher S.D. and the better recognition score, with a mean of 4.69. ‘Content’ was therefore dropped in favour of ‘satisfied’. This left two remaining items to be

selected for this quadrant, and since ‘relaxed’ was the weaker of the three near-synonymous items, it was discarded in favour of both ‘calm’ and ‘peaceful’. Though these items are located very close together, both are desirable because three items exist in the opposite quadrant, and therefore two different bipolar scales would be possible. This extra bipolar scale at a similar angle may be necessary to accommodate the special case of ‘anger’, whose qualitative difference to neighbouring items like ‘afraid’, ‘anxious’ and ‘tense’ should be given special consideration.

It was felt appropriate to err on the side of caution in retaining items rather than discarding them so that there would be more flexibility in pairing items for bipolar scales. Any superfluous items or scales which they comprise could then be dropped later on as findings dictated. After this process of selection and elimination was completed, 14 items of the original 26 item had been excluded, leaving a new set of just 12 items:

Satisfied	*Afraid*	*Sad*	*Sleepy*
Happy	*Angry*	*Miserable*	*Calm*
Excited	*Distressed*	*Bored*	*Peaceful*

Figure 5.6. *12 Item Subset Selected for Further Analysis*

5.3.6.3 Retesting with the 12 item subset. Having arrived at a much smaller group of stronger performing items, the next step was to rerun the experiment, firstly to check that a similar factor structure and plots of loadings confirm the essential coordinate system yielded by this first experiment, but also – as stated in the original objectives – to allow comparisons to be made between the performance of the actors with respect to the recognisability.

A meaningful comparison was not possible in this experiment because of the low number of datasets per actor image set, however the reduced number of items in this subset has substantial implications for the size of each dataset. Whereas this first experiment consisted of datasets each comprising 676 (26 x 26) responses to a single image set, the reduction of items now means that each dataset now contains only 144 (12 x 12) responses, reducing it almost fifth of its previous size. This means that a repeat experiment in which a similar total number of responses is collected would contain about five times the number of datasets. It was therefore decided that 500 of the smaller datasets would be collected for this next experiment.

Repeating the experiment with a greatly increased number of datasets will accomplish four things. Firstly, it should yield results that are more accurate and generalisable by eliminating sources of variance due to weaker items. Secondly it serves as a good test of convergent validity by essentially reproducing the experiment; though differences are to be expected with this more refined subset of items, the same essential structure should still exist, with both the factor structure and plots of items remaining fairly similar. Thirdly – and most crucially – this higher number of datasets would yield 25 per actor (rather than only 5), enough for meaningful statistical comparisons to be made between the patterns of recognition rates for each actor.

Finally, it was also an opportunity to strengthen the findings of the previous study by addressing shortcomings of the study sample used. As in many studies, the sample used in this experiment it was comprised primarily of university undergraduates, with a particularly high proportion of foreign students who were not native English speakers. Repeating the experiment was also an opportunity to recruit a far more diverse cross-section of the population with a higher proportion of people who speak English as a first language.

5.4 Part 3: Judging the 12-Item Photograph Subset

5.4.1 Sample size. As in Experiment 1, the question of sample size was a difficult one which demanded necessary compromises. Though it would be desirable to have a reasonably sized sample of individual respondents, the focus of this study was upon collecting a high number of datasets per actor. The greatly reduced number of responses per dataset introduced practical issues in terms of having tasks that were of sufficient length to allow incentives to be worthwhile. In Experiment 1, each task involved returning 676 (26^2) responses and took around 40 minutes to complete. A single dataset in this experiment, however, involved only 144 (12^2) responses (12 responses to 12 images of a particular actor). In order to keep task length and incentive amount on par with the previous experiment, tasks were grouped into image sets of five actors each, and – as in the previous experiment – participants were allowed to complete up to three of these. As with Experiment 1, the number of datasets was therefore the focus rather than the number of individual respondents.

5.4.2 Recruitment and participants. In order to obtain a more representative cross-section of English-speaking participants, recruitment of participants was conducted internationally. The study was advertised via online forums and social networking media, and by arranged canvassing in a small number of U.S. and Canadian cities and universities; it was also advertised locally by a small number of advertisements in location about University of Nottingham. As before, a small payment was offered for participation in an experimental task Selection criteria were that participants 1) were at least 18 years of age, and 2) were fluent in English. The data were provided by a total of 64 participants, 38 male and 26 female, aged 18 to 72 years (mean 33 years; S.D. = 14.8 years). Of these participants, 46 (72%) were

ethnically European, 6 (9%) were Asian or East Asian, 1 (2%) were of African descent and 3 (5%) described themselves as mixed race. 53 (83%) spoke English as a first language, while the remaining 11 (17%) spoke English fluently as a second language.

5.4.3 Method. The data collection, as before, was undertaken via the purpose-built web portal. The experiment remained essentially unchanged, except for the decrease in number of words/images being used, and a corresponding increase in the amount of images sets which comprised a single experiment. While in the first experiment each task constituted a set of 26 responses for a single, actor image set (26 photographs), for this second experiment, each task consisted of 12 responses to 5 image sets (each with 12 photographs), which were randomly assigned and ordered by the system.

As before, participants were first required to complete a brief sign-up process in which they set up a login and provided an email contact address. Details of gender, year of birth, ethnicity, and whether they speak English as a first language were collected as part of the sign-up process. Participants who do not speak English as a first language were asked to select their country of origin, and finally, participants were asked to indicate consent by checking boxes for each of five clauses describing terms and conditions of participation, and confirming that they were 18 years of age or above.

The main page for the experiment portal consisted of a list of open experiments (tasks) which had been assigned by the experimenter. Each experiment consisted of 5 sets of images, each containing the 12 mood states posed by that single actor (60 images per task). For the experiment, these images were presented, in turn, on a separate page, accompanied by 12, 7-point Likert scales corresponding to the

mood words used as stimuli for the posed expressions (see Figure 5.7). For each image, the participant rated each of the mood words according to how well they deemed it to agree with the image. Again, both the order of presentation of the images, and the order in which the mood words were listed were randomised, with consecutive pages containing mood words listed in a newly randomised order. As before, participants were free to save their session at any point and return to it later. Participants were initially assigned one experiment, but were offered to complete up to three separate experiments, which they were free to select at random from the list.

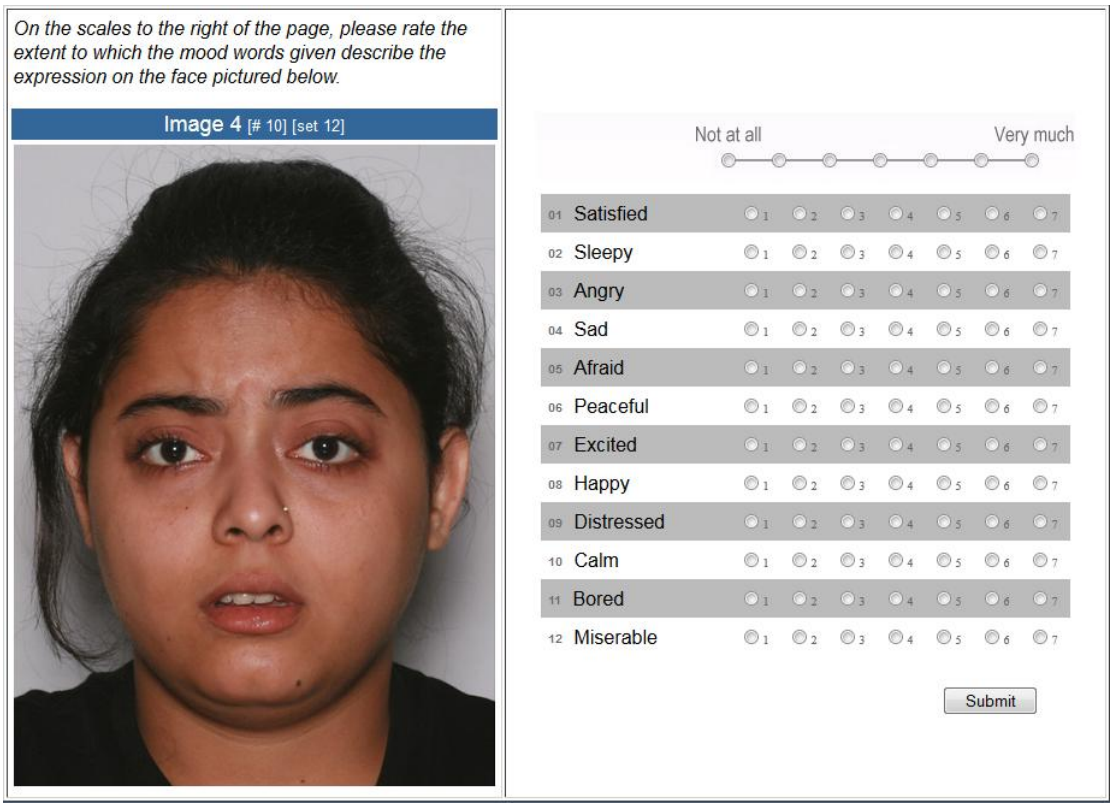


Figure 5.7. *Experiment 2 - Response Page for the Judgement Task from the Project Website*

5.4.4 Analysis. A total of 540 datasets were completed, with each participant completing a mean of 1.7 experiments each (S.D.= 1.11). As before, data was centred to adjust for systematic response biases, and a summary table of means calculated, with mood words ordered according to their theoretical positions on the circumplex.

For the first part of the analysis, a principle component analysis (PCA) was conducted on the table of mean data for all 540 datasets. For this analysis, only *r-type* PCAs (plots of mood words) were performed, as the results from the last judgement experiment supported the view that this projection was appropriate for the methodology employed (single image, multiple word judgements). As before, the PCA analysis was then also performed on the 6,480 rows comprising the cell-level data, with the solutions remaining un-rotated for both analyses. It was predicted that with the weaker items removed, the resulting plots for both levels of PCA analysis would be very similar.

Table 5.6. *Summary of Mean Word Scores(centred) from 540 datasets (Likert score)*

			Total Word Score (by ID)											
Posed Mood	*ID*	*SD*	*1*	3	*10*	*12*	*13*	*14*	*15*	*18*	*20*	*22*	*23*	*24*
Happy	1	1.85	3.7	1.88	-1.4	-1.3	-1.4	-1.4	-1.4	-1.4	-1.3	1.18	0.71	2.23
Excited	3	1.83	3.65	2.81	-1.3	-1.1	-1.3	-1.3	-1.3	-1.3	-1.3	0.66	0.11	1.82
Afraid	10	1.11	-0.9	0.48	2.13	2.1	-0.3	0.05	0.42	-0.8	-1	-0.7	-0.6	-0.8
Distressed	12	0.99	-1.1	-0.5	0.34	2.04	0.41	0.99	1.2	-0.4	-0.7	-0.8	-0.5	-1
Angry	13	1.43	-1.2	-0.2	-0.5	1.8	3.45	0.22	1.09	-0.7	-1	-1	-0.8	-1.1
Sad	14	1.03	-1.2	-1.1	-0.4	1.17	-0.2	2.04	1.36	0.24	-0.2	-0.5	-0.1	-1
Miserable	15	0.98	-1.2	-1	-0.6	1.05	0.2	1.63	1.38	0.57	-0.3	-0.5	-0	-1.1
Bored	18	0.91	-1.1	-1.1	-1	0.32	0.1	0.7	0.82	1.88	0.14	-0.3	0.32	-0.8
Sleepy	20	1.44	-0.8	-1.5	-1.3	-0.9	-1.1	-0.5	-0.7	1.06	3.32	1.15	1.46	-0.3
Peaceful	22	1.31	1.13	-0.6	-1.3	-1.1	-1.2	-0.9	-1.1	-0.1	-0.4	1.89	2.2	1.49
Calm	23	1.2	0.99	-0.7	-1.2	-1	-1.1	-0.8	-1	-0	-0.4	1.7	2.09	1.36
Satisfied	24	1.55	2.4	0.25	-1.4	-1.2	-1.4	-1.2	-1.3	-0.8	-0.9	1.66	1.51	2.35

The data were then charted showing loadings of variables on the extracted components. These plots, accompanied with a qualitative examination of the summary table and the S.D. metrics therein were used to identify and eliminate ‘weaker’ items which either performed poorly as stimuli for recognisable facial expressions, or which were shown to be essentially indistinguishable from other, similar items.

5.4.5 Results.

5.4.5.1 Principle component analyses. As before, the data were subjected to both an indirect and cell-level principle components analysis. The parallel forms method was again used to identify an optimal Eigenvalue cut-off, which for this experiment yielded a threshold of 1.45. Both this calculated threshold and the default Kaiser criterion of 1.0 yielded identical results for both analyses, with two factors being extracted, and factors 1 and 2 corresponding respectively to the predicted valence and activation dimensions.

For the indirect PCA, these two factors accounted for 84.9% of the variance (Table 5.7a). Factor loadings are shown in Table 5.7b. Factor 1, which accounts for 60.2% of the variance, loads proportionally to the negative valence of mood words, while factor 2, which accounts for 24.7% of the variance, corresponds to the level of activation or arousal associated with a given mood word.

For the cell-level PCA, these two factors accounted for 62.2% of the variance (Table 5.8a), Factor loadings of mood words are shown in Table 5.8b. Factor 1, which accounts for 43.4% of the variance, loads proportionally to the negative valence of mood words, while factor 2, which accounts for 18.8% of the variance, corresponds to the level of activation or arousal associated with a given mood word.

The greatly reduced amount of variance accounted for in the cell-level PCA is expected, and the indirect PCA – as was discussed – has the effect of collapsing variance due to individual differences. The comparatively low amount of variance accounted in the latter analysis is understandable given that the analysis was of results from all twenty actors, whose individual performance varied considerably.

Table 5.7a. *Indirect PCA - Words*
Total Variance Explained

Comp-onent	Extraction Sums of Squared Loadings		
	Total	% Variance	Cumulative %
1	7.228	60.232	60.232
2	2.964	24.704	84.937

Table 5.8a. *Cell-level PCA - Words*
Total Variance Explained

Comp-onent	Extraction Sums of Squared Loadings		
	Total	% Variance	Cumulative %
1	5.206	43.386	43.386
2	2.256	18.798	62.184

Table 5.7b. *Indirect PCA - Words*
Factor Loadings

	Component	
	1	2
Satisfied	-.978	-.142
Peaceful	-.918	.340
Happy	-.914	-.365
Calm	-.741	.588
Excited	-.533	-.777
Sleepy	.090	.843
Bored	.399	.839
Afraid	.637	-.417
Angry	.737	-.217
Sad	.894	.163
Distressed	.941	-.291
Miserable	.979	.005

Table 5.8b. *Cell-level PCA - Words*
Factor Loadings

	Component	
	1	2
Satisfied	-.864	-.179
Happy	-.817	-.382
Peaceful	-.799	.260
Calm	-.667	.469
Excited	-.375	-.732
Sleepy	.071	.727
Bored	.251	.696
Afraid	.528	-.329
Angry	.613	-.212
Sad	.728	.122
Distressed	.795	-.271
Miserable	.827	-.021

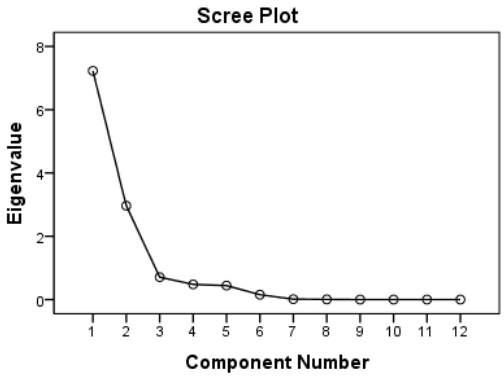


Figure 5.7a. *Indirect PCA - Words. Scree Plot*

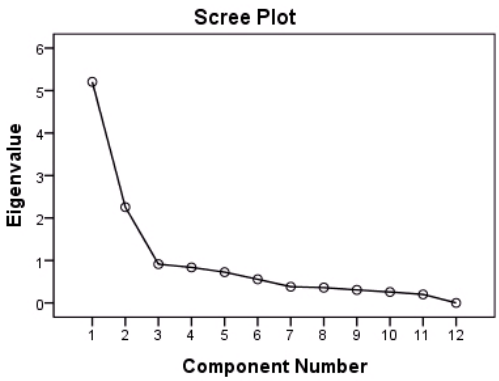


Figure 5.8a. *Cell-level PCA - Words Scree Plot*

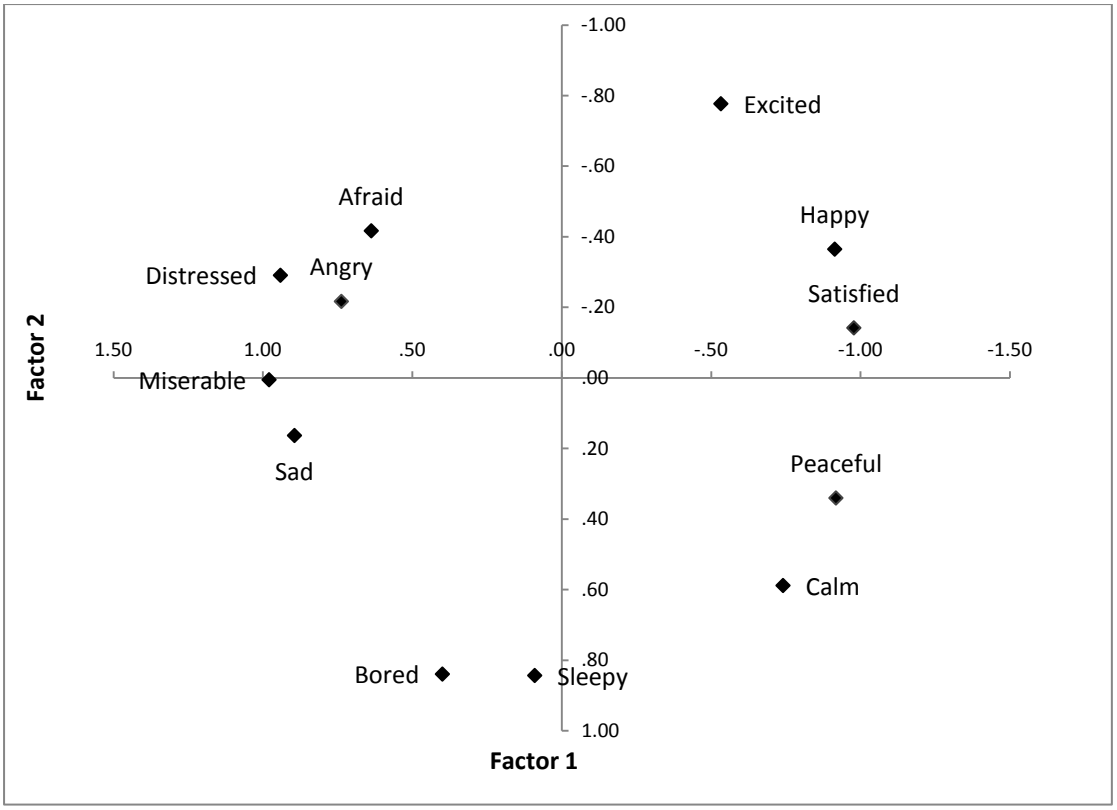


Figure 5.9a. *Indirect PCA - Words. Plot of factor loadings.*

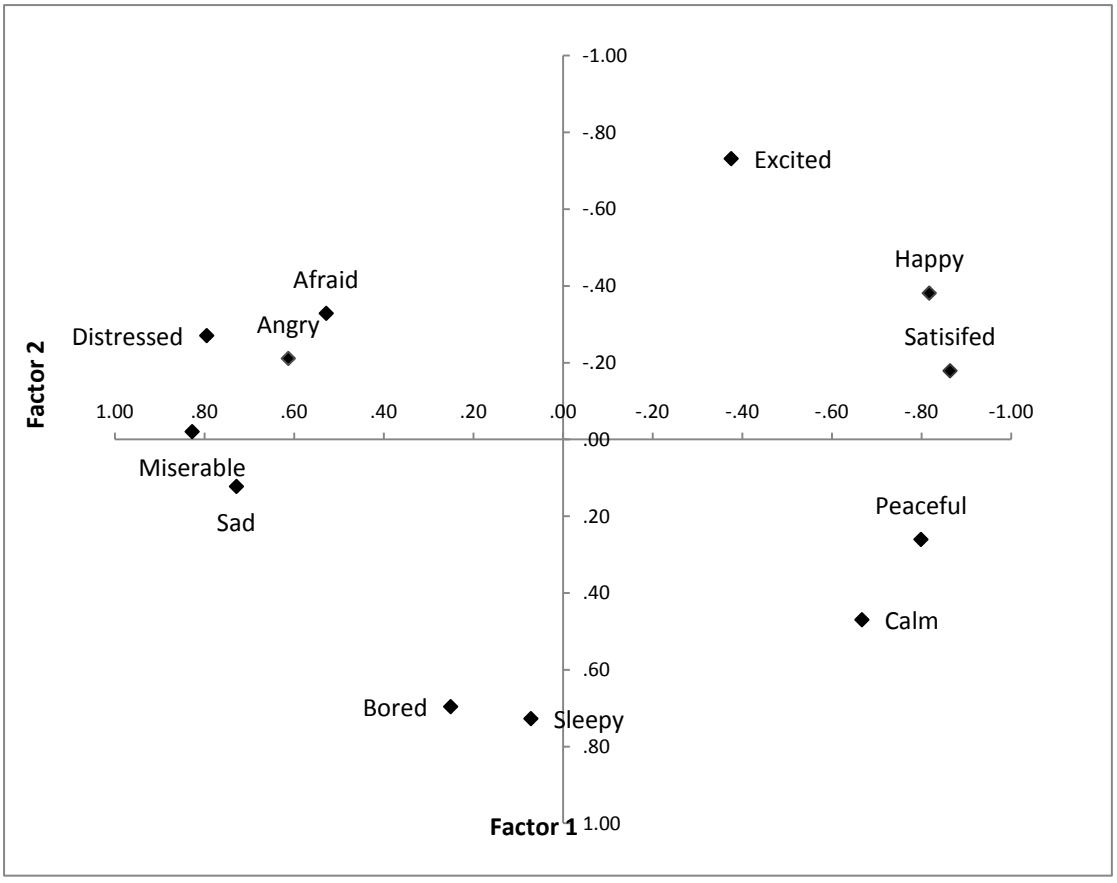


Figure 5.9b. *Cell-level PCA - Words. Plot of factor loadings.*

The plot of factor loadings (Figure 5.9a & Figure 5.9b) are, as predicted, very similar for both levels of analysis. The pattern of the plot is not as precisely circumplexical as it was in the previous experiment, though this is to be expected with the greatly reduced number of variables in this analysis.

5.4.5.2 Factor loadings correlations. This second experiment therefore gave us an opportunity to examine convergence of judgement data between two completely independent samples. The clear similarity of the plots is self-evident, however they can again be examined quantitatively by running Pearson *r* correlations of the factor loadings of words between the two studies. In order to do this, the factor loadings for valence and activation were correlated between the two plots for both the direct and indirect PCA analyses. Because Experiment 2 only had 12 of the original items, factor loadings for the other items were dropped for this analysis so that only the 12-items including in both studies were examined.

Looking at Table 5.8, it is clear that a near identical coordinate system was derived from the independent samples, which provides compelling convergent evidence that these words are, by and large, identified as being located at very similar positions about the circumplex.

Table 5.8. *Experiment #1/#2 Factor Loading Correlations for 'words' PCA (Pearson's* r*)*

	Valence	Activation
Indirect	.983	.982
Cell-level	.976	.975

Significant at the 0.01 level (1-tailed).

5.453 Discussion. The results of the judgement study for this subset of mood words essentially duplicates the findings of the previous experiment, further supporting a two-factor, valence-activation model. Furthermore, the very high correlations between the factor loadings from data from two independent samples is strong evidence of the reliability of the circumplex structure and the affect words mapped within it. The larger, more representative sample of English speakers used in this sample only strengthens these findings.

The results of these studies therefore not only strongly support the CMA, but demonstrate that the same affective space exists in judgments of facial expressions as it does with more traditional word judgement studies of similarity judgments of self-reported mood. Not only do judgements of facial expression reveal the same structure of affect, but the reliability with which they do so suggests strongly that they can be suitable basis for a coordinate system through which mood can be measured, and that facial expressions – and not just words – can be used to construct valid and reliable mood scales.

5.5 Construction of the Prototype Scales within the CMA

The judgement study of the 12-Item subset enabled the CMA to be retested and confirmed in a subset of suitable words which would be candidates for the final scales. They also provided plots and factor loadings which closely agreed with those of the original 26-item set in Experiment 1 (see Table 5.8). The words plot for this latter analysis would form the basis for constructing a number of bipolar scales that cross this factor space in a reasonably natural way, such that words are paired to form endpoints of scales which are approximately bipolar.

It's important at this point to recall the significance of the two different types of plot, the 'words' and 'expressions' plot. In the former plot, people's judgements of words are treated as the primary variables, and these are plotted in factor space that purely reflects how judgements fall in 'semantic space'. Though the words used to stimulate the expression all correspond to posed expressions, it is important to realise that the specific 'identity' of the facial expressions is not relevant. I could have a set of poses relating to expressions posed in response to a different – but representative – sample of mood words, and the word judgements should still fall in a similar pattern as they both delineate the same 'space'.

The alternate projection of this data is where we use the faces as the variables, and examine their position within the factor space. The distinction between these two projections is subtle but important: While PCA factor loadings plots for the former tell us where people place these words in 'semantic space', those of the latter map out where people place the faces posed as a result. However, because the posed expressions are imperfect reflections of the words posed for, we would expect the 'faces' projections to be weaker as, in a manner of speaking, between the words and the posed expression, the 'signal' has lost some fidelity in the transmission. This is exactly what we saw in the *part 2* judgement experiment: the variance accounted for in the word PCA analyses was substantially greater than those for the cell-level analysis.

As we saw in experiment 1, the 'faces' projections accounted for somewhat less of the variance than the 'words' projection, and for this stage it is this latter projection that should be used to guide the prospective mood scales across this space. However, for the final scales, a plot will be required specifically for the faces of the

‘top actors’ so that measurement using the scales of these images as reference points can be properly resolved into their respective components (valence and activation).

In guiding values for our coordinate system and the prospective scales within it, we can see how some of our mood words fall fairly naturally into approximately bipolar scales crossing this space, as envisaged. Beginning with horizontal scales and working towards the vertical, two predominately valence based scales are immediately evident. ‘Miserable’ and ‘satisfied’ form a near-horizontal scale comprising almost purely valence, with both of the words loading only by a very small amount on the activation dimension. ‘Sad’ and ‘happy’ form another valence-based scale, but this time at an incline, with ‘sad’ being slightly low activation, and ‘happy’ being moderately high activation, in line with plots from other studies (Russell, 1980; Yik et al., 2011).

‘Distressed’ and ‘angry’ both appear to form a counterpoint with the ‘peaceful’ item. The qualitative difference between these terms can be put down to the loading of ‘angry’ on a third, but less prominent dimension Dominance-Submissiveness which is present in complex emotions. In fact, one might view ‘angry’, ‘distressed’, and ‘afraid’ as being single arc across this dimension, with anger and fear representing dominance and submissiveness respectively, and with distress being a neutral point. Since there is no reason in principle why two scales cannot share a counterpoint, and the constraints of the items chosen seems warrant this, then two more scales *Distressed-Peaceful* and *Angry-Peaceful* seem to suggest themselves. The presence of ‘Afraid’ and ‘Calm’ as endpoints of a scale at a steeper inclination support this, and the surplus of high activation, low valence items mean that a shared endpoint in the opposite quadrant (low activation, positive valence) is needed.

At right angles to these last two scales, *Bored-Excited* presents itself as an obvious antithesis. However when it comes to producing a primarily activation-only scale, a problem presents itself: the 'aroused' item, which has been charted as a high activation, neutral valence in Russell's (1980; Yik et al., 2011). CMA, was located quite differently in the present study, perhaps due to a shift in the connotations the word and its use in language. This leaves us with no counterpoint to the 'sleepy' item that can give us a primarily activation-based scale.

However, it must be remembered that these scales will eventually be represented by posed facial expressions following trajectories from one endpoint to the other. It is therefore not essential that we have a precisely plotted word to represent the quality we want from the facial expression, we just have to explain what we mean in relation to this structure, a state of high activation but neutral valence, – something akin to surprise or alertness. Provisionally, though, this item will need to be called something, so as a convenient term of reference it will be labelled 'alert', which is reasonably close to the concept of psychological activation. The final scale, then, will be called *Sleepy-Alert*.

These seven scales are charted in Figure 5.10. Correlations between their respective endpoints based on PCA analyses from Experiments 1 & 2 are shown in Table 5.9, and attest to their bipolarity. Together, these form the essential structure of the DVAS-based scales envisaged herein, addressing the limitations of earlier attempts as described in 3.5. Unlike the original, paper-based VAMS, *Dynamic Visual Analogue Mood Scales* (D-VAMS) would be based on actual human faces. They would comprise an interactive system whereby images of facial expressions could be modulated to reflect points along a given scale. For every point on each DVAS a separate image would exist, allowing a person to receive feedback on what a

particular magnitude of a particular affect type 'looks like', without the need for the vagaries of scaling these notions mentally through a nebulous process of cognitive interpolation.

Table 5.9. *Correlations between Scale End-points from Indirect PCAs*

Scale	Pole 1	Pole 2	Exp 1	Exp 2
1	Miserable	Satisfied	-.95	-.94
2	Sad	Happy	-.89	-.82
3	Distressed	Peaceful	-.91	-.93
4	Bored	Excited	-.83	-.8
5	Afraid	Calm	-.7	-.62
6	Angry	Peaceful	-.75	-.73
7	Sleepy	[Alert]	-	-

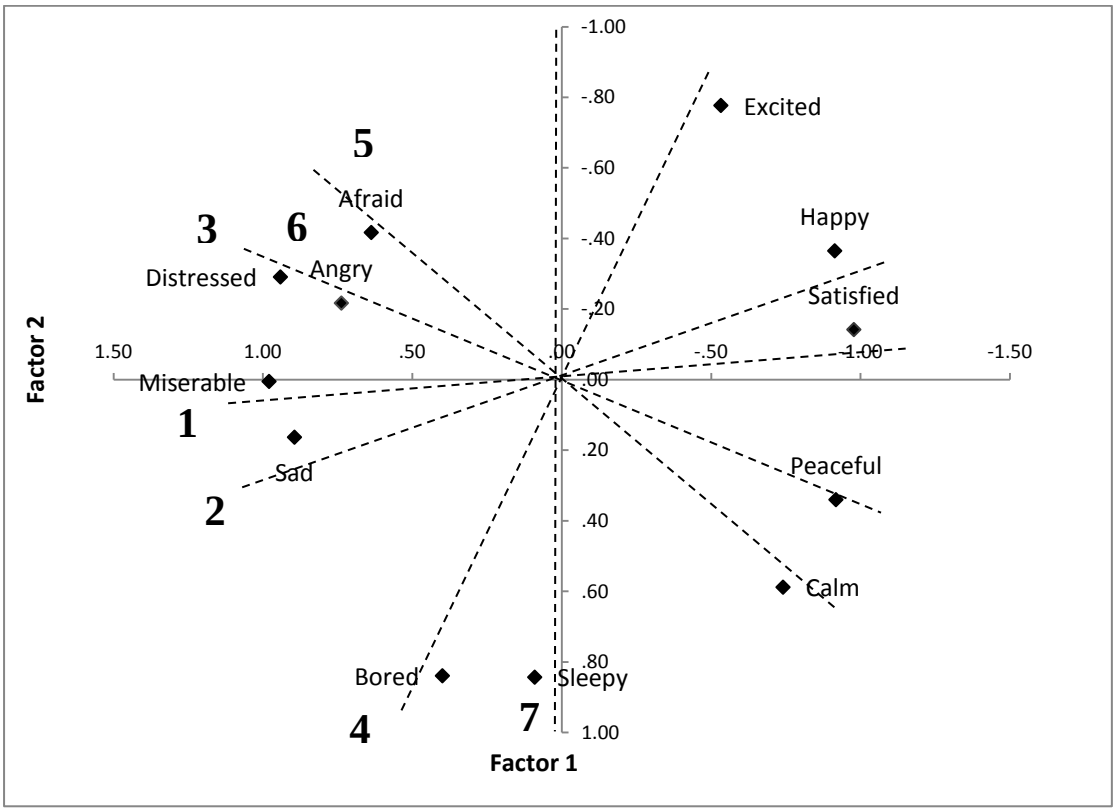


Figure 5.10. *Seven Prospective Scales assembled from the 12-Item Mood Word Subset*

Furthermore, as well as being separate measures tapping the many of the key facets of mood and emotion (as with categorically based accounts), each constituent scale would also exist as mathematically interconnected components in an underlying, CMA-derived coordinate system, through which they can be viewed as summed proportions of valence and activation. In this way, the power of multiple item responses tapping a common construct can be harnessed to give more fidelity and reliability to overall scores.

Finally, The D-VAMS would be of a form that could be run on most computers and handheld devices, taking advantage of the increasingly ubiquitous presence of portable, touch-screen devices, and their inevitable adoption as a routine part of medical assessments

Having arrived at a suitable theoretical model and a prototype for D-VAMS and its constituent scales, though, there remains the significant task of collecting images of posed facial expressions traversing each scale, and generating many frames of such images at a highly granular level. There also remains there task of choosing the 'high-performing' actors who would pose these expressions. This involves tackling the difficult issue of what criteria should be used to quantify the overall recognisability of images provided by our candidates, so that our actors can be ranked accordingly. It is to these matters that discussion will now turn.

6. Study 2: Creating Reference Image Sets for the Scales

The first study fulfilled a number of objectives. The first was to select and test a variety of mood words for their ability to evoke recognisable facial expression, and to eliminate poorer performing words by way of recognisability of corresponding facial expressions. The second objective was to confirm the CMA, and to verify that 'affect space' can be described by two dimensions, *valence* and *activation*, which would be suitable for use as a coordinate system within which to base bipolar scales of images of facial expression. The data from Experiment 1 achieved both of these objectives, confirming the CMA and giving data that enabled weaker mood words to be dropped, and a smaller pool of 12 candidate descriptors to be identified and retained for further study.

Finally, having established a smaller pool of candidate items, the experiment was duplicated using this selected sub-group of items and their corresponding images only. One objective of this (Experiment 2) was to refine the coordinate system and offer independent validation of the CMA in an independent sample, however this second judgement study was also necessary in order to produce sufficient data to enable the identification of actors who are particularly skilled at posing recognisable facial expressions, and who would be selected to produce further images that would form the prototype scales.

This data provided by Experiment 2 essentially corroborated that of Experiment 1, and provided a coordinate system within which to scales based on facial expressions. The results enabled key bipolar scales to be identified which would

form the basis of a set of prototype scales the *Dynamic Visual Analogue Mood Scales* (D-VAMS). These scales are: 1) *Miserable-Satisfied*, 2) *Sad-Happy*, 3) *Distressed-Peaceful*, 4) *Bored-Excited*, 5) *Afraid-Calm*, 6) *Angry-Peaceful* and 7) *Sleepy-Alert*.

Having established a set of prototype scales and a suitable coordinate system, it was next necessary to select actors for the prototype scales, and then to recall them to provide a series of new images. In the first photographic sitting, only the endpoint images of what were to become the prototype bipolar scales were available. The concept behind D-VAMS, however, called for a series of transitional images that would enable the facial expression at one end-point to gradually change into the one at its opposite end. This would call for a recall of suitable, high-scoring actors (by way of overall 'recognisability' of their facial expressions) to pose not only the endpoint images of the 7 scales, but also a series of facial expressions reflecting gradual transitions between one expression and the other. These new expressions would be experimentally located on their respective scales by a new judgement experiment, with morphing software used to generate the extra images required to create the appearance of a seamless transition between them.

This, however, called for actors judged to be better performing in terms of the overall 'recognisability' of their posed facial expression to be selected from the initial pool, and this, in turn, required a broad metric by which the performance of their images could be ranked.

6.1 Ranking the Actors by Recognition Task Performance

Crucially, though, the last study also offered a rationale for quantifying the 'recognisability' of facial expressions. Using mean S.D. scores as an indirect measure of the convergent and discriminant validity of particular mood words offered a key

index by which these words could be assessed and either retained or eliminated. In conjunction with examination of the correlation matrix of responses, they provided a satisfactory – though imperfect – way of quantifying the degree to which a word or actor image demonstrates a pattern of convergent and divergent values that reflect the strength with which judges associate an image with a given word.

6.1.1 Assessing actors' performance. Having removed 14 of the original 26 items, the remaining pool was small enough to allow a much greater number of participants to judge each of the 20 actors. For the initial pool of 26 items, only 5 participants judged each of the actors, which is not nearly enough to make meaningful comparisons between scores given to the mood images between separate actors. Experiment 2, however, which was less burdened by way of number of mood words, provided a total of 27 datasets per actor, allowing for meaningful statistical comparison between them.

This larger number was crucial to enable such meaningful comparisons to be made. However in order to assess the performance of actors in terms of the recognisability of their poses, we needed a basic metric to rate them. This metric was one that has already been used during the item selection process. When the summary tables were being examined to identify which items to dismiss and which to retain, two key methods were used to identify those words which yielded the most recognisable facial expressions. First, the summary matrix was examined in a qualitative way to identify which items showed the most pronounced peaks and troughs in responses in responses to words that were like or unlike those used to stimulate the depicted facial expressions. These patterns were also examined on an item by item basis to see how precisely the correspondences appeared to favour them in terms of their convergent and discriminant validity.

To guide this looser form of examination though, a statistic was introduced that in theory offers a more concrete index of the 'recognisability' of a particular item. In the context of the type of sine-wave pattern that would follow logically from the circumplexical nature of these type of data, the standard deviation of values across the peaks and troughs of responses to a particular mood image are a reasonable index of strength of recognisability of a particular item. An item whose values do not rise distinctively at one particular point on the circumference of the circumplex, and then fall again as they approach their bipolar opposite counterpart can be said to be less 'strong' than one which shows a sharp rise at a specific region of the circumplex, and corresponding decline at its antipode.

This same metric however, can also act as indicator of the success with which a particular actor was able to convey the mood word in question. The same method that enabled the 'strength' of mood words to be judged can also be used to assess the overall 'strength' of actors by way of their ability to pose expressions that are recognisable by participant judges as a whole. In Experiment 2, participants provided judgements of each actor image from which an S.D. value was computed. Differences between actors of S.D. values for these images can be seen as differences of 'strength' of the actor as a mediator of facial expressions of these mood states, and these values can be used as a basis of comparison between actors, while the mean of the S.D values for all of a given actor's images offer a statistic by which actors can be ranked.

6.1.2 Comparison of means. Though this method would seem to be a reasonable way of ranking the actors in terms of the recognisability of their facial expressions with respect to corresponding mood words, the statistic of mean S.D. must obviously be treated with some caution. However, examination of summary tables broken out by actor offers evidence of its utility. Those ranking high overall

tended to get the most consistently high S.D. scores for each of their posed expressions while those ranking lowest seem also to get the most uniformly poor S.D. scores across the posed expressions.

However in a ranking system of this kind there is the important issue of the statistical significance with respect to differences in scores of actors adjacent to one another in their assigned order. The rankings only have meaning insofar as they reflect statistically significant differences between the scores of one actor and those of another, but this is something that can easily be tested. By comparing the S.D. values of responses of all 27 respondent judges to the images of each actor, with the corresponding values given to those of an actor above or below them in the rankings, a simple one-tailed comparison of means between scores for different actors can shed light on the reliability of the assigned rankings.

Table 6.1. *Matrix of significance values for related, one-tailed T-Tests on ranked, top scoring actors.*

Rank	1 (A13)	2 (A17)	3 (A14)	4 (A04)	5 (A12)	6 (A07)	7 (A10)	8 (A16)	9 (A20)	10 (A01)
1	1	.284	**.017**	***.001***	**.019**	***.000***	***.000***	***.000***	***.000***	***.000***
2		1	.082	**.012**	.066	***.001***	***.001***	***.000***	***.000***	***.000***
3			1	.236	.493	**.039**	**.022**	***.000***	**.003**	***.000***
4				1	.226	.111	.097	**.004**	**.006**	**.002**
5					1	**.031**	**.027**	***.000***	***.001***	***.001***
6						1	.443	.051	.151	.080
7							1	.076	.206	.094
8								1	.238	.400
9									1	.353

p<.05, ***p<.01***

To examine this, S.D. scores of the top ten actors, as defined by this metric, were subjected to a comparison of means test with the scores of every other actor, in the form of a related, one-tailed T-test. The hypothesis, naturally, was that the scores

for each actor would be significantly higher than those of any actors below themselves in the rankings. The resulting significance matrix is charted in Table 6.1.

Reading across the table, a pattern of p-values is clearly evident which supports the essential veracity of the assigned rankings. Generally, S.D. values for an actor assigned one rank were significantly greater than those of actors in ranks below it, with the differences being in the predicted direction. Neighbouring ranks, however, did not always show a significant difference in S.D. scores. The number #1 ranking actor did not have significantly greater scores than the #2 ranking actor for example, but *did* have significantly greater scores than the #3 ranking actor; while the number #2 ranking actor did not have significantly greater scores than the #3 ranking actor, but *did* have significantly greater scores than the #4 ranking actor. Though the number of datasets per actor based on which to make a comparison for Experiment 2 (*n=27*) was a great improvement on that of the first experiment, it probably left the statistical test slightly underpowered. If the constraints of the study had permitted a greater number of datasets per actor, then in all likelihood differences between consecutively ranked actors would have reached statistical significance.

This shortcoming aside, though, the ranking undeniably reflects a clear and strong effect by way of statistically significant differences through the descending ranks, and serves the purpose of narrowing down the prospective actors for the prototype scales.

6.1.3 Correlation of actor rankings. Though the data from Experiment 1 was of itself unsuitable for making a reliable assessment of actor rankings, further supporting evidence might still be found by examining the results of Experiment 1 for a similar pattern. If two independent samples of people were to return a similar

pattern of ratings, then this would offer strong supporting evidence of the calculated ranks. In more precise terms, if the mean actor ratings given by participants in one sample were to correlate with the mean actor ratings from another, then this would offer valuable convergent validity.

In both of these experiments, there was a trade-off between the number of actors, the number of mood words that could be examined, and the number of respondents judging the expressions corresponding to words presented to each actor. For both of these, it was necessary to use reasonable number of actors so as to offer a good chance of discovering individuals who were particularly good at portraying recognisable facial expressions.

In Experiment 2, the reduced number of 12 mood words meant that 27 respondent datasets could be returned for each of the 20 actor images sets. However, in Experiment 1 it was necessary to examine a much larger, initial pool of 26 mood words, with a correspondingly greater number of images per image set. The resulting, smaller number of 100 datasets meant that only 5 datasets were returned for each of the 20 actors, not nearly enough to make a meaningful comparison between ratings of actors. However by comparing the 12-item data from Experiment 2 with the corresponding data (12 of the 26 image/word response sets) from Experiment 1, it would be possible to use the same method to produce side-by-side actor rankings for data from both of the independent samples. A correlation between the two can be calculated.

Similar S.D. metrics were therefore calculated from mean responses to the 12 items, for the 5 datasets provided in by each actor in Experiment 1 (see Table 6.2). Though it was not expected that a significant correlation would be obtained from data arising from such a small sample size, the strength, direction and significance of the

resulting statistic might at least offer tentative support for convergent validity in these rankings.

However, even with such an underpowered test, a small correlation of 0.299 (Person's *r*, one-tailed) was found between the mean S.D scores of the actors from the 26 x 26 image/word judgement study, and those from the 12 x 12 image/word judgement study. Though the correlation fell short of statistical significance (*p=.100*), the convergence of S.D. scores from two separate studies using independent samples offers further reassurance that the ranking of the actors determined by this method has an objective basis.

Table 6.2. *Actor rankings and mean S.D. values for Experiment 2, with corresponding mean, 12-item S.D. values for Experiment 1*

Rank	*ID*	*Exp 2*	*Exp 1*
1	A013	1.68	1.46
2	A017	1.58	1.50
3	A014	1.56	1.42
4	A004	1.55	1.60
5	A012	1.52	1.50
6	A007	1.46	1.23
7	A010	1.45	1.26
8	A016	1.42	1.31
9	A020	1.41	1.64
10	A001	1.41	1.48
11	A002	1.40	1.51
12	A015	1.36	1.46
13	A005	1.32	1.22
14	A011	1.32	1.34
15	A018	1.32	1.58
16	A009	1.31	1.51
17	A003	1.31	1.19
18	A006	1.27	1.09
19	A008	1.24	1.17
20	A019	1.20	1.58

6.2 Part 1: Producing Transitional Scale Images

In order to realise the vision of a dynamic version of a VAS (DVAS), where an image dynamically changes in response the position of a slider, we must first address the question of its granularity. In order to allow as free and nuanced a response to the scale as possible, there must be a sufficient number of images so as to allow an almost seamless transition from one end of the scale to the other. Since the traditional form of a VAS has always been of the form of a score in the range of 0-100, as measured by the position of a mark along a 100mm line, it makes sense for a DVAS to adopt a similar level of granularity.

One option for achieving this level of resolution would be to record a video clip of the actors changing their expression from one endpoint expression to the other, and then to key the display of frames of this video to particular points of a slider.

There are problems with this approach, however. Firstly, the quality of still-frames extracted from video are notably poor. It is important to have images that are as high quality and resolution as possible, so that all the detailed nuances of facial expression and the underlying musculature are captured. Ordinary, still-frame photographic images from a good camera offer image quality that is far superior to still-frames from any affordable video equipment. Secondly, we have already established that in order to create a scale that is at least approximately interval level, the positions of transitional images along a scale would be decided by normative data from where participant judges placed them in an experiment. It therefore made sense to start from a series of transitional, high-quality images, and then use morphing software to generate the additional images required by a 0-100 point VAS. This would offer the highest quality images and allow for precise calibration of the scales, giving the smoothest and most natural animation from one end of a scale to the other.

As well as providing a new set of images for the final, prototype scales, though, these would also be used for a refined coordinate system specific to the chosen actors.

6.2.1 Recall of highest scoring actors for photographic sitting. Having arrived at an appropriate method of ranking the actors and their respective mood images, it was next necessary to try and recall the best actors to provide more images that would enable the final scales to be assembled. Out of the top four actors, only two were available to take part: the actors designated A014, and A017, who were ranked 2nd and 3rd respectively. These actors – designated #1 and #2 respectively – were recalled to pose a set of photographs of varying intensities of expression along the seven scales selected for further development.

This second photographic sitting would also serve to fill in a notable gap with respect to current set of images. If you recall from the PCA plots of Study 1, there was a notable absence of any mood words or corresponding images which could act as an opposite endpoint to the 'tired' item. In every plot, the high activation end is sparsely populated, with no word/image that represents a high activation, neutral valence item that can serve this purpose. The word 'aroused' was initially intended as the marker for high activation, neutral valence, but with hindsight it was a poor choice of word, as its meaning in a technical, psychological context is at odds with its significance in common parlance. In everyday language the word has connotations that render it positively valenced, as is attested to by its position within the CMA plots. This second photographic sitting would also provide an opportunity to plug this gap by obtaining suitable, high activation, neutrally valenced images which could be used to complete the *Sleepy-Alert* scale.

6.2.1.1 Continuity considerations. For both of these actors, over a year had elapsed since they had posed for the mood words in Study 1, part 1. Because the expressions of these actors were validated by the previous data as well-recognised endpoints to these scales, it had initially been anticipated that endpoint images for the first study could be blended with those collected in this phase of the study, as it did not seem likely that a few months would make much difference to a person's facial appearance. However, when the actors appeared for their second sitting, there were clear and notable differences that were evident as the previous photographs representing the endpoints of the scales were displayed. The male had significantly lost weight, and his face was noticeably leaner, while the female had acquired a mild suntan which stood in contrast to her paler skin-tone in the previous year's sittings (even though October was chosen to avoid this issue).

For the purposes of producing a natural, linear scale in which images flowed seamlessly from one to the other from scale endpoint to the other, it was essential to minimise any discrepancy between the endpoints images and the transitional ones which were to be taken in the second sitting. In addition to minor differences in physical appearance, other factors would also come into play here. Even the closest attention to technical detail in the setup of the camera equipment cannot fully control for subtle differences in lighting conditions, or for minor variations in vantage point, and angle relating to the posture of the actor and the tilt, turn and lean of their head in relation to the camera.

Since the coordinate system for the endpoint images for these top-scoring actors needed to be individually – and more accurately – charted in a further experiment, however, it was not essential to retain the original images from the first sitting. Though using the new endpoint images would mean that data from the

previous judgement studies could no longer contribute towards the data for coordinate systems specific to the two actors, the continuity issues meant that it was better to start from scratch and collect both endpoint and transitional images for each of the scales. To reduce the impact of replacing the endpoint images, the actors would be guided to simply reproduce as closely as possible the facial expressions that they had posed in the original sitting.

6.2.1.2 Method. Approval for this study, including the judgement tasks detailed in 6.3 and 6.4 was granted by the Faculty of Medicine and Health Sciences Research Ethics Committee, University of Nottingham (Ref : I10102013).

The actors' second sitting was fairly similar to the first sitting (see 5.2), but with some key modifications. Prior to the sitting, the two recalled actors were given a brief explanation of the circumplex model that was being used, and the mapping of the seven prototype scales in relation to the structure, and provided with an information sheet (Appendix VI) and an instruction sheet (Appendix VII) explaining the task. As before, the actor was sat at a table approximately two metres in front of the camera equipment. They were provided with a mirror which they could use to guide the appearance of their poses of facial expression, and there was also a monitor to the side of the camera, which was used to display the actors' endpoint expressions from their previous sitting. For each of the scales, the actor was shown their expressions from the previous sitting which would denote the two endpoints for the scale for which they were to pose. For the 'Sad-Happy' scale, for example, they were shown the expressions which they had posed in their previous sitting in response to the words "sad" and "happy".

One of the endpoints of a given scale was then chosen to begin from, and the actor was instructed to begin by reproducing this expression as closely as possible.

Photographs were taken, and the actor was then asked to slowly modulate the degree of intensity of the expression so that it faded gradually towards a 'neutral' expression, with a sequence of photographs being taken as they did this. As their expression approached the 'neutral' point, they were then asked to slowly begin to express the bipolar opposite expression – again using their photograph from the previous sitting as a guide – gradually intensifying the expression to its fullest extent. Again, photographs were taken as they did this. This process was then repeated in reverse, starting with a pose for the latter, opposite expression, and returning slowly to the starting point expression, all the while have multiple photographs taken of the transitional expressions. In order to capture as clear as possible a transformation from one pose to another, actors were sometimes guided in the expression of particular, fine feature transitions, such as how open or closed their mouths or eyes were, how bared the teeth, how raised the eyebrows, or the extent of a scowl or a snarl.

This process was repeated as often as necessary to capture as smooth as possible a set of transitional facial expressions for each scale, before proceeding to the next. The order in which endpoints and neutral states for the scales were posed, and the direction in which facial expressions were modulated varied from one run to the next, with the direction of expression changes being reversed or repeated as necessary to accommodate individual performance on the scale being posed.

Particular emphasis was placed on recapturing as accurately as possible the original endpoint expressions; these were the images based on which the actors scored well on the metrics used to rank them, and therefore reproducing them as closely as possible was a particular priority. Where necessary, actors were given specific instructions on what aspects of their expression to modulate so as to offer as close an imitation as possible.

There was also the matter of the ‘gap’ in the circumplex, where no image was charted which could act as a polar opposite of ‘sleepy’ for the *Sleepy-Alert* scale. For this image, the actors were instructed on the type of image that was required, bearing in mind the briefing. The expression was described as something akin to ‘alert’ or ‘surprised’, – highly activated like ‘excited’ or ‘afraid’, but neutrally valenced. It was emphasised that it was important to produce an expression that was as neutrally valenced, but as highly activated as possible; if the expression were slightly negative, it was liable to be interpreted as terror, whereas a positively biased image would look more like elation or joy. Special attention was therefore given to capturing a good sample of such images, and of their transitions along the *Sleepy-Alert* scale.

6.2.1.3 Selecting candidate scale images. Before selecting candidate images for the scales, the entire pool was first examined for quality control; any that were obviously defective were eliminated from the pool and permanently deleted. These photographs included blanks (where the flash had failed to trigger), exposure test photographs, and any resulting from other technical problems that rendered the images of inadequate quality. After this, a total of 209 (mean 29.9 per scale) and 283 (mean 40.4 per scale) photographs remained for the female and male actors (#1 and #2) respectively.

6.2.2 Ordering the images into sequences. Of the pool of images remaining for the new series many endpoint images existed that closely reproduced the original endpoint expressions. The only notable exception to this was the ‘distressed’ pose for Actor #1, which was not as intense as the one originally posed. This was rectified by using Photoshop to ‘transplant’ the face from the original ‘distressed’ pose onto the best-fit, endpoint image from the new image set, thus creating a new, manufactured

endpoint image. This acted as a replacement for existing image, which then acted as its neighbour a little further along the *Distressed-Peaceful* scale.

This pool also, of course, contained numerous runs of transitional images covering the same scales, and some of the runs were better than others. Some of the runs appeared to form good continua for one part of the scale, but not for other parts of the scale. Furthermore, different runs would consist of poses that were subtly different, with turns or tilts of the head throughout which limited their ability to be mixed with other sequences.

The task of assembling these images into a single set per scale was therefore fairly complex. In some cases, a single run of consecutively taken photos could be used from end to end, while in others, two or more combined runs were sometimes needed to cover the full span of the scale. By a process of selection and elimination, images were sorted into a smaller number of candidate images which formed a sequence that was as smooth as possible.

Selection of images was not always guided purely by issues of continuity, however, but by elements that held theoretically importance from the point of view of the underlying CMA. For the *Sleepy-Alert* scale, for example, particular attention was paid to identifying runs of images which were as neutrally valenced as possible, This was of particular importance for the 'alert' end of the scale, as the higher the level of activation, the stronger appeared the tendency for an expression to slip in a positive or negative direction.

Once this selection process had been completed for each of the scales, sets of reference images of varying number remained for each of them; these sets were something of a patchwork, but assembled into the form of a reasonably tidy and smooth succession of images spanning the length of each respective scale. Most of the

images from these runs would form the keyframe images based on which the final scales would be generated (see Appendix VIII & Appendix IX).

6.2.3 Re-mastering and equalising the images. Like their chemical-film predecessors, modern, professional, digital cameras also have the digital equivalent of a 'negative' file, which gives a far more detailed and nuanced level of information about the image at the time the photograph was taken. The JPEG files generated by the camera are a rendition of this RAW information in much the same ways that chemically developed photographs are a rendition of the negative, but modern cameras also allow you to manually create optimal settings for the type of image you want, and the quality of a JPEG file manually rendered from its RAW file is frequently better than the one using the camera's default 'best guess' algorithm, as the requirements for images quality vary dramatically with the situation.

Crucially, this technology also allows the photograph – to a limited degree – to be 'retaken' with different exposure settings (or even, with some cameras, a completely different focus). This turned out to be particularly critical for this study because of intermittent fluctuations in flash intensity or timing of the camera equipment that would occasionally occur; the result was that consecutive photographs in a series would sometimes appear to be slightly brighter or dimmer than one another, which would interfere with the appearance of a smooth transition between the two in the final blend. Though editing the rendered the JPEGs offer limited ability to correct for this, it can only be effectively accomplished from the RAW files. Corrections by 'reprinting' the RAW files to JPEG yields far superior results, and preserves fine detail much better.

The RAW files for each series were therefore re-rendered using 'darkroom' software, with equalisations to the lighting conditions applied, and suitable levels of

contrast, brightness, tint, hue and other settings to bring the final image to an acceptable level of clarity and definition.

6.2.4 Centring and cropping the images. Once the images had been remastered and normalised for lighting levels, the next stage was to centre and crop the images. In a scale comprising an animation between a number of consecutive images, the anchor images must be carefully aligned to ensure a smooth transition from one image to the next.

It was decided that all the animations should keep as their main focus one (or if possible both) of the eyes. In sequences where there is a lateral tilt of the head (that is, sideways, such that the face rotates about the plane of the image) alignment was maintained with the upper of the two eyes, so as to preserve any downward slump present between one endpoint image to the other. For forward tipping of the head, however, it was not possible to show the downward turn without sacrificing a consistent alignment of the eyes throughout a series. It was decided that using the eyes as the central point about which the other features changed was more important than allowing the point of view to move upwards or downwards to reflect the absolute position of the head within the frame of the photograph, as the focus of our vision naturally falls upon the eyes of a face whose expression we want to interpret. Secondary cues, such as the visible forwards, or rearwards tilt of the head are readily apparent in the images and should suffice in furnishing these peripheral nonverbal cues.

To accomplish this, all the adjusted images were loaded into Photoshop as different layers, and labelled accordingly with their assigned sequence code. Each layer was turned into a transparency, and then carefully aligned with the next in the sequence by superimposing the image at an agreed point, usually the pupil of one of

the eyes, zooming in as far as necessary to examine the superimposition in fine detail. Once this part of the image was at maximum focus, the layer positions were set, and the same procedure was completed with the next two layers, and so on until all the images were aligned. Finally a cropping tool set to 4:3 portrait aspect-ratio was used to crop the image in such a way that the entire series were well framed and precisely aligned with one another.

6.2.5 Summary. In a second photographic sitting, a new series of images for two, high-scoring actors was collected and processed. These images provided a series of transitional facial expressions marking intervals along each of the scales from which further transition images would be interpolated; they also provided new, bipolar, endpoint images which would enable coordinate systems to be created in which to combine the separate scale scores. The next stage, though, was to run two more judgement experiments. The first would provide the PCA plots required to create the coordinate systems, while the second would allow scaling data for the transitional images to be collected based on their perceived locations on their respective scales. Together, they would furnish the data required to generate a complete set of 100-interval scales, and a coordinate system within which the separate scores could be mathematically combined into a total score.

6.3 Part 2: Judgement Task for the New Endpoint Image Sets

As was discussed in Chapter 5, the 'words' plots (see Figure 5.9a and Figure 5.9b) from Study 1, part 3 offered a basis from which to derive a set of bipolar scales, each of which represent a trajectory across factor space between suitably located points about the circumference of the affect circumplex. Though the coordinate system

yielded by this exploratory study offered a good starting point from which to empirically locate facial expressions within such a coordinate system, it was not a suitable basis for a deriving coordinate systems pertaining to specific images sets.

This exploratory plot, remember, was based on composite data from images of posed facial expressions from all 20 actors who participated in this initial, exploratory study. Even if only data relevant to the selected actors were chosen, the modest sample size of just 17 datasets for each actor 12-image set would limit the accuracy and external validity of any coordinate system created from a PCA analysis derived thereof.

Furthermore such a coordinate system would apply only to the original endpoint images provided by our actors. For reasons already discussed, it was deemed necessary to recreate these endpoint images as part of the full scales comprising transitional images, and though every effort was made to ensure that the expressions in new set of endpoint images were as close a match as possible to the former, some discrepancies were inevitable, which would no doubt affect their positions in the resulting coordinate system. In any case, the problem of the absence of a suitable complementary endpoint expression for the ‘sleepy’ item demanded an extra, custom-created expression, which would also need to be located in the coordinate systems for the respective actors.

For all of these reasons, it was necessary to perform a further, more extensive judgement study for this new, final set of endpoints images that would comprise the final scales, so that an accurate corresponding coordinate systems could be generated.

6.3.1 Sample Size. Since this part of the study again involved generating plots of factor loadings, the issue of sample size was not as straightforward as alternative modes of analysis, and thus require a slightly more roundabout rationale.

The experiments conducted in study two offered plots and correlation matrices that gave a reasonably clear idea of the effect size. Examination of the pattern of observation would attest to the effect size being at least in the range of 'medium-to large'. This judgement was based on Cohen's criteria for assessing effect size in observations (Cohen, 1992), but erred on the side of understating the effect size so as to minimise the risk of under-powering the experiment. It was the data provided by this study which would be used for the important, final coordinate system in which the data of the different scales would be combined into a single metric, and so it was important to get as accurate and representative data as possible.

Bearing in mind this constraint, a number of figures were examined which generally offer sufficient power (0.8 or more) for studies of a medium to large effect size to detect statistically significant effects using a range of standard parametric and non-parametric tests (Faul et al., 2007). Based on this examination a sample size of 100-120 was established as a reasonable range for this study.

6.3.2 Participants and recruitment. The judgement task for the new endpoint images sets (Experiment 3) was run concurrently with another (Experiment 4), in which judgement data was collected for the position of transitional scale images (see 6.4). Participants were permitted to complete one or both of the tasks; 87% of participants who completed Experiment 3 also completed Experiment 4.

As in Experiment 2, recruitment of participants was conducted internationally in order to obtain a reasonably representative cross-section of English-speaking participants. The study was advertised on online forums and through social networking media such as Twitter and Facebook, and through canvassing of a small number of U.S. and Canadian cities and universities. The study was also advertised locally by a number of advertisements displayed at locations about University of

Nottingham. As before, a small payment was offered for participation in an experimental task. Selection criteria were that participants 1) were at least 18 years of age, and 2) were fluent in English. The data were provided by a total of 110 participants, 59 male and 51 female, aged 18 to 67 years (mean 29.9 years; S.D. = 10.2 years). Of these participants, 75 (68%) were ethnically European, 25 (23%) were Asian or East Asian, 4 (3.6%) were of African descent and 6 (5.5%) described themselves as mixed race. 94 (85%) spoke English as a first language, while the remaining 16 (15%) spoke English fluently as a second language.

6.3.3 Method. The method for this experiment essentially duplicated that of Experiment 2, except only the newly collected endpoint images for actors #1 and #2 were used, including the new, 'alert' expression that would form the endpoint opposite to 'sleepy' on the *Sleepy-Alert* scale. Mid-point 'neutral' images were also included to provide additional scaling data.

As before, the task was administered via the purpose-built web portal. Before taking part, participants were required to complete a brief sign-up process in which a login was set up and an email contact address provided. Details of gender, year of birth, ethnicity, and whether they speak English as a first language were collected, and participants who do not speak English as a first language were asked to select their country of origin from a dropdown list. Finally, participants were asked to indicate consent by checking boxes by each of five clauses listing terms of participation (see Appendix IX), confirming that they had read the study information sheet (see Appendix X) and were aged 18 years or above.

For each task, a total of 50 images were presented. 26 were from Actor #1, and the other 24 were from Actor #2. The images included endpoint images and neutral

images from complementary ends of the scales. For technical reasons, images for the *Angry-Peaceful* scale for Actor #2 were not available at this point.

For the experimental task, the images were presented, in turn, on a separate page, accompanied by 12, 7-point Likert scales corresponding to the 12 mood words selected for the final scales. For each image, the participant rated each of the mood words according to how well they deemed it to agree with the image. As before, both the order of presentation of the images, and the order in which the mood words were listed were randomised, with consecutive pages containing mood words listed in a newly randomised order. As before, participants were free to save their session at any point and return to it later. At the end of the task, a message was presented thanking the participant for their time.

6.3.4 Analysis & results. A total of 110 datasets were collected, this endpoint being decided by examination of successive means for the judgement task as detailed in Section 6.4.4. For each candidate image, participants were asked to adjust the slider to reflect where along the corresponding scale they judged it to lie, and then hit the 'submit' button to continue to the next page. As before, participants were free to save their session at any point and return to it later, and at the end of the task the participant was presented with a message thanking them for their time. Because of the high correlations between factor loadings for cell-level and direct PCA analyses, and the superior model fit of plots from the indirect PCA in the results of Study 1 (see 5.3.4.3 & 5.3.4.4), indirect PCA analyses were performed on the centred data to create plots of words and expressions, based on which the respective coordinate systems would be derived.

Looking to the 'words' projections, we can see how judgements of words in this experiment are very similar for the two actors, and how closely the two plots

denote affective space. The ‘angry’ word is slightly adrift, but this was unsurprising, as no ‘angry’ face for Actor #2 was present for this experiment: It is the very high scores of the word ‘angry’ in response to the angry face that help to identify its location within this space; without the key corresponding image for this plot, the ‘angry’ item cannot be so accurately located, here.

In almost every other respect, though, the plot is very close, and we can see how the seven bipolar scales emerging from Study 1 (see Table 5.9) hold up very well. As evident from the plot, however there was no word here that could act as an endpoint for the *Sleepy-Alert* scale in this projection; for the purposes of our scale, actors were simply instructed on the type of high activation, neutrally valenced mood state (‘alert’ or ‘surprised’) for which an expression was needed.

The key to charting the D-VAMS scales within this space, however, lies within the ‘expressions’ plots, as it is these plots that tell us what proportion of valence and activation each endpoint image was judged to possess, and therefore whereabouts within this space they are located. Once the endpoint images have been located, then the scales themselves will then be charted as trajectories across this space, giving a common coordinate system within which scores on the separate scales are combined.

When we confine ourselves to the data of poses for just two individuals, however, significant individual differences present themselves. When data for 20 actors were taken, the expressions projection was much like that of the words, with expressions broadly agreeing with the position of their respective word markers in factor space. Looking at the plots for each individual (see Figure 6.3 and Figure 6.4), we can see some key differences in the way the endpoint expressions are located about the CMA.

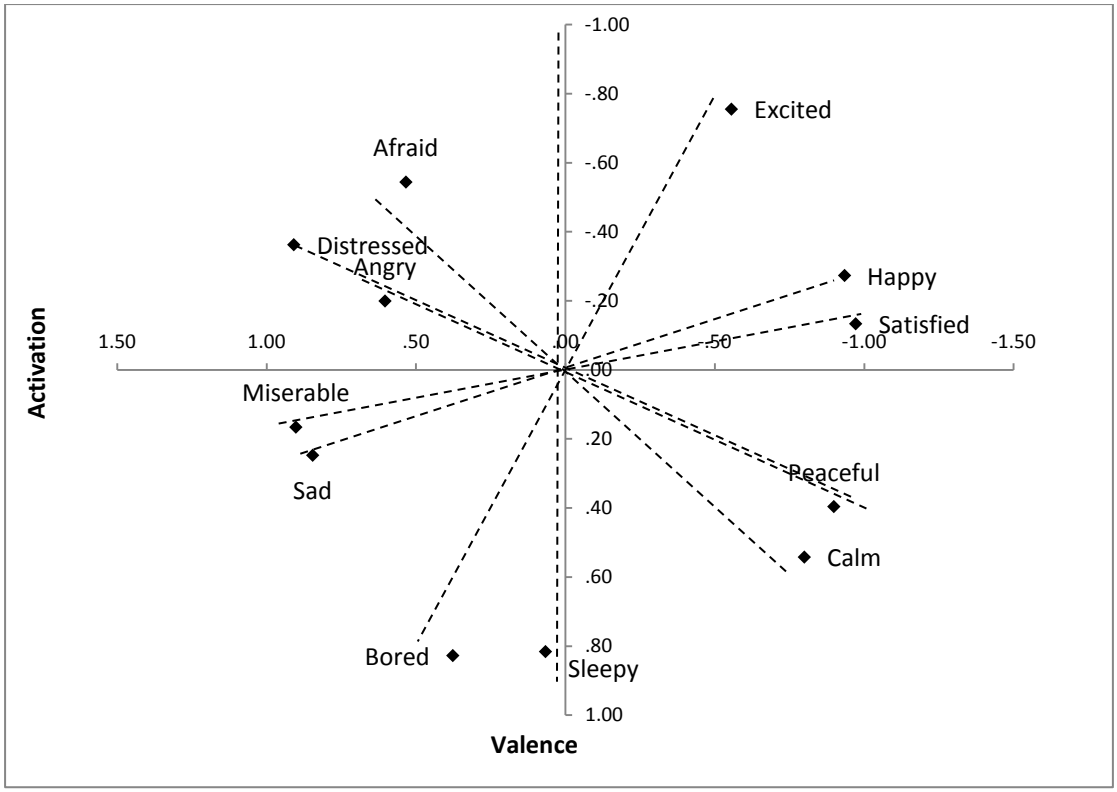


Figure 6.1. *Actor #1: Indirect PCA - Words. Plot of factor loadings.*

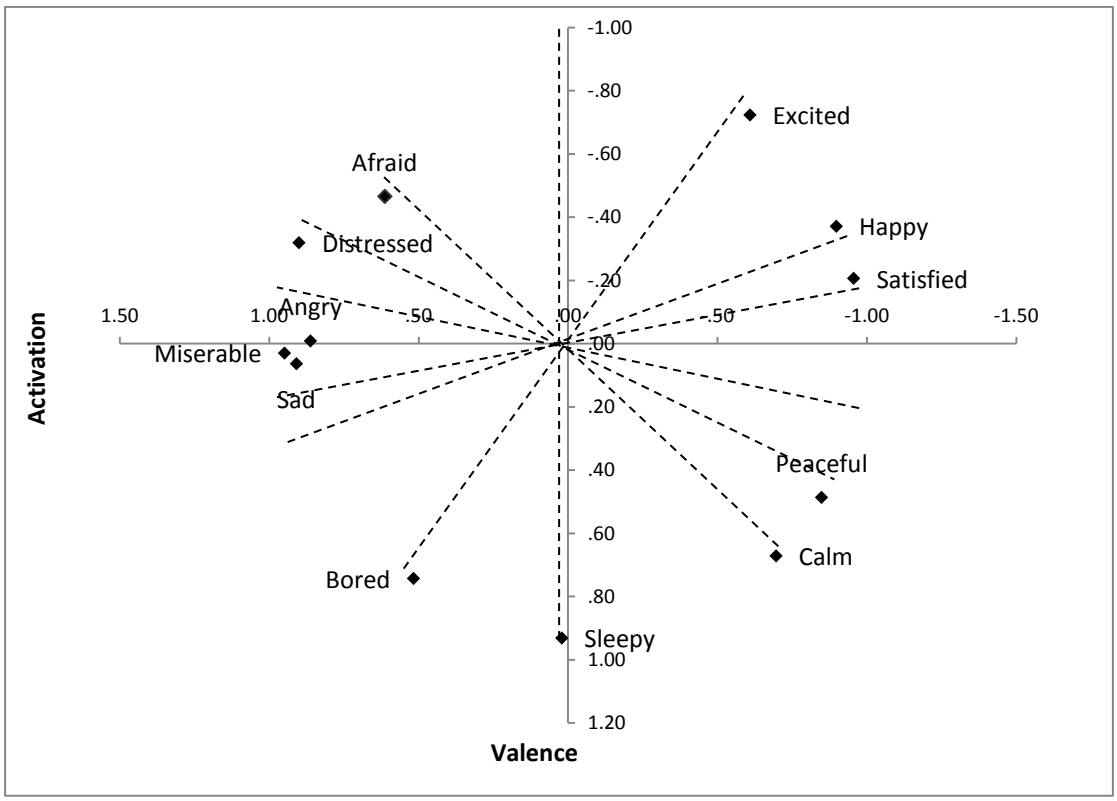


Figure 6.2. *Actor #2: Indirect PCA - Words. Plot of factor loadings.*

Though the opposite ends of the seven scales are all located approximately bipolar to one another, and the plots as a whole broadly agree, there are some irregularities. First, the plot of the high activation, neutral valence expression for Actor #1 was rather close to the plot for her 'afraid' expression, suggesting that the expression was interpreted as slightly negatively valenced, rather than as a neutrally valenced expression. This seems to be a notable problem with high activation expressions. With so much of perceived affective quality being informed by context, there is a tendency for high activation expressions to become ambiguous and prone to a certain degree of projection. A look of intense surprise might be interpreted as terror or elation, depending on imagined context, and gender stereotypes may also introduce subtle biases. The location of *Sleepy* at the opposing end of the *Sleepy-Alert* scale was almost exactly neutral, however, representing an excellent, pure-activation, marker. Because of the off-centre 'alert' marker, the *Sleepy-Alert* scale for this actor was therefore mildly skewed, about 8° from the line of best fit for the respective endpoints.

For Actor #2, the *Sleepy-Alert* scale was also slightly off centre, with a similar bias in the inclination of its line of best fit. Though the high valence, 'alert' item was judged as considerably more valence-neutral, its 'sleepy' counterpart was placed towards the positively valenced, low activation quadrant, closer towards 'calm' than in the plot for Actor #2.

This much higher activation of 'afraid' of Actor #1 also impacted on the goodness of fit of the *Afraid-Calm* scale across the CMA. With 'calm' being plotted as lower activation and more positively valenced than in the 'words' projection. For Actor #2, however, the 'afraid' and 'calm' items were almost perfectly bipolar, forming a very clean line of best fit.

The *Distressed-Peaceful* scale fared better for both actors. Though the 'distressed' face for Actor #1 was placed as more neutral than high activated, the line of best fit forms a reasonably clean, bipolar trajectory, with its 'peaceful' counterpart (±5° from best fit). This is even more pronounced for Actor #2, for whom 'distressed' and 'peaceful' were plotted as almost perfectly opposite one another.

Some notable skew was also evident for the *Bored-Excited* scale for Actor #1. For this actor, the 'excited' expression was judged to be much lower activated than in the 'words' projection, resulting in a particularly sharp departure (±12°) of this scale from its line of best fit. Again, however, the scale for Actor #2 was remarkably close to perfectly bipolar, within 1° of its line of best fit.

Plots of expressions for both the *Miserable-Satisfied* and *Sad-Happy* scales were only approximately similar to their placement in original 'words' projection. As with the Experiment 2 plots, 'happy' was judged as higher activation than 'satisfied', while 'sad' was judged to be lower activation than 'miserable', with *Sad-Happy* appearing at a slightly steeper inclination to *Miserable-Satisfied*, which appeared generally to align more with the valence axis. Plots of the faces for these scales by both actors, however, do not appear to reflect this nuance, with the respective posed expressions varying in their fit. 'Miserable' and 'Sad' expressions are identically located for Actor #1, but reversed from their 'words' projection (and those of Experiment 2) for Actor #2. However the 'Satisfied' and 'Happy' positions seem to broadly agree with the pattern of inclination in the 'words' (and Experiment 2) plots. Overall, however, both of these scales are slightly skewed from the line of best fit from their theoretical counterparts; *Miserable-Satisfied* was ±9° for Actor #1 but only ±1° for Actor #2, while *Sad-Happy* was ±6° for Actor #1 and ±9° for Actor #2.

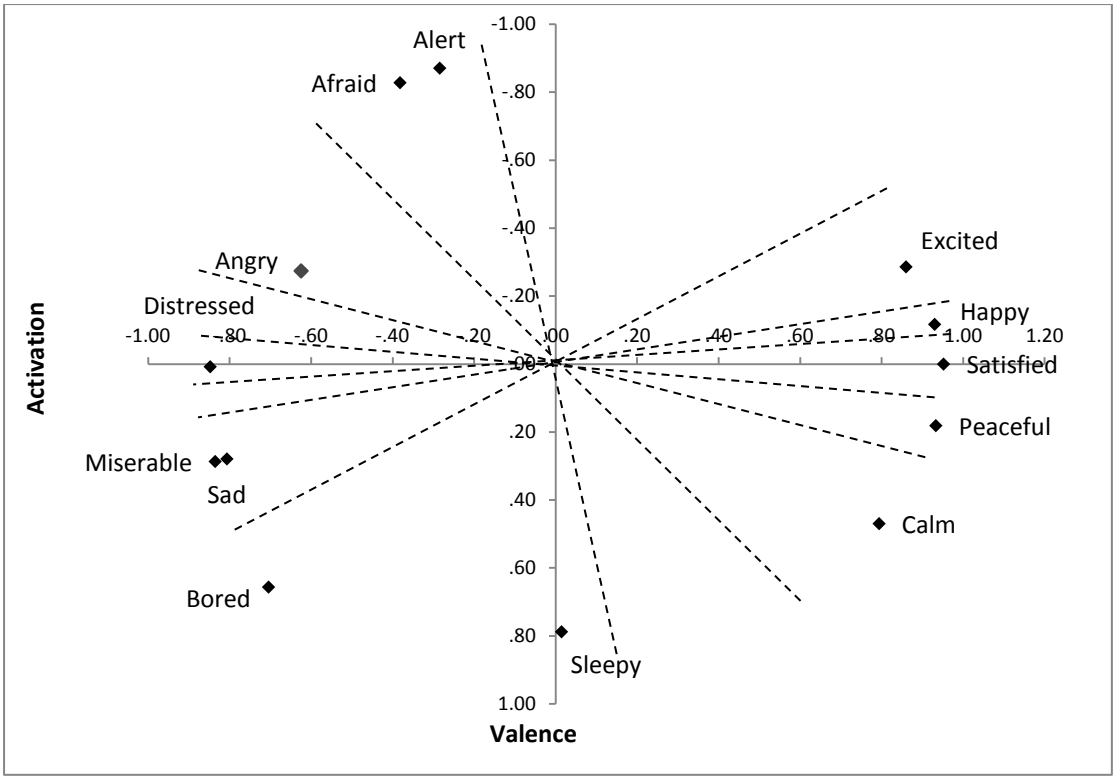


Figure 6.3. *Actor #1: Indirect PCA - Expressions. Plot of factor loadings.*

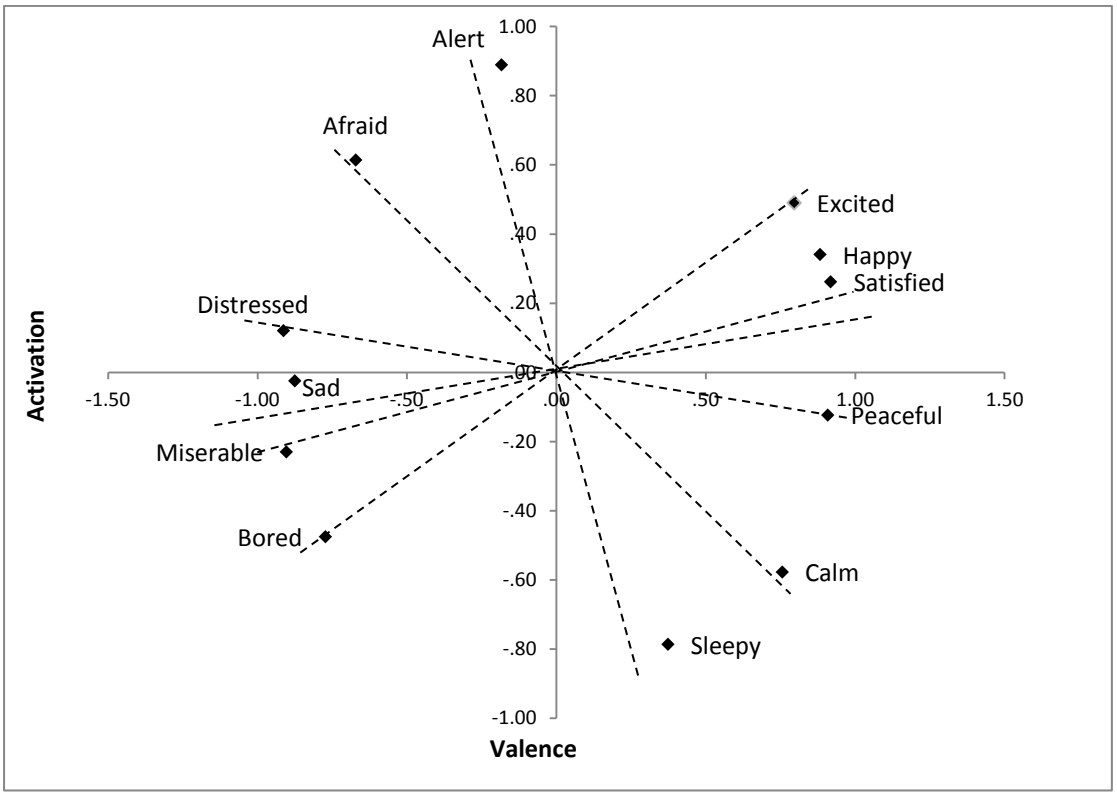


Figure 6.4. *Actor #2: Indirect PCA - Expressions. Plot of factor loadings.*

The *Angry-Peaceful* scale formed a fairly, good-fit bipolar scale of ±7° for Actor #1. For Actor #2, whose 'angry' image was not available until later, this angle was interpolated from data from the Study 1 'expressions' plots; the resulting bipolar scale also yielded a line of best fit of ±7°.

6.3.5 Discussion. So though the proposed scales concur reasonably well with the lines of best fit for the expressions plots for both of the two, finalist actors, there are also some notable discrepancies and irregularities throughout. This raises the significant conundrum of how to handle discrepancies in bipolarity of the scales when it comes to charting them as part of a unified coordinate system. This question can be addressed by asking how it is that the scales are construed by people making the judgements. For example, if presented with a scale with a neutrally valenced 'sleepy' face at one end, but a slightly negatively valenced 'alert' face at the other, how would somebody interpret the scale and its trajectory across affective space?

The first possibility is that the scales would be interpreted as they are, that is, as *approximately* bipolar scales that each consist of two unipolar scales end-to-end, meeting at the neutral centre-point of the circumplex. A measure above the midpoint (score=50) of a single DVAS scale would represent one angle of incidence within factor space, whilst measures below it would represent another that is not necessarily 180° from its counterpart.

The alternative, however, is that the assembly of approximately bipolar scales end-to-end would be perceived differently, with the scale represented psychologically as if it were perfectly bipolar. This might occur as an artefact of expectation regarding the nature of these scales, with the assumption of bipolarity artificially 'straightening' the scales. In this latter case, it would be better to treat the scales is they fell along their lines of best fit, calculating angles that artificially render the scales as perfectly

bipolar, and using these to resolve scores on separate scales into components that can be combined.

These two alternative modes of rendering the scales will be referred to as *unipolar* and *bipolar* renditions. One might envisage a number of arguments both for and against either of this systems, but since no ready answer exists as to the superiority of one over the other, both of these renditions will be applied so that the results of the two can be empirically compared in the course of the validation study of the prototype D-VAMS. A full description of these coordinate systems and the way in which the scale scores are rendered mathematically the into a combined metric will be discussed in the next chapter in the section detailing the charting and construction of the prototype scales (see 7.1).

As well as providing the key plots for these coordinate systems that would enable scale scores to be charted and combined, this experiment offered another source of convergent validity in the form of data that could be compared with that of Study 2. It also allowed this structure to be compared between the judgement data for the two separate actors. To compare these coordinate systems, Pearson's *r* correlations were computed to examine the correlations between factor loadings for the PCA of the present judgement study with those of Experiment 2. This correlation was applied to factor loadings for the 'words' projection, as this is the one that references affective space in terms of the meaning of the words, rather than that delineated by the corresponding facial expressions (see Table 6.3).

The very high correlations for both factor 1 (valence), and factor 2 (activation) between the two experiments – with their independent samples of participants – attest further to the fidelity of affective space, while the slightly lower correlations on the activation dimension are consistent with the larger amount of random error and

reduced proportion of variance accounted for by this second factor. The slightly superior correlations of Actor #1 over Actor #2 are also consistent with the rankings assigned to the respective actors, as the superior performance of actors' poses will also be reflected in the structure of underlying affective space extracted by PCA. A series of neutral or vague facial expressions, for example, will introduce more error into the judgements data than a set of clearly defined expressions.

Table 6.3. *Actor #1/#2 Factor Loading Correlations (Pearson's* r*), with Experiment 2('words' PCA)*

	Valence	Activation
Actor #1	.999	.989
Actor #2	.997	.983

Significant at the 0.01 level (1-tailed).

Another source of convergent validity can be taken from examining the correlations between PCA factor loadings for the two actors, in both 'words' and the 'faces' projections. This form of validation is very important from the point of view of convergence, as the data collected in this experiment can be seen as two concurrent, identical studies conducted on sets of images from two separate actors. Pearson's *r* correlations were therefore also computed for these.

Table 6.4. *Factor Loading Correlations (Pearson's* r*) between Actor #1 and Actor #2('faces' PCA)*

	Valence	Activation
Words	.995	.974
Expressions	.983	.958

Significant at the 0.01 level (1-tailed).

In this respect, too, the results were fairly decisive (see Table 6.4): In the 'words' projection, extremely high correlations were found between factor loadings for both the valence and activation dimensions, attesting to the fact that they embody the same affective space. The expressions projections showed slightly lower correlations, however this is unsurprising as we would expect different actors posing expression in response to the same mood word to produce expressions that were close, though not identical. As was seen in the corresponding plots (see Figure 6.3 & Figure 6.4), there were some differences between the way their expressions were judged, and consequent variations in their individual coordinate systems. As with the comparison of factor loadings for 'words' and 'expressions' in Study 1 (see Table 5.5) the data supported the 'expressions' projection as being more subject to random error because of these individual differences in posed facial expressions.

The results of Experiment 3 have therefore provided the coordinates systems through which separate scale scores will be combined into a single metric. The scales, however, still remained to be constructed, with additional, second-order morphs filling in the many images required for a 0–100 point DVAS. Before these extra images could be generated, however, the positions of these transitional images along their respective scales needed to be quantified, and this required a further judgement study.

6.4 Part 3: Judgement Task to Locate Scale Positions of Transitional Images

The transitional scale images created in the second photographic sitting (see Appendix VI & VII) represent a clear sequence of images showing various stages of transition between one endpoint image and another. The position of the endpoint images can be approximately mapped within the circumplex coordinate system

described in the last section, using PCA factor loadings plots of the faces projection of data from the judgement task described in the last section. However, there remains the question of what intervals these successive images represent in terms of a notional quantity of a particular affect type. How much anger, for example, does image AG-P/02 ‘have’, compared to, say, AG-P/03? At what intervals should each of these keyframe images be placed along their respective scales? The scales, remember, will eventually consist of 101 images each (indexed 0-100), and though the scales can be made more granular by morphing successive images to create an arbitrary number of digitally interpolated (morphed) second-order images, we need to arrive at a way of estimating the placement of transitional these images along the scales at an approximately interval level.

To answer this question we therefore need to find out how people rate these images in terms of where they would judge them to be along their respective scales. Naturally, we can expect differences in the way that people transpose a perceived facial expression into a notional quantity of an affect type, but in order to arrive at the as accurate a coordinate system as possible, the placement of these images should be based on the means of judgements returned by a representative sample of people.

To accomplish this, a second task was therefore devised to collect this data so that the transitional images could be mapped quantitatively.

6.4.1 Sample size. Again, there was the question of what sample size to use to offer an acceptable degree of accuracy of the population means under examination. In this respect, the same rationale as that described in 6.3 can also be applied here, with the observed effect sizes and power requirements warranting a figure in the range of 100–120.

Another consideration is the margin of error that we might consider acceptable in the placement of these items on the respective scale. A 95% confidence interval for a 10% margin of error would seem like a reasonable, minimum level of accuracy, and this translates to a minimum sample size of 96, which agrees reasonably well with this latter estimate. Since these data are critical to the process of establishing accurate information about the way images are judged to be positioned on these scales, a target sample size of 100-120 was therefore chosen in order to ensure reasonable quality of normative data. This part of the study ran concurrently with Part 2 of the study (6.3).

But there is also another rationale which can be used to further guide the sample size. This rationale is based on the resolution of the scale, which – as with an ordinary VAS – is granular to 1% intervals. This granularity offers a somewhat more concrete way to determine the point at which further data will not make any significant difference to the mean judgments for the sample.

With each dataset collected, successive means change by progressively smaller values, regressing exponentially towards a theoretical true mean. The mean values, rounded to the nearest 1%. will therefore become successively less likely to change as fluctuations from one value to the next shrink to less than 1%. A reasonable rule of thumb – bearing in mind the proportions of the project – is that once less than 10% of the rounded, 1% mean values change with successive datasets, the means can be deemed reasonably stable. Successive means were therefore also examined to chart their trajectory as data was collected, and to allow the proportion of values changing at the 1% level to be examined at each successive interval.

6.4.2 Participants and recruitment. The judgement task for the scale positions of transitional items (Experiment 4) was run concurrently with Experiment 3, in which judgement data was collected for the position of transitional scale images

(see 6.3). Participants were permitted to complete one or both of the tasks; 88% of participants who completed Experiment 4 also completed Experiment 3.

As with Experiment 2 and 3, recruitment of participants was conducted internationally. The study was advertised by posts on online message boards and through social networking media such as Twitter and Facebook. Adverts were also displayed at a small number of U.S. and Canadian cities and universities, and at locations about University of Nottingham. As before, a small payment was offered for participation in an experimental task Selection criteria were that participants 1) were at least 18 years of age, and 2) were fluent in English. The data were provided by a total of 110 participants, 57 male and 53 female, aged 18 to 67 years (mean 29.4 years; S.D. = 9.6 years). Of these participants, 75 (68%) were ethnically European, 25 (23%) were Asian or East Asian, 5 (4.5%) were of African descent and 5 (4.5%) described themselves as mixed race. 95 (86%) spoke English as a first language, while the remaining 15 (14%) spoke English as a second language.

6.4.3 Method. As with prior experiments, the task was administered via the purpose-built web portal. Before taking part, participants were required to complete a brief sign-up process which allowed them to set up a username on the system and provide a contact, email address. Participant gender, year of birth, ethnicity, and whether English was spoken as a first language were collected, and participants not speaking English as a first language were asked to identify their country of origin. Finally, participants indicated consent by checking boxes by five clauses listing terms of participation, and confirming that they were 18 years of age or above.

For the task itself – which was again implemented via the web portal – participants were presented with a total of 174 images representing each of the scale transitional images for both the male and female actors respectively. For technical

reasons, scale 6 of the male scales could not be collected until later so were judged in a separate micro-study (*n*=50) later on.

The images to be judged were each presented, in random order, on consecutive web pages. Beneath each of the images were the end-point images for its respective scale, with a horizontal slider between the two. The positions of the endpoints were reversed at random to counterbalance for any systematic biases due to the sequence of presentation.

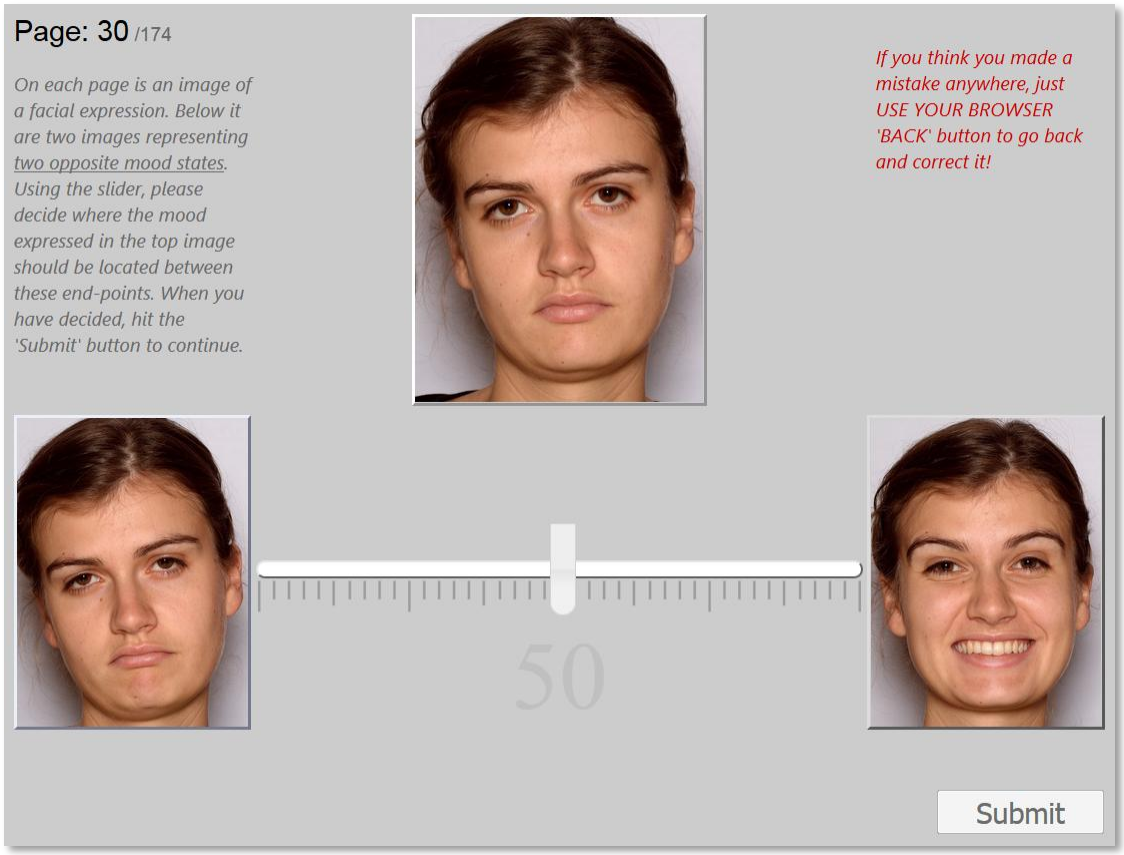


Figure 6.5. *Experiment 4 - Response Page for the Scaling Task from the Project Website*

For each candidate image, participants were asked to adjust the slider to reflect where along the corresponding scale they judged it to lie, and then hit the 'submit' button to continue to the next page. As before, participants were free to save their

session at any point and return to it later, and at the end of the task the participant was presented with a message thanking them for their time.

6.4.4 Results & discussion. The median duration of the task for participants who completed it in a single sitting was 36 minutes ($\tilde{x}$=43 minutes). Charts of successive means of judgement data returned by the participants were examined as the sample size approached the target range of 100–120. By the time data from 105 participants was collected the number of percentage-rounded values changing with each successive dataset was consistently below 10%, and so data collection was terminated at n=110.

Once the data had been collected, the positions of the images along their scales were charted based on means of returned judgment data for each image. These were then examined to finalise the selection of images to be used as keyframes for the final scales, and to enable the number of interpolated, morphed images necessary to complete the scales to be calculated. Examining these preliminary charts, it was evident that the means for their judged positions for the most part concurred with the sequences in which the photos were taken, and followed the sequence of the expressions across their respective scales. The standard deviation, however, was fairly high (mean of 11.3% across all images), and so it came as no surprise that some exceptions existed where adjacent images appeared out of sequence. Observing individual differences in the pattern of judgements returned by participants, it was clear that though the baseline sequences of expressions along the scales were for the most part preserved, the way that people scaled them did vary somewhat, with the distributions differing to varying degrees.

Hence, though the ordinal structure of the scales by and large corroborated the sequence of the facial expression transitions photographed, a small number of

irregularities were noted. These was particularly notable where many images with very subtle differences were clustered close together within just a few percent of one another, with many judged as identically positioned with or extremely close to adjacent images. Where this occurred, images were examined and those deemed to be surplus to requirements were omitted. Out of an original pool of 184 transitional images, 27 were discarded, and the positions of a further 11 anomalous placements resulting in sequence failures were adjusted manually, guided by close examination of facial expression changes between consecutive images.

Once the redundant, transitional images had been dropped from the scales, the remaining 158 became the keyframe images from which (along with their endpoint images) the scales in their entirety would be generated. Charts of the scale keyframes for both actors are shown in Figure 7.7 and Figure 7.8.

Having created and selected groups of images to represent transitions across the scales chosen, and mapped the scale locations within the CMA and the positions of the keyframe images on their respective scales, two phases of the project remained. The first phase would involve the actual construction of the scales based on the data collected in this study. This would involve interpolating the additional images required between the keyframe images charted, and building an interface to house them and support an automated assessment run using images from either actor. An algorithm would also need to be devised to combine all of the scales scores into a unitary measure based on the CMA coordinate system. The second phase would comprise a crucial validation study, in which its psychometric qualities of D-VAMS would be experimentally assessed in a sample of stroke survivors. These phases will be covered in the next and final chapter.

7. Study 3: Construction and Validation of the Prototype Scales

In the previous chapter, a metric was created to approximately assess actors' performance in communicating their affective state using their facial expression. The metric reflected the pattern of sensitivity and specificity with which participant judges identified the words used to evoke the facial expressions. These metrics, in turn, allowed the actors to be ranked, and for the significance of differences between the scores to be tested and charted in a correlation matrix (see Table 6.1). The pattern of correlations supports the actor ranking reflecting a significant trend through successive ranked actors. The rankings given were offered some support by a modest correlation in an underpowered comparison with corresponding scores with from the exploratory, 26-item study (see 5.3).

The two highest scoring, available actors, one female (#1) and one male (#2) were then recalled to repose the original expressions for the 12 selected words (see 5.3), but this time as part of scales constructed from the coordinate system in 5.5. For each of the scales, transitional images were created of poses along the continuum formed as the actors changed their expression from one endpoint expression to the other. A third, judgment study was then undertaken to generate faces plots specifically to act as coordinate systems for actors #1 and #2, while a fourth, concurrent study enabled transitional images to be charted along their respective scales at an interval level, identifying redundant images that could be excluded to leave a reasonable spread of consecutive images along each scale. Finally keyframe images were

selected from which full, 0–100 point continua would be created by morphing adjacent scale images.

7.1 Charting and Constructing the D-VAMS Scales

The next stage of the study was to use the information from Study 2 to create a coordinate system within which to mathematically combine the separate scales, interpolate the additional images required, and then construct them as a software interface to run on a tablet, laptop or desktop computer.

First, the data from the judgement study for the new endpoint images (see 6.3) would be used to map the scales for the top-scoring actors within the circumplex structure, enabling scale scores to be expressed as vectors from the origin of the circumplex. These vectors would then be combined mathematically into a mean vector representing the intensity and quality of a person's mood as represented within this two-factor space, and a scoring method would be derived based on the degree to which this vector concurs with mapped constructs central to depression and low mood. Next, the data locating the positions of keyframe images along the respective scales (see 6.4) would be used as reference points based on which additional, morphed images would be generated. These additional images would enable the assembly of approximately interval, 0–100 point scales. Finally, these images would be assembled into interactive HTML pages that would allow the expression of the faces for each of the scales to be modified by adjusting a slider, and the slider setting to be recorded as a score for each respective scale.

7.1.1 Using the coordinate Systems to combine scale scores. First, the PCA 'expressions' plots examined in Study 2, Part 2 were used to derive coordinate systems for actors #1 and #2. As discussed in 6.3.5, two possible renderings of the

scales onto the coordinate scales were possible, one in which the scales are treated as two approximately unipolar scales end-to-end, and the other in which they are artificially straightened into true bipolar scales by means of a line of best fit.

Each of the scales consists of a bipolar VAS with a range of 0 to 100 units. By convention, these are keyed such that 0 corresponds to negative valence, and 100 to positive valence (with the exception of scale 7, which was keyed to negative activation). When these scales are laid across the circumplex at the angles derived from PCA factor loadings, they define a coordinate system comprising axes centred on the 50-point mark of each of the scales, with the 0 and 100 endpoints of each scale meeting opposite points about the circumference. Figure 7.1 to 7.4 show both the unipolar and bipolar rendition of the coordinate systems for both of the actors, with angles calculated from the plots.

The origin of these axes thereby represents a theoretical neutral point in affect space in terms of both valence and activation, and scores on each scale can be viewed as vectors $(\Delta x, \Delta y)$ originating from this point. In transposing scale scores to the coordinate system, each becomes represented as a vector from the centre of a circumplex which, for convenience, is given a radius of 100 units. The length of a vector representing a given scale score can easily be computed by multiplying the score's distance from the scale's midpoint (50) by two, to bring its magnitude within a range of 0-100. So for a score *a* of scale number *n*, the vector size *v* will be denoted by:

$$v_n = 2(a_n - 50)$$

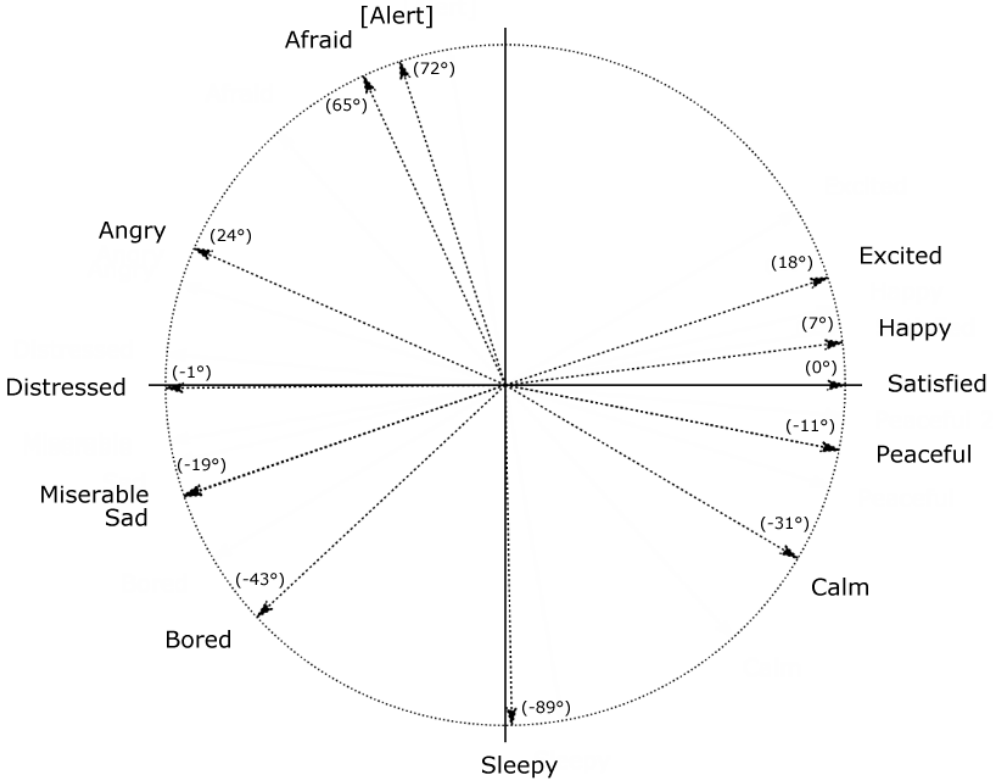


Figure 7.1. *Actor #1: Unipolar Coordinate System*

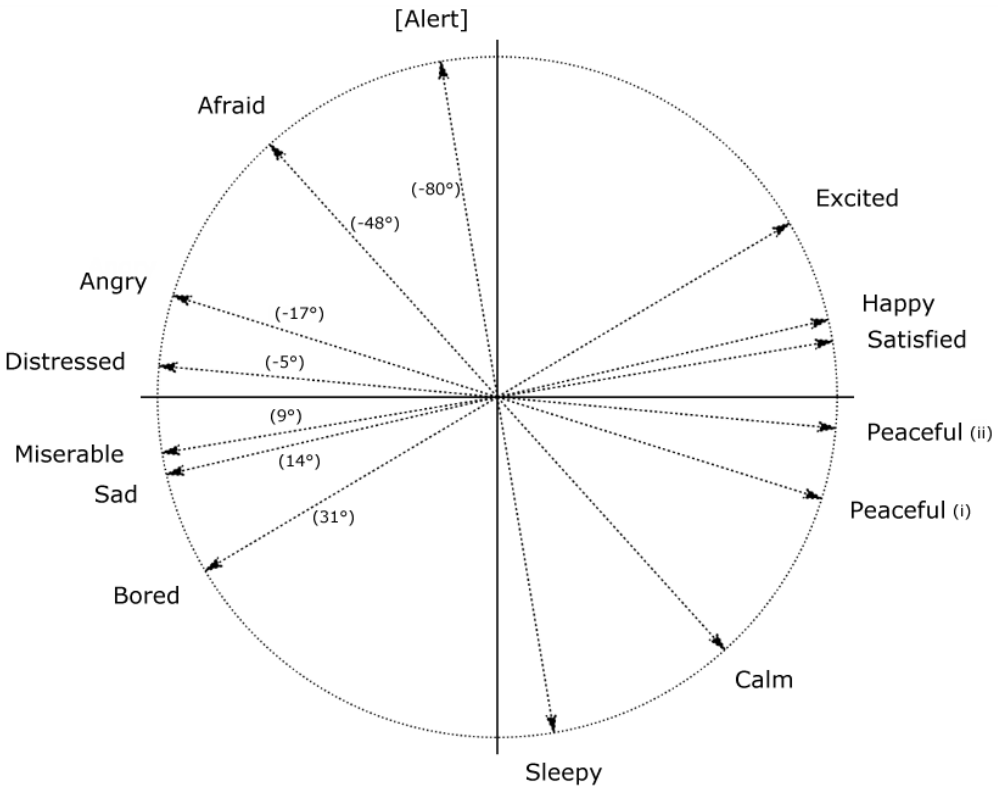


Figure 7.2. *Actor #1: Bipolar Coordinate System*

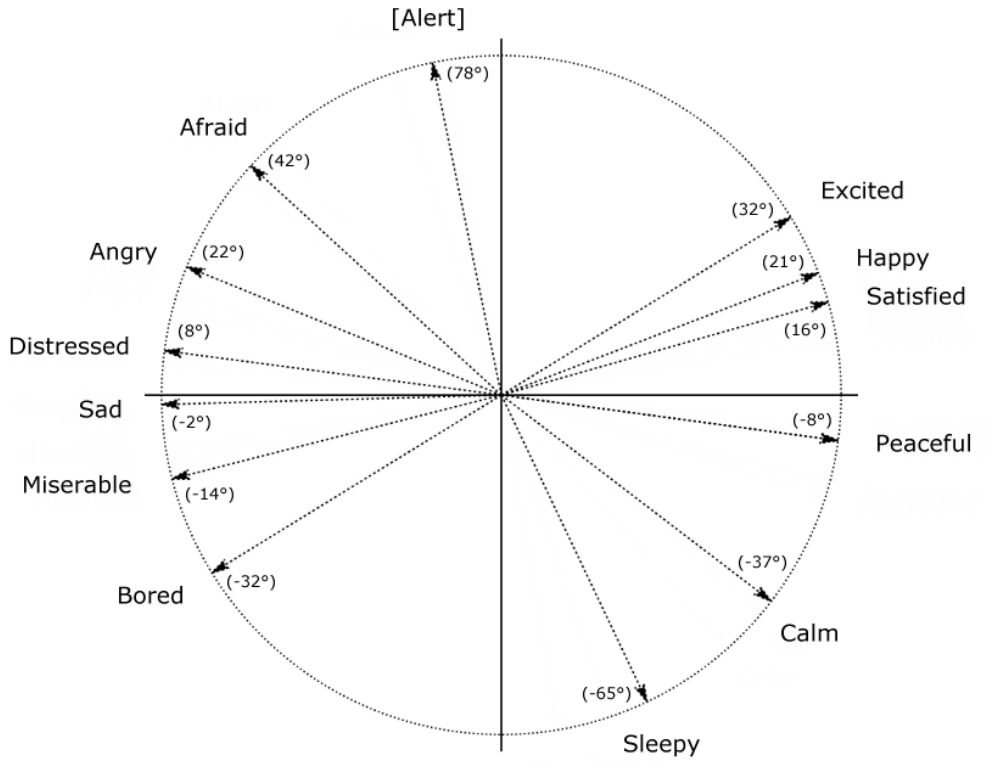


Figure 7.3. *Actor #2: Unipolar Coordinate System*

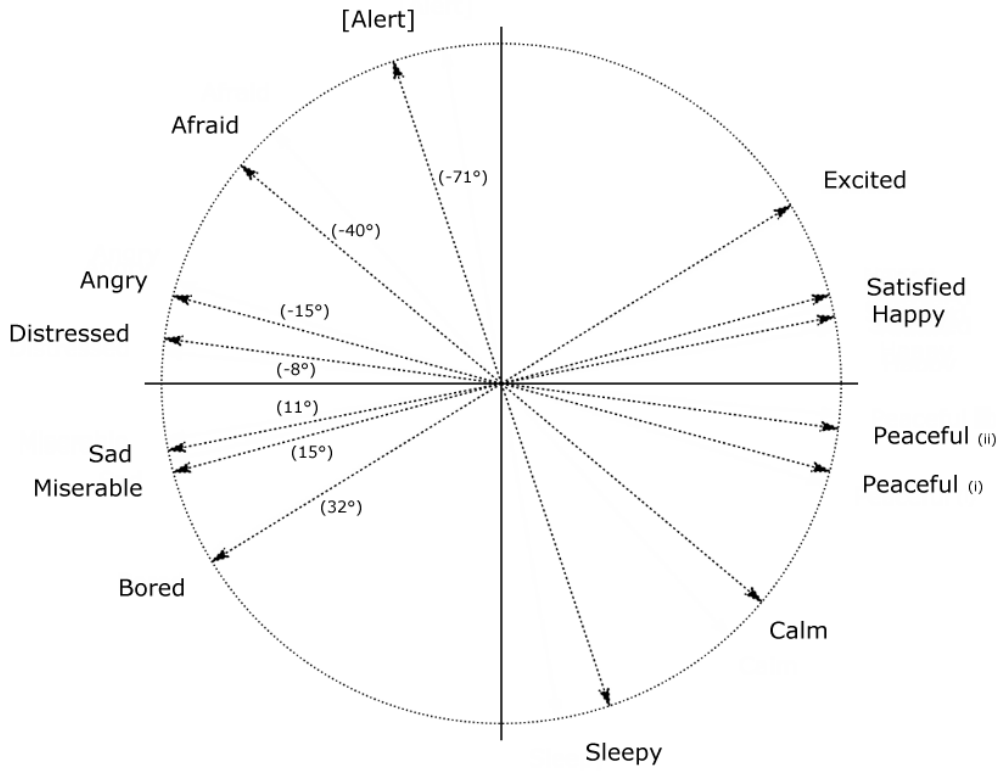


Figure 7.4. *Actor #2: Bipolar Coordinate System*

Having calculated a vector magnitude for each scale score, *v* now needs to be rendered into its valence (Δx) and activation (Δy) components, as defined by each scale's angle of trajectory across the circumplex. For the sake of simplicity, we will use the bipolar coordinate system, here, in which each scale has a single angle of incidence defined by its line of best fit (see Figure 7.2 & 7.4).

In this way, each scale score is broken out into a vector (Δx, Δy), with all seven scales forming a series of points or vectors comprising *x*, *y* values. These are then resolved into a score total in the form of a mean vector which represents the amount and proportion of valence and activation overall (see Figure 7.5).

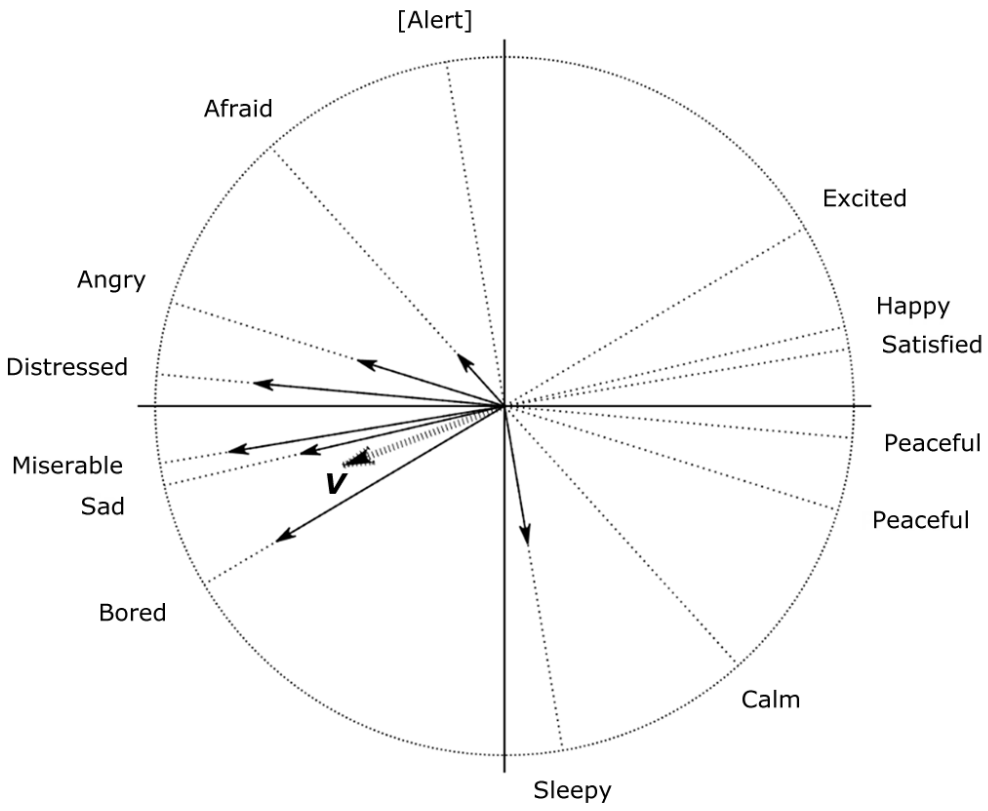


Figure 7.5. *Scale Scores as Vectors within Circumplex, with Mean Vector* V

In order to arrive at this total score, it is a relatively simple matter of calculating the mean of the vectors for all of the scale scores. Where $a_{\{1\ldots7\}}$ are the scale scores as vectors from origin (0,0) in the range of 0 to 100 units, and $\theta_{\{1\ldots7\}}$ are the angles of the scales across affect space, mean valence (Δx) and activation (Δy) components can be calculated thus:

$$\overline{\Delta x} = \sum_{n=1}^{7} 2(a_n - 50)\cos\theta_n/7$$

$$\overline{\Delta y} = \sum_{n=1}^{7} 2(a_n - 50)\sin\theta_n/7$$

The seven scales, however, are only a small sampling of the potential trajectories across the underlying affect space. The affect space which they delineate is therefore distorted due to differences in the total range of activation and valence permitted by them. Were the scales to cross affect space in a regularly distributed range of angles, thereby tapping equal amounts of valence and activation (Δx and Δy), the mean vector could be considered as the mean of the absolute valence and activation components of the respective scores. However, the scales tend toward an overall bias towards the horizontal plane, thus favouring the valence dimension. A multiplier, *K*, must therefore be introduced to compensate for this and render a mean vector that reflects the circumplexical structure of the coordinate system used here to quantify affect space. *K* is calculated by summing the total of the respective valence and activation components of the scales based on their angles, and dividing one by the other, like so:

$$X = \sum_{n=1}^{7} \cos \theta_n$$

$$Y = \sum_{n=1}^{7} \sin \theta_n$$

$$K = \frac{X}{Y}$$

The resulting multiplier is then applied to the calculation to arrive at a vector representing the mean scores, of magnitude V_m and angle V_θ.

$$V_m = \sqrt{(\overline{\Delta x}^2 + K\overline{\Delta y}^2)}$$

$$V_\theta = \tan^{-1} \frac{\overline{\Delta y}}{K\overline{\Delta x}}$$

The result of this adjustment is a mean vector of magnitude V_m and angle V_θ. It can also be represented as the vector $(\overline{\Delta x}, K\overline{\Delta y})$, with the coordinates representing the amounts of valence and activation respectively.

But having calculated this mean vector, what do we do with it to give us a score that we can use to quantify a specific mood, depression-related or otherwise? Well, this is a key advantages of using this type of system: the mean vector can be given a simple multiplier that can yield an index of any mood charted within the circumplex. Since the correlation of two vectors is a function of the cosine of the

angles between them, then all we have to do is multiply the magnitude of the mean vector V_m by the cosine of the difference of angles between it and the theoretical location of any named construct within the CMA. In this way, the D-VAMS can give us not only an index of *sadness*, but also an index of *boredom*, an index of *excitement*, an index of *calmness* and so on, acting as something of a Swiss Army Knife of mood measures.

To do this, we therefore need only have the theoretical angle of a given construct to create any quotient we want. Where C_θ is the angle of the construct *C* within the CMA, and V_m and V_θ are the magnitude and angle of our mean vector, the quotient Q_c (our construct 'score') can be defined by:

$$Q_c = V_m \cos(V_\theta - C_\theta)$$

In this way, in addition to the individual scales scores of the D-VAMS, we can produce more robust construct indices that are based on the scale scores as a whole, and which offer a more reliable metric.

7.1.2 Creating indices for depression and anxiety from the mean vector.

This, however, brings us to a critical question. What key construct or constructs would be most useful in examining depression or low mood in a clinical setting? The primary aim of this project was to create an instrument that could be used as a screening measure for depression after stroke, however 'low mood' more broadly is also something that is of great interest. For the purposes of a screening measure that would be suitable for assessing depression, 'depression' is obviously the construct of key interest, however another construct of interest is that of anxiety, which was

charted – close to ‘afraid’ ‘nervous’ and ‘tense’ – within the high activated, negative valence quadrant of the circumplex (see p.139). Anxiety is of particular interest, as depression and anxiety frequently go hand in hand. With depression and its neighbouring constructs (‘sad’, ‘miserable’ ‘disappointed’) being charted generally towards the low activation, negative valenced area of the circumplex, both can be viewed theoretically as high and low-activated ‘flavours’ of negative valence, which together encompass the concept of ‘low mood’ as a whole.

So, as well as deriving a metric that could act as a depression score (‘DVAMS-D’), there could also be another representing an anxiety score (‘DVAMS-A’). This would be particularly desirable, as the instrument that was selected for the validation study (the HADS) includes both depression and anxiety subscales (the HADS-D and the HADS-A), and a cross-comparison of these subscale-type scores would be enormously valuable in assessing the psychometric properties of the D-VAMS.

But how should we go about deciding where within our circumplex the mood components of depression and anxiety should be charted? The plots from the first two judgement studies (Experiments 1 & 2) offer an abundance of data in this respect, so the positions of terms relating to anxiety and depression were examined in the ‘words’ projection of these plots. Looking first at the ‘depression’ construct, the findings of Experiment 1 show how the words ‘depressed’, ‘sad’, ‘miserable’ and ‘disappointed’ cluster fairly tightly together in the ‘words’ projection, particularly in the cell-level plot (see Figure 5.1b & Figure 5.2b). As noted before, ‘depressed’ and ‘sad’ appear to be almost interchangeable in terms of the way they are used to rate faces, and are characterised by strong negative valence with mildly lowered activation. Taken from the ‘9 O'clock’ position of the circumplex (negative valence, neutral activation), the

mean angle of all of these related terms across the indirect and cell-level analysis was about 7°. For Experiment 2, with its cut down, 12-item word set, only the 'sad' item was examined, with the relevant plot returning a slightly higher value of about 11° (see Table 7.1a & Table 7.1b).

Table 7.1a. '*Depression' Items: Angles from Experiment 1 'words' plots*

	Indirect	Cell-level	
Depressed	-8.58°	-7.04°	
Sad	-8.15°	-9.17°	
Miserable	-2.55°	-5.95°	
Disappointed	-2.47°	-5.96°	
Mean	-6.14°	-7.56°	$\bar{x} = -6.85°$

Table 7.1b. '*Sad' Item: Angles from Experiment 2 'words' plots*

	Indirect	Cell-level	
Sad	-10.33°	-11.38°	$\bar{x} = -10.9°$

Turning now to the 'anxiety' construct, plots for this item from Experiment 1 were examined, along with its conceptually similar 'nervous', 'tense', and 'afraid' neighbours, while from Experiment 2, plots for the 'afraid' item were studied. The means of the Experiment 1 items showed some discrepancy between projections, however the aggregated mean for the Experiment 1 angles was quite close to that of the 'afraid' item of Experiment 2, 30.5° for the former, and 32.6° for the latter (see Table 7.2a & Table 7.2b), indicating stronger between-experiment convergence than the angles derived for depression.

Table 7.2a. *'Anxiety' Items: Angles from Experiment 1 'words' plots (-x, +y)*

	Indirect	Cell-level	
Anxious	27.1°	36.4°	
Nervous	31.0°	38.1°	
Tense	23.8°	31.8°	
Afraid	25.2°	30.2°	
Mean	26.8°	34.1°	$\bar{x} = 30.5°$

Table 7.2b. '*Afraid' Item: Angles from Experiment 2 'words' plots (-x, +y)*

	Indirect	Cell-level	
Afraid	33.2°	31.9°	$\bar{x} = 32.6°$

Another source of guidance in deciding suitable empirical angles for these constructs is through reference to existing literature in which the location of scales and their constructs have been examined. In a more recent study, Yik et al. (2011) offered an update of Russell's earlier (1980) CMA in the form of a 12-point structure the 12-point Affect Circumplex (12-PAC), charting a total of 30 mood scales and 38 personality scales within this structure using two separate procedures (Cosine method and CIRCUM-extension). Examining the mood scales included in their analysis, three scales: 'Sadness', 'Fear' and 'Tension' stand out as items of interest in relation to our Depression and Anxiety constructs (Yik et al., 2011, p. 720). The angles in their analysis are charted from the right of the CMA (at '3 O'clock') running anticlockwise, so must first be subtracted from 180° to bring them into line with the negative valenced reference system used here.

In this respect, the results are again encouraging. Turning to the 'Sadness' scale, the authors arrive at angles of -9° and -10° using their Cosine and CIRCUM

methods respectively. These figures are very close to Experiment 1 values for the 'sad' and 'depressed' items and fairly close to the composite depression index derived from all four items (-6.9°) (see Table 7.1a). For experiment 2, the figures are remarkably close to our values for the 'sad' item, with a mean value of -10.9° (see Table 7.1b).

For the 'Tension' scale, the authors arrive at angles of 37° and 36° in their respectively, figures which are close to our Experiment 1 plots for 'anxious', 'nervous; and 'tense' (see Table 7.2a), though only in the cell-level analysis. The figures given for the 'Fear' scale were particularly interesting, with a figure of 31° emerging from both methods. These are almost identical to the figures arrived at for both the Experiment 1 'afraid' item (cell-level) and the mean, composite anxiety index derived from all four items (30.5°). It also agrees closely with the mean angle for 'afraid' derived from Experiment 2 (see Table 7.2b).

The convergence between these values between experiments described herein, and values arrived at independently from other key research in this area is compelling; it comprises particularly strong convergent validity and attests further to the essential veracity of this CMA-based coordinate system.

It also allowed appropriate theoretical locations of our 'depression' and 'anxiety' constructs to be charted with some confidence. After careful consideration, the Experiment 2 plots were used as the basis for assigned reference angles for 'depression' and 'anxiety' within the CMA. These results were used firstly because the indirect and direct values agreed much more closely than those in Experiment 1, and secondly because of their particularly close agreement with equivalent scales in Yik et al. (2011). The angle between them is also consistent with the high correlations noted between scores of anxiety and depression, which typically exceed r=.65

(Brumfitt & Sheeran, 1999; Feldman, 1995), with a cosine of around 0.7 reflecting this fairly closely.

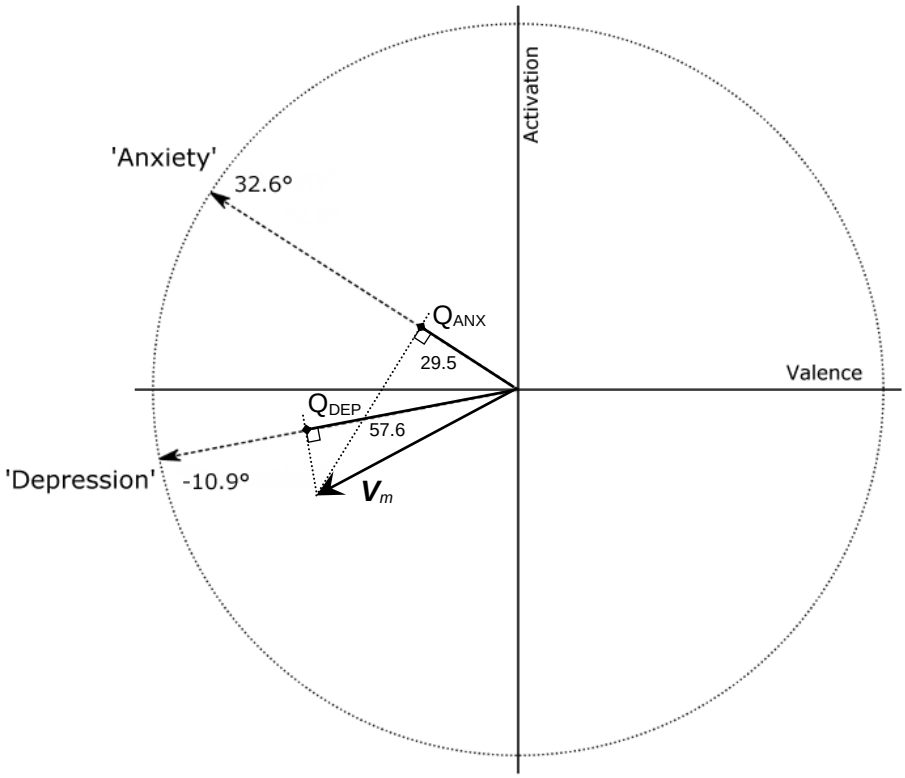


Figure 7.6. *Theoretical Locations of Depression and Anxiety Constructs within CMA, and calculation of Indices from Mean Vector*

Figure 7.6 shows the theoretical locations of depression and anxiety within the CMA structure, and demonstrates how the mean vector calculated from all of the scales scores is resolved into the respective indices of Q_{DEP} and Q_{ANX}. For this example, let us imagine a mean vector *V* of magnitude 60 at an angle of -28°. As described before, the index for any CMA construct is a function of the magnitude of the mean vector V_m and the cosine of the angle between it that construct's theoretical CMA location. Thus, for our index of depression, we first calculate the angle between the mean vector and that of depression (28-10.9=17.1°) and apply the multiplier

cos(17.1°) to the vector size 60. Being such a small angle, this gives us 0.96×60, for a depression index of ***57.6***. For our index of anxiety, the angle between the mean vector and the theoretical anxiety angle is much larger, at (28+32.6=60.6°), therefore our multiplier cos(60.6°) is much smaller, yielding an anxiety index of 0.49×60=***29.5***.

Both of these indices can be seen in Figure 7.6, in which they are visualised as tangents against the theoretical reference lines for depression and anxiety. Similar indices can also be envisaged for any other constructs within the CMA. From here on, the depression and anxiety metrics (Q_{DEP} and Q_{ANX}) given by the D-VAMS will be referred to as 'DVAMS-A' and 'DVAMS-D' respectively, making them analogous to subscales like those used in instruments such as the HADS.

The mathematics of this system, of combining the scores is necessarily complex, however, and some may prefer a simpler and transparent treatment of the scale scores. There are however, alternate, ways of arriving at a sum total score. The scales were all deliberately arranged such that they run from the negatively valenced endpoint to the positively valenced one. Because of this arrangement it is a simple matter to create a basic score of valence by simply taking the mean of our scale scores. However, there is one problem in this respect; Scale 7 (*Sleepy-Alert*) is not aligned along a valence axis *per se*, but along the activation dimension which is theoretically valence-neutral. For a purely valence based total score, we could omit Scale 7 and simply give the mean of the first 6 scales, in the form of a D-VAMS $Mean^{-SA}$ Score, the '-SA' denoting the omission of this scale value. However, since depression is also characterised by low energy as well as negative valence, there is also an argument for keeping the Scale 7 score included, and using the mean of all of the scales instead, in the form of the D-VAMS Mean Score.

In addition, the scale scores can also be used individually in a way that's tailored to the requirements of a particular assessment process. For some purposes, just using one or two of the scale scores – such as *Sad-Happy* or *Afraid-Calm* – might be within acceptable bounds of accuracy, allowing for a briefer assessment process.

Which of these scores offers the best correlates to other measures of depression and anxiety was something that remained to be seen in the validation study (see 7.2), but since they could all be dynamically generated with ease, all were included in the D-VAMS assessment results. As well as the DVAMS-A and DVAMS-D, the Mean and Mean^{-SA} scores were also computed and displayed alongside the individual scale scores.

7.1.3 Charting scale keyframe images and generating morphed transitions.

Having established coordinate systems from the two actors in which scale scores can be resolved into unified measures of anxiety and depression, there still remained the task of turning a small number of keyframe images comprising the scales into a full set of 0–100 point scales with images for each of the 1% increments (see Appendix X & Appendix XI).

In 6.4, the location of keyframe images on their respective scales were charted using data from the judgement study, with the number of required interpolated images between them being mapped in preparation for the final stage of constructing the D-VAMS. Figure 7.7 and Figure 7.8 show the final scales mapped with endpoint images, keyframe images and range markers indicating the number of interpolated images to be generated.

Figure 7.7. *Locations of Scale Keyframe Images: Actor 1*

Figure 7.8. *Locations of Scale Keyframe Images: Actor 2*

Obviously, were there to have been a smaller number of keyframe images judged to be at precise intervals along the scales, then a less granular scale – such as ones based on 7 or 9 images – could have been created. The images collected and judgements thereof, however, were at necessarily irregular intervals, thereby requiring additional images to make them part of a more granular series that could act as an approximately interval scale. Since the morphing process by which these images would be created can generate an arbitrary number of transitions between keyframe images, it made sense to retain the original 1% granularity of a VAS, which in its traditional form has been shown to be an effective way of gathering judgements of interval level data (see 3.3.1). The highly granular nature of the resulting scale is greatly advantageous in that the user is unconstrained by multiple choice categories, and is free to return finely discriminated measurements that are as subtle and nuanced as desired. No doubt such fine resolution is beyond the fidelity of the judgments generally made, but it is better to have too much resolution than too little.

So for this next phase, images between adjacent keyframe images were generated by using morphing software to interpolate smooth transitions between the consecutive facial expressions on each of the scales. In all, 177 morphs were performed between a total of 194 endpoint and keyframe images. For each of the morphs, adjacent images were tagged with between 150 and 200 anchor-points which were used to cross-locate points across the surface geometry of the faces portrayed (see Figure 7.9). These markers were focussed primarily about the face, following the features of the eyes, nose, lips, and the contours of the skin, cross-referencing their outlines and forms by mapping distinctive surface features such as moles, blemishes, or other micro-features so that their textures could be smoothly interpolated. The placement of these markers was a laborious and time-consuming process in which

areas of adjacent images were magnified and aligned, and markers placed as accurately as possible. Close attention was paid to this part of the process in order to allow as smooth and animate a transition as possible between images, which would imitate closely the natural movement of the face as it changes between one endpoint expression and its counterpoint.

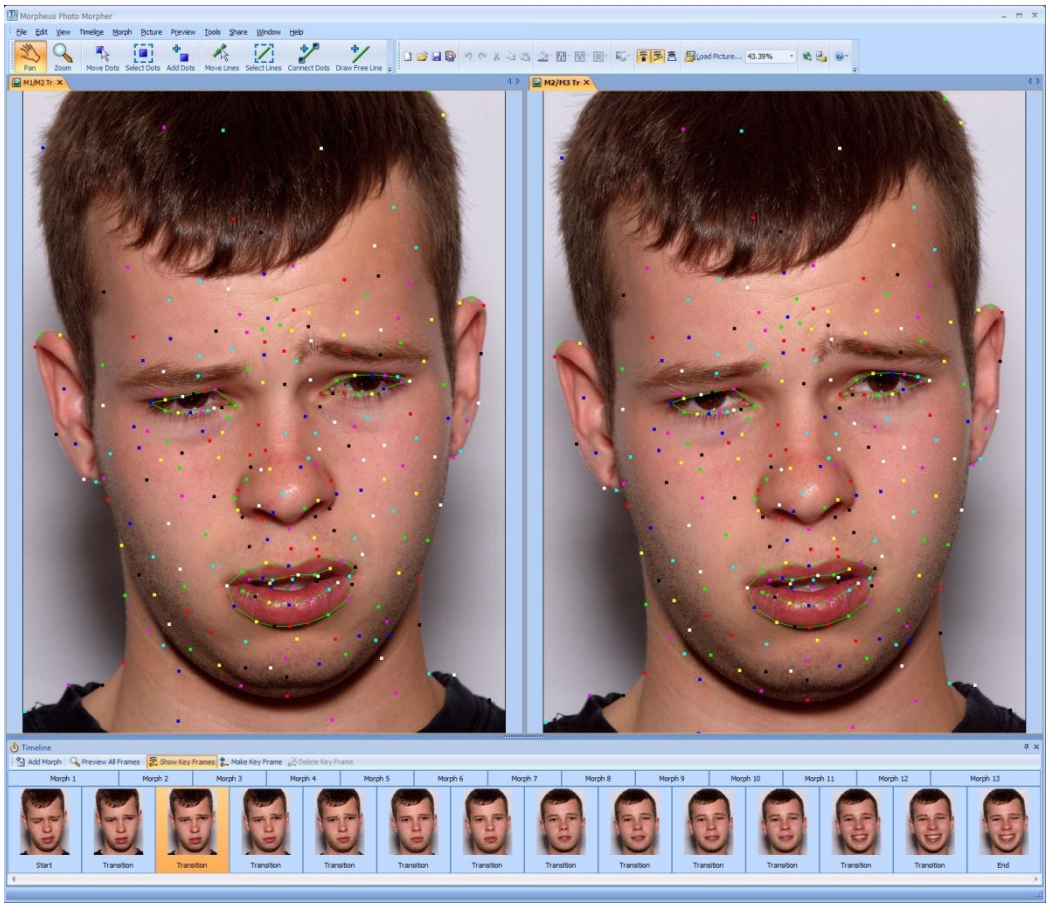


Figure 7.9. *Morphing Images using Surface Reference Points*

7.1.4 Constructing D-VAMS. Having generated all the images required for each of the 14 scales (seven for each actor), there was next the question of how to implement them in a form which would offer the greatest versatility and convenience. In addition to the PCs, laptops, notebooks and tablets on which many varieties of

software run, there are also a growing range of multi-function devices, running a variety of operating systems each with their own demands regarding the way that software is programmed and implemented. Naturally, it would be desirable to have D-VAMS constructed in such a way that it could run on as many devices as possible, regardless of the hardware or operating system. It would also be good if the instrument could be made publicly and remotely accessible to any device with an internet connection.

Both of these requirement immediately suggested the current, *de facto* method for managing interactive content to an agreed universal standard, that is, as a standard HTML web page that can be accessed and run by a web browser. Not only would this enable D-VAMS to run on any conventional computer or device with a browser, but it would also allow it to be run as a locally installed service without requiring access to the internet.

D-VAMS was therefore written as a simple interface using HTML and Javascript/JQuery. Each scale comprises a page with a picture in the centre, and a vertical slider to its right. The vertical slider is accompanied by division markers much like the ones on a ruler, as a visual cue to clearly identify the interface as a measurement device.

The scales are bipolar, and the slider is set at the scale midpoint (50) by default, with scales all oriented such that the bottom end of the scale is negatively valenced (‘sad’, ‘distressed’, ‘afraid’ etc.) and the top end is positively valenced (‘happy’, ‘excited’, ‘peaceful’). The only exception to this was for the *Sleepy-Alert* scale; because this was theoretically valence-neutral, it was, instead aligned with the activation dimension, as the upward direction is taken to naturally connote positivity or greater amount. This is consistent with the findings of Kontou et al. (2012), in

which reversing the VAMS happy and energetic items such that 'happy' and 'energetic' faces were at the top of the scales, was found to improve the performance of the VAMS.

Above and below the picture are words representing the scale endpoints, such as "Happy" and "Sad", clearly labelling the scales for those who have some intact language ability. The picture is initially set to the midpoint image corresponding to the respective scale, as is the slider; as the slider is moved, the image is changed to the one corresponding to the point-position on the respective scale. When the slider is moved up, the expression on the face changes towards the positively valenced (or in the case of the *Sleepy-Alert* scale, more activated) end of the scale, and when it is moved downwards the expression changes towards the negatively valenced (or in the case of the *Sleepy-Alert* scale, less activated) end of the scale (see Appendix XVII - 1). The scale value appears as a greyed number at the bottom right of the page which dynamically changes according to the slider position.

A main menu allows the scales to be browsed and examined individually, or used as part of a mood assessment run in which all are completed and charted results returned, while a 'Scale Browser' button brings up a scale menu page in which the buttons for the scales are presented in two columns, one for each actor; pressing a button brings up a page for the respective scale. Each scale page scale (see Appendix XVII) contains a large image of the face representing the midpoint (scale value of 50) of that scale. To the right of the face is a vertical slider which can be adjusted using the touch-screen either by dragging the slider control, or by swiping up or down over the image. Alternatively, the slider can be controlled using a mouse pointer by dragging the slider lever up or down, clicking any point along the scale, or using the

mouse wheel while the pointer is over the image to adjust the slider position. The up and down arrows of a keyboard can also be used.

For an assessment run, two buttons are displayed at the base of the main page, specifying which faces are to be used, male or female. Upon starting a run, a user is presented, in order, with each of the scales for that actor. For an assessment run, however, beneath the slider is a button with a forward-arrow icon, which is greyed and disabled until the slider has been moved. Once the slider has been moved, the button turns green and become enabled, allowing the user to press it in order to continue to the next page. As a value is entered for each of the scale pages, a chevron in a status bar at the base of the page also turns green to indicate progress throughout the task.

Upon finishing the final scale, the main view of a results screen scale is displayed. This faces view shows a display of the faces chosen, below which is a series of traffic-light style of bar charts displaying the scales in both unipolar and bipolar format (see Appendix XVII - 2). In chart mode (activated by the middle button at the base of the page), the scale results are charted as vectors within the CMA, alongside the calculated, standardised mean vector and its parameters (see Appendix XVII - 3/4). Below this, DVAMS-D (depression) and DVAMS-A (anxiety) scores are displayed. These values are derived from this vector based on its magnitude and position within the circumplex, and its relation to the theoretical positions of anxiety and depression within this structure (around 150° and 190° respectively).

As detailed in 6.3.5, the calculation of the mean vector may take two forms. Bearing in mind that the scales are not perfectly bipolar, there are two ways of mathematically resolving the scales in this coordinate space. In the first, there is the assumption that though the endpoints of the scales may not be perfectly bipolar,

people notionally correct for this in their use of a given scale, treating it as though it were purely bipolar. For this case, complementary scale angles used to split the scales scores into their valence (*x*) and activation (*y*) components were artificially 'straightened' by splitting any discrepancy (angles above or below the 180° that would theoretically exist between them) and sharing the difference between them through a line of best fit.

In the second form, the complementary ends of the scales, such as *Afraid-Neutral* and *Calm-Neutral*, retain their absolute position given by the plot, and instead form 'bi-unipolar' scales consisting of the imperfectly opposite, separate unipolar scales that comprise them. In this unipolar arrangement, the data are treated as if they were from a number of unipolar scales each running from a position on the circumference to the centre of the CMA, and no adjustments are made to coerce them into a coordinate system wherein they are perfectly bipolar. By default, the unipolar coordinate system is used, and the DVAMS-D and DVAMS-A are calculated without any adjustment of scale item angles within the CMA to contrive each scale as exact bipolar opposites (see Appendix XVII - 3). However a button is available to toggle the views of the coordinate system between unipolar and bipolar, with this latter view greyed out to visually contrast the two results (see Appendix XVII - 4). In addition to the CMA based indices, a purely valence based score is also returned, in which the mean is simply taken of all of the scale scores. In an alternate version of this a value ('Mean^{-SA}') is returned which does not include the *Sleepy-Alert* scale, which is theoretically valence neutral. One might expect the mean 'Mean^{-SA}', which includes only the first six scales to fare better as a measure of valence (and therefore as a correlate of depression) than the one excluding the *Sleepy-Alert* scale; while the former all have a substantial valence component, this last scale loads strongly on

activation but has near zero valence, therefore should be omitted if we want an accurate mean valence metric.

A ‘chart’ button at the base of the display allows the user to toggle between the faces view and the chart view. In the faces view, the faces selected from each of the scales are presented together above a colour-coded bar-chart of scale scores using a traffic-light style coding system. Together, these views provide an at-a-glance profile of a person’s mood which can be interpreted without any understanding of written or spoken language.

7.2 Validating D-VAMS in a Sample of Stroke Survivors

Having designed two sets of 7-item scales (one female and one male actor), the next step was to test them in a suitably designed experiment in an appropriate cross-section of stroke survivors. This validation study was crucial, as it is essential to properly assess the suitability of any new instrument for the purpose for which it was devised, and to gain as clear a picture as possible as to its strengths and weaknesses.

7.2.1 Design.

7.2.1.1 Testing psychometric properties. A test of a psychometric instrument should normally involve the examination of at least three key properties: *construct validity*, *criterion validity*, and *reliability*.

The first, *construct validity*, relates to whether our scales are indeed measuring what they are supposed to be measuring. When somebody rates themselves on the *angry-peaceful* scale, for example, we need to be sure that they do indeed recognise the respective endpoint faces as denoting ‘angry’ and ‘peaceful’ mood states respectively. The findings of the judgement studies of Chapter 5 yielded 12 mood

words which have been shown to be systematically related to a set of corresponding facial expressions derived from them, however the scales comprising the corresponding faces now need to be properly validated.

One test of the construct validity of the face scales is to simply compare how people rate themselves on them to how they rate themselves on similar scales which use the mood words instead of morphing faces. If the scales are recognised as being conceptually equivalent in terms of the affective continua that they represent, then we should see high correlations between people's self-ratings on the face scales and those given for corresponding word scales. So, as a test of this first, key psychometric property, it was decided that the experimental task would involve the use of both types of scales presented in random order.

Another way of assessing the validity of an instrument is to examine how well the results correspond to those of other, validated measures already in use. If the depression or anxiety scores returned by the D-VAMS, for example, were to correlate well with a conventional measure of depression or anxiety, then this would offer strong evidence that they were measuring the same thing.

A *criterion* measure was therefore necessary, against which the D-VAMS could be compared. This measure would ideally be quick-to-administer, and capable of quantifying low mood in general, as well as quantifying elements relating to the experience of a depressed state. It would also need to be a measure suitable for use on patients with a range of physical symptoms, where it is important for any assessment of depressed or anxious mood not to be confounded by symptoms relating to the presence of physical illness.

The HADS meets all these criteria. It is a well validated and widely used instrument whose psychometric properties are well known. Despite being titled as a

‘hospital’ instrument, the key feature of the HADS is its focus on features of depression and anxiety that exclude symptoms which may be due to coexisting physical illness. Since it was specifically designed for use with clinical populations, HADS may be considered appropriate for use in people with a chronic or persistent medical disorder regardless of the immediate environment, and therefore suitable for use in community as well as inpatient settings. It is brief, quick to administer and performs as well as the longer GHQ or BDI instruments, and appears to perform well even in non-clinical populations (McDowell, 2006). Its suitability for use with stroke patients is attested to by its use in studies of depression following stroke (Lewis et al., 2001; B. S. Townend et al., 2007a; Ayerbe et al., 2011; Hackett & Pickles, 2014) and its adoption as a criterion measure for validation studies of assessment instruments in this population (Sutcliffe & Lincoln, 1998; Brumfitt & Sheeran, 1999; Bennett et al., 2006).

The HADS anxiety (HADS-A) and depression (HADS-D) subscales also have direct, theoretical equivalents in the form of the D-VAMS anxiety and depression indices (DVAMS-A and DVAMS-D) calculated from the circumplex model, so a comparison of HADS scores against the D-VAMS would provide a particularly strong test of convergent validity.

Finally, there is the question of the reliability of an instrument, that is, the consistency with which it measures what it is supposed to be measuring. A thermometer, for example, should consistently return a reading of 100°C when it is placed in boiling water; any significant fluctuations from this figure from one reading to the next – all else being equal – would indicate that the thermometer is not a very good one. Test–retest reliability is therefore something that is very important to assess

in a psychometric instrument, and this is usually done by simply administering the same test twice separated by an appropriate time interval.

The issue of what constitutes an appropriate time interval for assessments of self-rated mood is a contentious one, but for the purposes of this study we must only be reasonably sure that the ratings given for the D-VAMS scales are far enough apart that recollections of ratings returned for the first test do not unduly influence those given for the retest. Since it was established that the construct validity test would consist of both word and face versions of the scales presented in random order, it would seem reasonable that only a relatively small period of time should be required between test and retest, as the randomisation of the items for the retest would likely disrupt detailed memory of the earlier run.

It was important for practical purposes to confine the validation study to a single task that could be done in one sitting, but there should also be distinct break between the test and retest tasks that would allow some distance from the memory of the earlier responses. Since it has already been established that the validation study would also consist of responding to the HADS as a test of criterion validity, this suggested a simple, three-stage task that would enable all three properties to be studied simultaneously in an experiment comprising a final, validation study:

The first part of the task would consist of a random mixture of the 7 scales: word-based versions, with words such as “happy” and “sad” at the scale endpoints, and face versions (from the D-VAMS) where a facial expression is adjusted using a slider. These would be presented in random order on 14 consecutive pages.

The second part of the task would involve responding to 14 questions from the Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983). For the

final part, the first part of the task would be repeated, but with items in a different, random sequence.

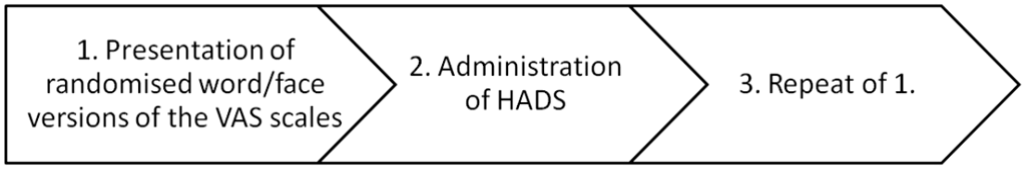


Figure 7.10. *Validation Study Protocol*

For the test and retest phases of the D-VAMS scales (word/face versions), DVAMS-D and DVAMS-A coefficients (for depression and anxiety respectively) would be computed, as well as D-VAMS Mean and Mean^{-SA} scores as described in section 7.1.2. For the second part, the HADS administration, the HADS-D and HADS-A subscale scores would be also be calculated. The objective of this validation study would be to examine the construct validity, criterion validity, and test-retest reliability of D-VAMS by examining correlations between the different measures made.

Construct validity would be assessed by examining correlations between the scores for the 'face' and 'word' versions of the scales during the test and retest phases. Criterion validity would be assessed using the Multitrait-Multimethod approach to examine the convergent and discriminant validity of HADS-D/HADS-A and DVAMS-D/DVAMS-A, as well as the D-VAMS Mean and Mean^{-SA} scores. Reliability would be assessed by examining correlations between D-VAMS scale scores from the test and retest phases. Content validity will also be examined from a qualitative and theoretical perspective.

7.2.1.2 Prior consultation. As part of consultation prior to implementing the study, a presentation was given to the Division of Rehabilitation and Ageing Protocol Planning Committee meeting (October 6th, 2014) to obtain feedback from academics and other professionals. A further presentation was then given to Nottingham Stroke Research Partnership Group (November 10th, 2014) to obtain input from stroke survivors, carers and medical and rehabilitation professionals. Feedback was taken from group members which was subsequently incorporated into the design and implementation of the validation study. The group stated that this research is worth pursuing and were supportive of it.

As a result of these consultations, two main amendments were made. A debriefing form was added to the online task, to thank participants, reiterate the importance of participation in research and offer an opportunity for the participant to offer feedback by way of a text box. It was also agreed that all participants would be followed up and informed of the results of the study once the analysis was complete. More importantly, a contingency plan was also developed in the event that participants returned HADS scores which raised concerns about their psychological wellbeing. In the event of a participant returning a HADS-D or HADS-A score of 11 or more, the participant would be contacted to discuss the score of concern, and the experimenter would offer to contact the participant's GP on their behalf to inform them of HADS results. Full details of these contingencies are described in the D-VAMS Working Practise Document (see Appendix XVI).

7.2.2 Recruitment and Participants. The next question was what type of people the scales should be tested on. Ideally, the study sample would consist of the kind of people with whom they D-VAMS would eventually be used, that is, people with absent or greatly impaired ability to use language following stroke. However in

order to properly assess a nonverbal mood assessment instrument of this kind, it must be validated by comparing responses on the D-VAMS to scores returned using conventional, language-based measures. This necessarily raises a paradox that cannot be fully resolved by any choice of sample characteristics, because people with impaired ability to use language will also be impaired in their ability to use language-based measures against which the instrument is being assessed. Some studies have attempted a compromise by employing people with aphasia who score well on single word recognition (Arruda et al., 1999), however this strategy is only suitable when the criterion measure against which a tool is being assessed consists purely of single words.

It was therefore not practicable to employ people with significant aphasia as part of the present study sample, however, the context of the use of D-VAMS suggested that participants should be stroke survivors at some stage following stroke. Ethics committee requirements demand that prospective participants must be adults capable of informed consent, so this was also stipulated as a requirement. Finally, fluent English was required because this was the language of the concurrent mood measures against which the D-VAMS was to be validated. The inclusion criteria were therefore that participants:

1) be English-speakers who are over 18 years of age
2) have had a stroke
3) be capable of giving informed consent.

(see Appendix XII)

In order for the findings of this study to be generalisable to people and settings in which the D-VAMS might be used, it was important to employ a reasonably

representative cross section of people from a variety of sources. One source of recruitment – and one that is increasingly influential – is via the internet. Since the tasks comprising the study were fully automated via the online, D-VAMS project portal and did not require supervision, permission was sought for advertisements to be posted on the Stroke Association 'Talkstroke' forum and the brain injury research section of the Headway website. A Twitter account 'Strokewellbeing' was set up and used to socially network with numerous online stroke groups and individual campaigners and encourage snowball sampling. Announcements were frequently made through the Twitter feed and publicised further by 'retweets' through the network. Advertisement and announcements were also made on relevant Facebook pages and in forums of other online communities.

Another source of recruitment was through local community stroke groups. A number of groups in the Nottingham area were contacted, and appointments made to attend meetings and deliver a short presentation about the study, followed by an invitation for those interested in taking part to be contacted by the researcher. Participating groups included the Nottingham Stroke Club, Sherwood Stroke Club, Mansfield and District Stroke, Club, and Ashfield Stroke Group.

People in stroke clubs, however, are not representative of people who have had a stroke, as they are generally much further into the recovery process, and are unlikely to contain people who have only very recently had a stroke. In any correlational analysis, it is very important to have a diverse spread of scores; if there are too many people with high scores, or too many people with low scores, it can attenuate any resulting correlations. It can also make it difficult to derive accurate measures of sensitivity and specificity and to determine suitable cut-off points.

Since D-VAMS should be suitable as a screening instrument for depression in people in the early stages of stroke, as well as later, it was decided to recruit a third group from within a rehabilitation setting. This would make for a more representative group of people, and enable scores to be obtained from people whose mood may be lower. The combination of data from these three sub-groups should offer a reasonably representative cross-section of stroke survivors that reflect the general population. Recruitment was therefore also pursued with the cooperation with staff at the Citycare Partnership, an NHS Community Stroke Discharge and Rehabilitation Service.

Ethics Committee approval was granted by the Faculty of Medicine and Health Sciences Research Ethics Committee, University of Nottingham (Ref : I10102013). Ethics Committee approval was also sought with the NUH via NHS proportionate review and R&D review, and approved by West of Scotland Research Ethics Service (WoSRES) (Ref: 15/WS/0239), with a research passport being issued to allow the researcher access to Citycare Partnership.

To estimate sample sizes, power calculations for this experiment were performed using the G*Power software package (Faul et al., 2007). Calculated values were studied based on low-end (dz=0.35) to high-end (dz=0.45), medium to large effect sizes, for a power of 0.8 and an α of 0.05. An appropriate target sample size for the validation study was estimated to be in the range of *n=26* to *n=46*, with the aim of recruiting approximately 40 participants across the three recruitment methods.

Table 5.0. *Sample sizes from study power calculations*

	dz=0.35	dz=0.4	dz=0.45
Power = 0.8	46	34	26

In total 46 participants (28 male, 18 female) were recruited by the means described above: 15 (33%) responded to online canvassing, 20 (33%) were recruited via local stroke clubs, and 11 (24%) were recruited via the NHS.

7.2.3 Method. For participants recruited via the internet (*n*=15), the task was completed online and unsupervised through the D-VAMS Project Portal. Participants recruited from stroke clubs were offered a choice of taking part in the experiment either at some point during their visit to the premises (the stroke club venue) or during a home visit by the researcher. Participants recruited via the NHS were first approached by staff at Citycare Partnership and given a Participant Information Sheet (Appendix VIII) explaining the nature and purpose of the study. If the service user was interested in participating, they were then provided with a consent-to-contact form (Appendix XIV) in which they were asked to provide their name and a contact phone number or email, and to sign and date the form at the bottom. These details were then passed on to the principal investigator, who contacted prospective participants to arrange a home visit at a later date.

When home visits were conducted, they were conducted and documented as per UoN lone working policy. When the task was completed during a home visit or at a stroke club, it was done using a Nexus 10 tablet under the supervision of the experimenter. The tablet was equipped with a wireless, 4G, broadband router, to give remote access to the D-VAMS Project Portal from wherever the task was conducted. If a participant expressed difficulty reading any of the paperwork or on-screen instructions or other text, then this was read out loud by the experimenter.

Participants were first given (or, if online, directed to) a Participant Information Sheet to read. Participants recruited online or via stroke groups received the School of Medicine version (Appendix XII), while participants recruited via the

NHS received one in NHS format (Appendix XIII). Participants were then asked to provide details of their gender, year of birth, ethnicity and time elapsed since their stroke, and to check boxes on an online consent form (Appendix V and Appendix XV). NHS participants were asked to also sign a paper copy (Appendix XV).

The task was then completed as described in Figure 7.10. The first part of the task consisted of completing a random mixture of the seven, face-based scales, along with seven, equivalent, word-based scales. Whereas in the D-VAMS face scales a facial expression was adjusted using a slider, the word-based scales differed in that there was no image of a face, only words describing the scale endpoints. For the *Sad-Happy* scale, for example the word "happy" was displayed at the top of the page, and the "sad" was at the bottom, with the slider being used to indicate where between these extremes a person considered themselves to be, mood-wise. The gender of the face scales used during the task were assigned at random by the project portal, and remained constant from the test to the retest phases of the construct validation task.

These scales were presented in random order on 14 consecutive pages. To set a value, the participant dragged a slider to a point on the scale representing their mood over the course of the last week. If the participant expressed difficulty adjusting the slider, then the experimenter assisted, asking, “Up or down?”, adjusting the slider position until the participant was satisfied. Once the slider had been moved, a ‘next page’ button lighted up green and became enabled, allowing the participant to proceed to the next page. Whenever a face scale was presented, the experimenter first manually moved the slider the length of the scale from one end to the other and back again, to clearly demonstrate the endpoint facial expressions delineating the scale. Once a face scale had been demonstrated, the slider was then set back to its midpoint on the respective scale, and the participant was asked to select a suitable face.

Once this first run had been completed, the Hospital Anxiety and Depression Scale (HADS) was completed in a page-per-question format, with the instructor assisting if the participant had problems selecting their chosen option or the 'next page' button. For the final part, the first part of the task was repeated, but with items presented in a different, random sequence. When the task was complete, a debriefing ("thank you") page was presented, giving positive feedback and providing a field on the form where the participant could give feedback if they wished.

7.2.4 Results. Scale scores were collated and divided into groups so they could be analysed by run (test/retest), type of scale (word/face) and recruitment group. DVAMS-D and DVAMS-A metrics (for depression and anxiety respectively) were computed for all participants, along with D-VAMS Mean and Mean^{-SA} scores. The former were computed from the valence-activation coordinate system as described in 7.1.1, while the latter rendered them as simple averages reflecting overall valence. The second of these mean scores excluded the seventh, *Sleepy-Alert* scale ('Mean^{-SA}'), which is valence-neutral (see 7.1.2). These data were first examined to assess their distribution and decide what kinds of tests of significance would be appropriate. Descriptives yielded no evidence of skew or kurtosis of the examined data, and a Q-Q plot of D-VAMS scores was consistent with normally distributed data, therefore parametric tests were chosen.

7.2.4.1 Sample Characteristics. Sample characteristics are summarised in Table 7.3. There were more males than females (28 male, vs. 18 female), and online participants were generally much younger than stroke club/rehabilitation groups (48 years on average, vs. about 70 years for stroke club/rehabilitation group). Stroke club participants had the highest time elapsed since stroke (about 5 years on average), followed by the online group (about 3 years) and rehabilitation group (7 months).

Table 7.3. *Summary Sample Characteristics*

	Online	Stroke Club	Rehabilitation	ALL
n	15 [10♂/5♀]	20 [13♂/7♀]	11 [5♂/6♀]	46 [28♂/18♀]
Age	48.1 (*9.3*)	72.2 (*9.7*)	70.5 (*11.4*)	63.8 (*14.7*)
Time since stroke (yrs)	3.0 (*1.6*)	5.2 (*4.5*)	0.6 (*0.4*)	3.4 (*3.6*)
HADS-D	8.7 (*3.4*)	5.2 (*3.3*)	7.5 (*4.4*)	6.9 (*4.0*)
HADS-A	9.4 (*4.8*)	7.0 (*4.2*)	9.3 (*3.7*)	8.4 (*4.7*)
DVAMS-D	-8.0 (*30.6*)	-33.9 (*28.9*)	-3.0 (*31.1*)	-18.1 (*32.6*)
DVAMS-A	-13.5 (*26.1*)	-23.8 (*21.3*)	-3.9 (*19.4*)	-16.2 (*23.2*)
D-VAMS	54.5 (*20*)	71 (*18.2*)	50.9 (*20.4*)	60.8 (*21.0*)
D-VAMS^{-SA}	52.5 (*18.9*)	71 (*18.1*)	52.1 (*19.4*)	60.5 (*20.5*)
HADS-D≥11	47%	5%	18%	22%

Mean *(S.D.)*

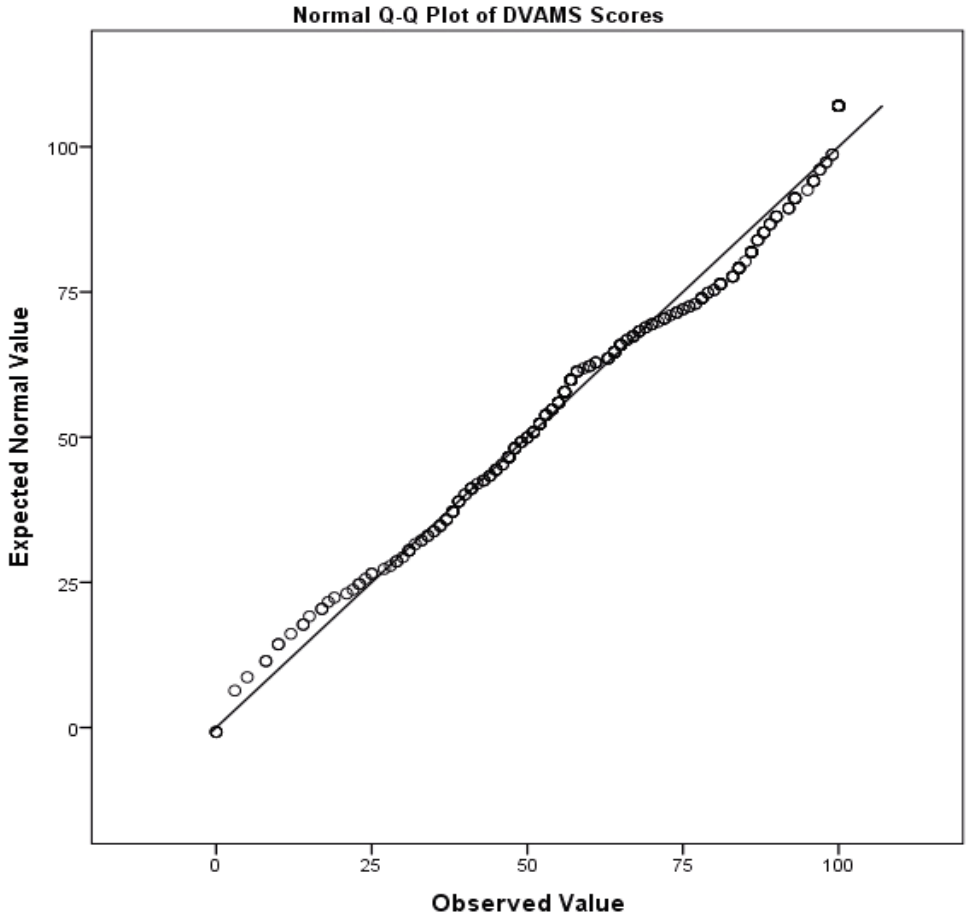


Figure 7.11. *Q-Q Plot of D-VAMS (Face) Scores, Run 2*

The rehabilitation & online groups generally had much higher HADS-D scores than the stroke club group. The online group scores were highest (8.7 mean), followed by the rehabilitation group (7.5), compared to 5.2 for the stroke club group. The online group had the highest proportion of 'depressed' people (nearly half) as flagged by the 'scores of concern' criteria of HADS≥11, followed by the rehab group, with nearly 20%. Only one person from the Stroke Club group scored as depressed.

7.2.4.2 Construct validity. Construct validity was assessed by comparing the scores returned for the face scales to those returned by the corresponding word versions. Though the repeat run was performed primarily for the purposes of test-retest reliability, it also provided an opportunity to examine construct validity. The data for the scale judgements were also analysed separately by scale, as well as combined into Run1/Run 2 data, and again with data from both runs combined. At each of these levels (scale, run and total) Pearson's *r* correlation coefficients were computed between word and face scale scores to arrive at figures for construct validity. The results are presented in Table 7.4.

Table 7.4. *Construct Validity by Total, Run and Scale*

	Run 1	*Run 2*	
Miserable–Satisfied	.78	.76	
Sad–Happy	.67	.79	
Distressed–Peaceful	.68	.76	
Bored–Excited	.62	.70	
Afraid–Calm	.66	.79	
Angry–Peaceful	.40	.79	
Sleepy–Alert	.40	.73	
Total by Run	.59	.76	*Total = .66*

Significant at the 0.01 level (1-tailed).

Construct validity from all the pooled data was .66. Construct validity for Run 1 varied from moderate to very good, at .59 overall, while construct validity for Run 2 was uniformly very good, at .76 overall.

Table 7.5. *Between-scale Correlations for Face Scales (Run 2)*

	Scale 1 *Mis-Sat*	Scale 2 *Sad-Hap*	Scale 3 *Dis-Pea*	Scale 4 *Bor-Exc*	Scale 5 *Afr-Cal*	Scale 6 *Ang-Pea*	Scale 7 *Sle-Ale*
Scale 1	1	.77	.73	.77	.76	.70	.73
Scale 2		1	.74	.88	.77	.73	.73
Scale 3			1	.71	.79	.81	.58
Scale 4				1	.70	.72	.77
Scale 5					1	.75	.62
Scale 6						1	.72
Scale 7							1

Significant at the 0.01 level (1-tailed).

Discriminant validity between the scales, however, was extremely poor, as evidenced by the consistently high cross-correlations between scales (see Table 7.5). Internal consistency was very high, with Cronbach's α values of over .9 for both face and word version of the scales across both runs (see Table 7.6).

Table 7.6. *Cronbach's Alpha for Face/Word D-VAMS and D-VAMS^{-SA}*

	Run 1		*Run 2*	
	DVAMS	DVAMS^{-SA}	DVAMS	DVAMS^{-SA}
Face	.920	. 933	.950	.948
Word	.920	.918	.951	.954

To examine the factor structure of D-VAMS scores, a PCA analysis was performed on data from both runs of both versions of the scales. Only one significant factor was extracted, with the remainder having eigenvalues below the commonly used Kaiser criterion of 1.0, and – by extension – any eigenvalue cut-off derived by the more stringent parallel forms method (see 5.4.5). This single factor is consistent with valence, and accounts for a very high proportion of the variance (see Table 7.7a). Factor loadings by scale and type are shown in Table 7.7b.

Table 7.7a. *% of Variance Accounted for by Factor 1 - Face/Word DVAMS and DVAMS^{-SA}*

	Run 1		*Run 2*	
	DVAMS	DVAMS^{-SA}	DVAMS	DVAMS^{-SA}
Face	69	76	77	80
Word	68	71	78	82

Table 7.7b. *Factor Loadings for Factor 1 by Scale - Face/Word DVAMS and DVAMS^{-SA}*

		Run 1		*Run 2*	
		DVAMS	DVAMS^{-SA}	DVAMS	DVAMS^{-SA}
Face	*Mis-Sat*	.940	.944	.887	.884
	Sad-Hap	.921	.908	.914	.915
	Dis-Pea	.839	.847	.870	.892
	Bor-Exc	.924	.917	.900	.892
	Afr-Cal	.839	.854	.874	.889
	Ang-Pea	.699	.730	.880	.878
	Sle-Ale	.578	-	.833	-
Word	*Mis-Sat*	.887	.863	.935	.941
	Sad-Hap	.876	.867	.939	.948
	Dis-Pea	.888	.901	.936	.955
	Bor-Exc	.756	.777	.761	.709
	Afr-Cal	.768	.786	.901	.917
	Ang-Pea	.843	.863	.906	.926
	Sle-Ale	.746	-	.783	-

7.2.4.3 Criterion validity. An important means of assessing convergent validity is to examine how D-VAMS scores compare to those of the criterion instrument of choice, the HADS. First the DVAMS-D and DVAMS-A scores were compared to those of the HADS-D and HADS-A. Bearing in mind that two different renditions of the coordinate system exist, unipolar and bipolar, two different calculations are possible for these scores. To assess criterion validity across runs of the D-VAMS under both of these renditions, Pearson's *r* correlations were performed between the HADS-D/HADS-A scores, and DVAMS-D/DVAMS-A scores of the two runs, first using the unipolar rendition, and then using the bipolar one.

Correlations of the HADS and D-VAMS depression/anxiety metrics yielded two sets of multitrait-multimethod matrices for each calculation method, shown in Table 7.8a and Table 7.8b. The results were very similar for both coordinate systems, and made little difference to the correlations found.

Examining the cross-correlations between the D-VAMS and the HADS, high correlations were noted between HADS-D and DVAMS-D scores, demonstrating very good convergent validity between the respective depression metrics. Cross-correlations between the HADS-D and DVAMS-A, and the HADS-A and DVAMS-D were also significant, though substantially lower. Correlations between HADS-A and DVAMS-A, though significant, were poorest.

Table 7.8a. *D-VAMS–HADS Correlations by Run and Subtype - Unipolar Coordinate System*

	Run 1		*Run 2*	
	DVAMS-D	DVAMS-A	DVAMS-D	DVAMS-A
HADS-D	.70	.56	.71	.56
HADS-A	.60	.46	.70	.55

Significant at the 0.01 level (1-tailed).

Table 7.8b. *D-VAMS–HADS Correlations by Run and Subtype - Bipolar Coordinate System*

	Run 1		*Run 2*	
	DVAMS-D	DVAMS-A	DVAMS-D	DVAMS-A
HADS-D	.69	.58	.72	.55
HADS-A	.60	.47	.70	.56

Significant at the 0.01 level (1-tailed).

Finally, correlations were examined using only the D-VAMS mean scale scores. Since the D-VAMS scales run from negative to positive valence on a scale of 0 to 100, higher scores mean a positive mood, therefore the predicted direction of the correlation was negative, since low HADS-D scores indicate positive mood, and high scores indicate negative mood. Correlations between HADS-D and D-VAMS (scale scores and means) are shown in Table 7.9. Of note, also, was the high correlation between the HADS-D and HADS-A scores (r=.64, p<0.001).

Table 7.9. *DVAMS/HADS-D Correlations*

	Run 1		*Run 2*	
	Face	*Word*	Face	*Word*
Miserable–Satisfied	-.64	-.*66*	-.63	-.*72*
Sad–Happy	-.72	-.*69*	-.73	-.*68*
Distressed–Peaceful	-.59	-.*76*	-.51	-.*74*
Bored–Excited	-.68	-.*68*	-.73	-.*70*
Afraid–Calm	-.53	-.*65*	-.61	-.*64*
Angry–Peaceful	-.41	-.*81*	-.56	-.*67*
Sleepy–Alert	-.50	-.*53*	-.71	-.*62*
D-VAMS Mean	**-.59**	-.*68*	**-.70**	-.*70*
D-VAMS Mean^{-SA}	**-.72**	-.*72*	**-.73**	-.*73*

Significant at the 0.01 level (1-tailed).

7.2.4.4 Test-retest reliability. Tests of reliability were then performed on both the individual scale scores and the summary metrics. Again, Pearson's *r* correlations were performed between D-VAMS scale data for the test and retest conditions.

Table 7.10. *Test-retest Reliability (Face Scale Scores)*

	Run 1/2
Miserable–Satisfied	.79
Sad–Happy	.81
Distressed–Peaceful	.71
Bored–Excited	.84
Afraid–Calm	.75
Angry–Peaceful	.71
Sleepy–Alert	.62
Total	.71

Significant at the 0.01 level (1-tailed).

The D-VAMS scale scores individually had good to excellent reliability, with a total correlation of .71 between data from test and retest runs (see Table 7.10). The reliabilities of the combined metrics, were of course, better. The DVAMS-D and DVAMS-A had a reliability of .89, and .80 respectively, while the D-VAMS Mean and Mean^{-SA} reliabilities were .89 and .90 (see Table 7.11).

Table 7.11. *Test-retest Validity (D-VAMS Summary Metrics)*

	Run 1/2
DVAMS-D	.90
DVAMS-A	.84
D-VAMS Mean^{-SA}	.91
D-VAMS Mean	.91

Significant at the 0.01 level (1-tailed).

7.2.4.5 Sensitivity and specificity. High correlations between a scale under investigation and a criterion measure, however, do not alone guarantee that it will perform well as a screening instrument. We must also assess its ability to correctly identify positive and negative cases (i.e: 'depressed' and 'not depressed') identified by the criterion instrument based on various cut-off values.

HADS-D cut-offs recommended for stroke patients range from ≥4 (Sagen et al., 2009) to ≥7 (Aben et al., 2002) (Lincoln et al., 2012b). Receiver operating characteristics (ROC) were therefore calculated for both the DVAMS-D and Mean^{-SA} statistics against HADS-D cut-offs of 4–8. Optimal cut-offs for Run 2 are listed in Table 7.12a and Table 7.12b and ROC curves for HADS-D cut-offs 4–7 are shown in Figure 7.1.2a and Figure 7.12b.

A comprehensive analysis was also performed on data including individual scale scores for Run 2. ROC curves for scale scores at HADS-D cut-offs 4–7 are shown in Figure 7.13.

Table 7.12a. *Sensitivity/Specificity Cut-offs: DVAMS-D against HADS-D*

	DVAMS-D			
	ROCS AUC	Cut-off	Sens %	Spec %
HADS-D ≥ 4	86.8%	≥ -45	89	82
HADS-D ≥ 5	88.2%	≥ -45	97	70
HADS-D ≥ 6	89.5%	≥ -30	93	79
HADS-D ≥ 7	81.6%	≥ -23	83	65
HADS-D ≥ 8	80.5%	≥ -14	81	68

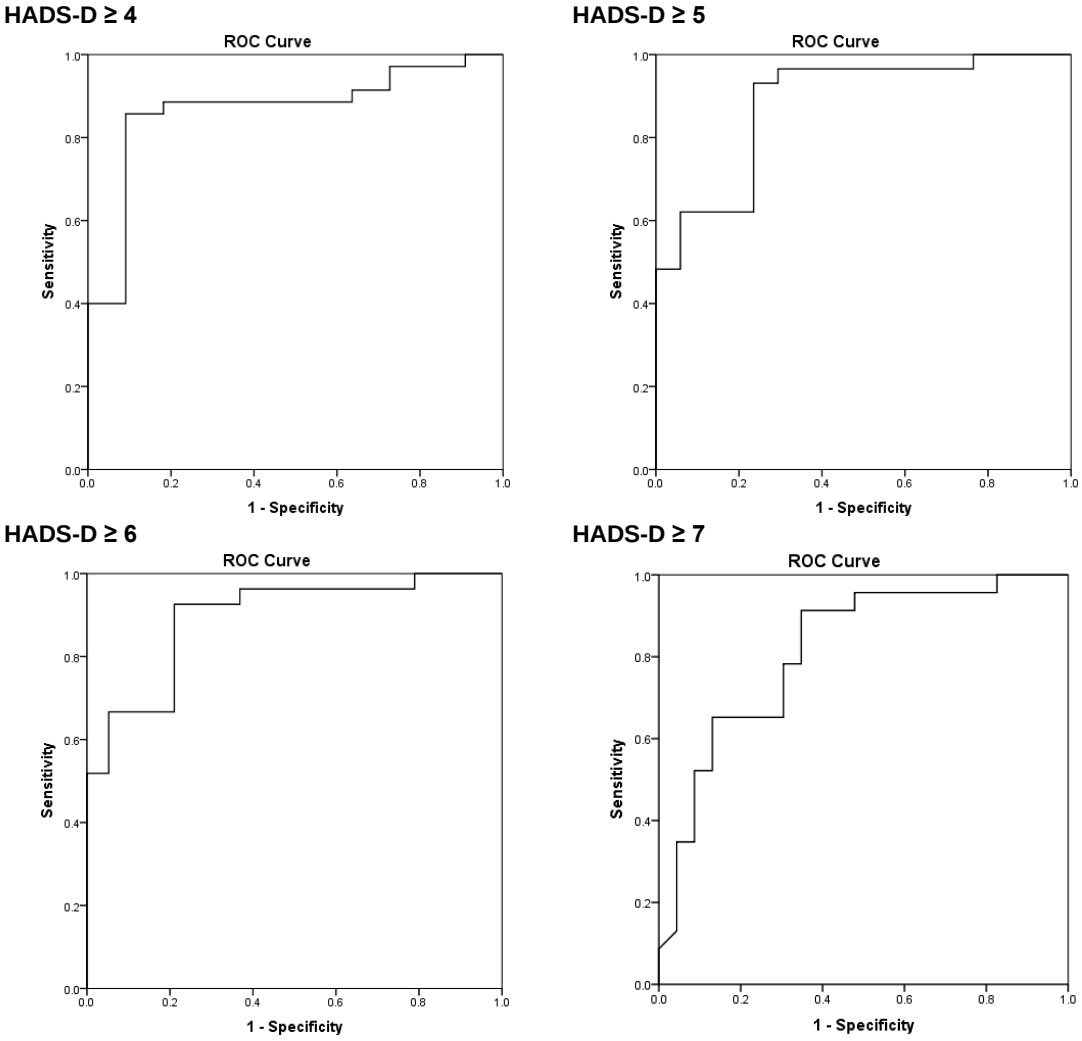


Figure 7.1.2a. *ROC Curves for DVAMS-D against HADS-D Cut-offs 4 –7*

Table 7.12b. *Sensitivity/Specificity Cut-offs: D-VAMS Mean^{-SA} against HADS-D*

	D-VAMS Mean^{-SA}			
	ROCS AUC	Cut-off	Sens %	Spec %
HADS-D ≥ 4	87.8%	≤ 74	89	91
HADS-D ≥ 5	89.5%	≤ 74	97	77
HADS-D ≥ 6	91.0%	≤ 69	96	79
HADS-D ≥ 7	83.6%	≤ 63	83	65
HADS-D ≥ 8	82.1%	≤ 59	81	68

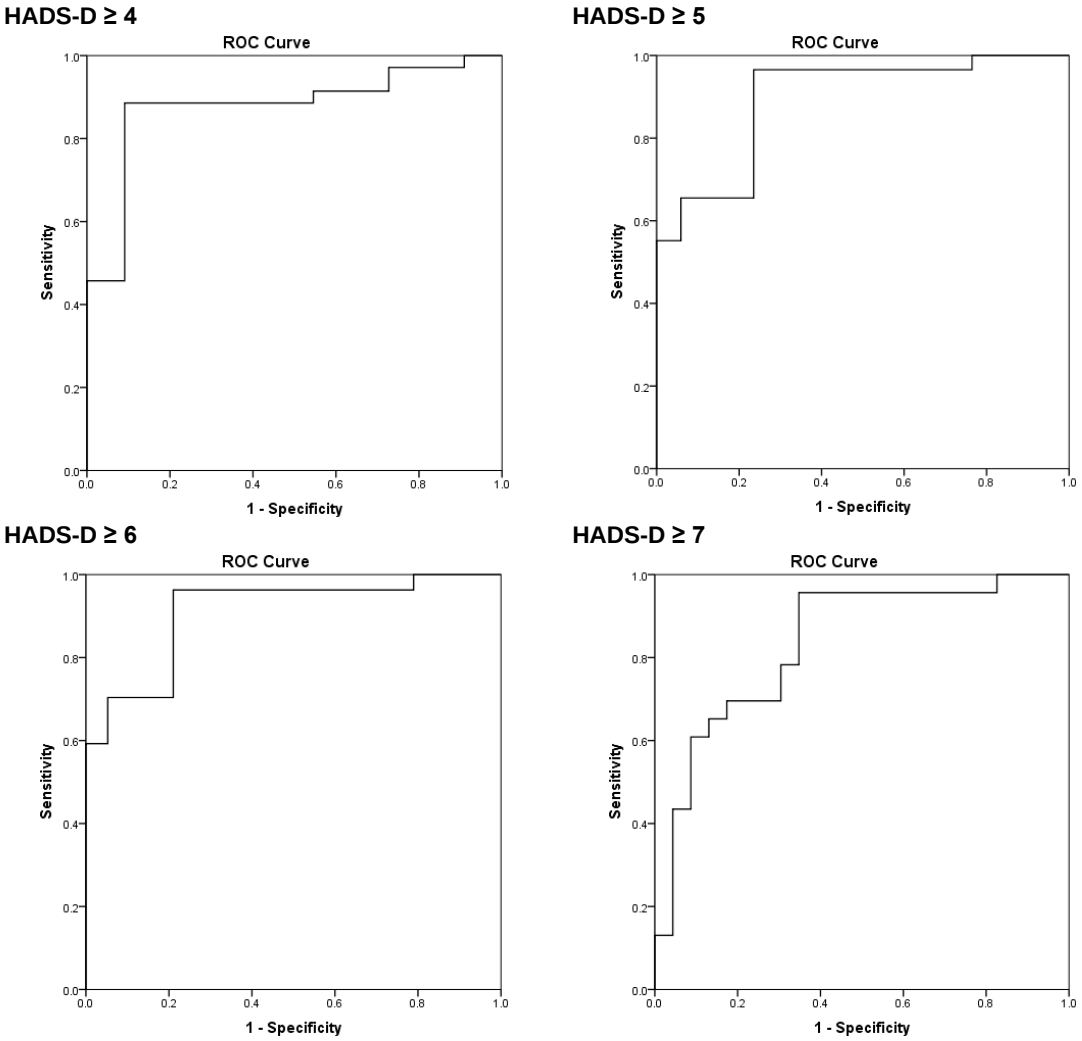
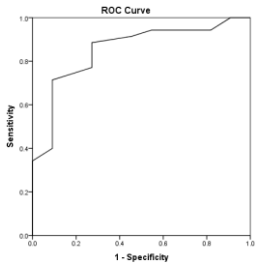


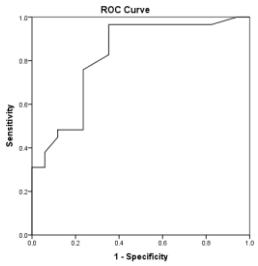
Figure 7.12b. *ROC Curves for D-VAMS Mean^{-SA} against HADS-D Cut-offs 4 –7*

Scale 1 – *Miserable-Satisfied*

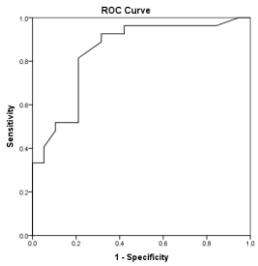
HADS-D ≥ 4



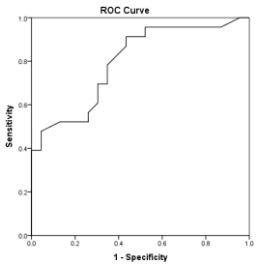
HADS-D ≥ 5



HADS-D ≥ 6

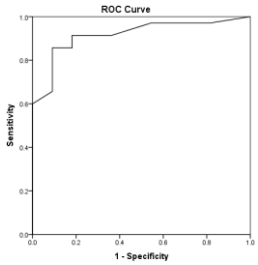


HADS-D ≥ 7

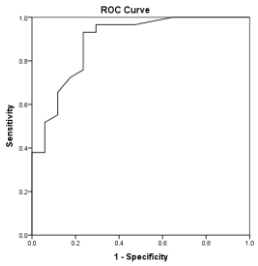


Scale 2 – *Sad-Happy*

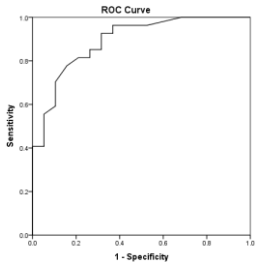
HADS-D ≥ 4



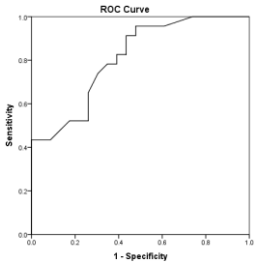
HADS-D ≥ 5



HADS-D ≥ 6

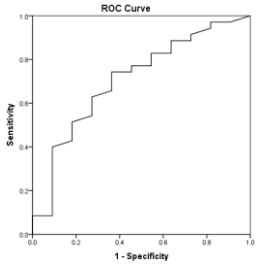


HADS-D ≥ 7

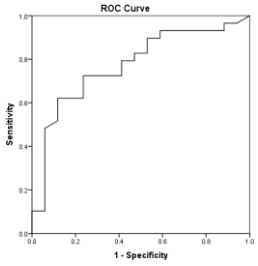


Scale 3 – *Distressed-Peaceful*

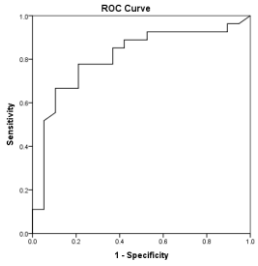
HADS-D ≥ 4



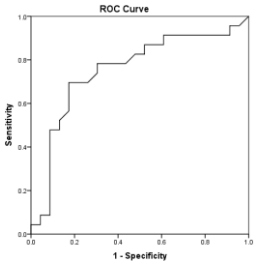
HADS-D ≥ 5



HADS-D ≥ 6

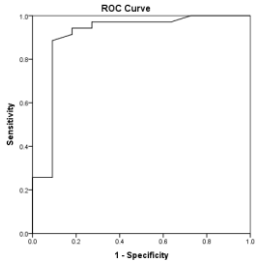


HADS-D ≥ 7

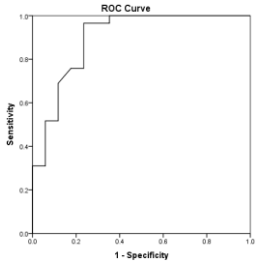


Scale 4 – *Bored-Excited*

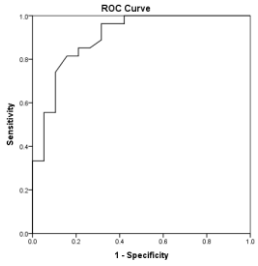
HADS-D ≥ 4



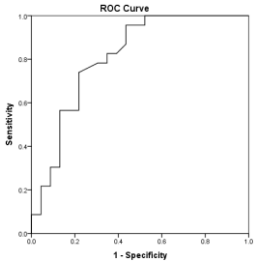
HADS-D ≥ 5



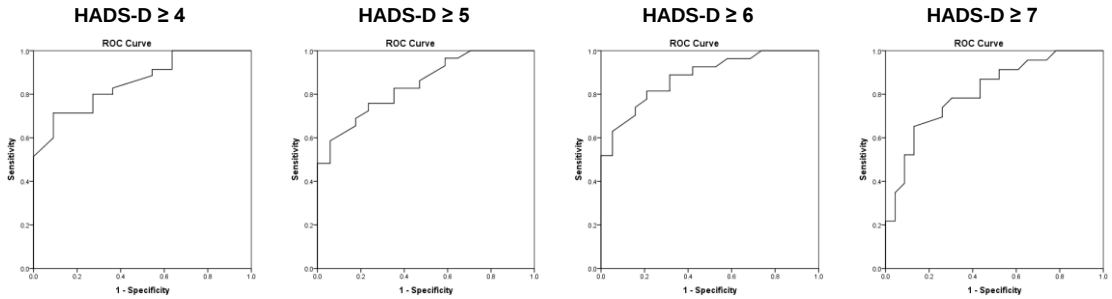
HADS-D ≥ 6



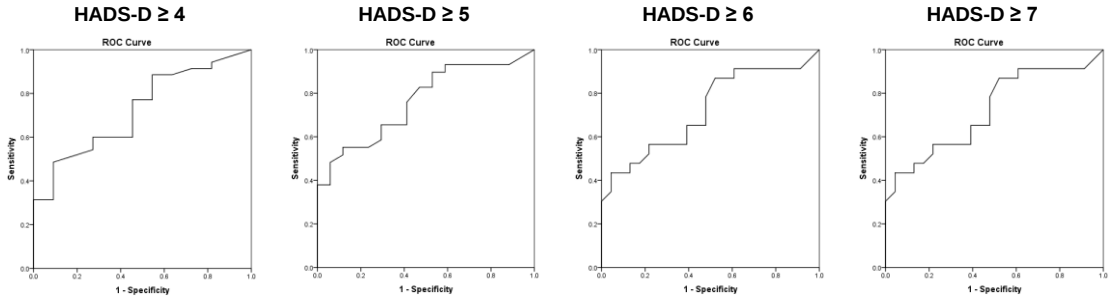
HADS-D ≥ 7



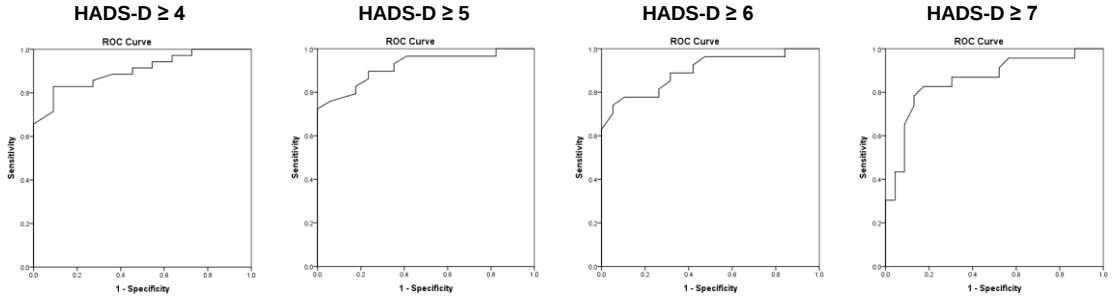
Scale 5 – *Afraid-Calm*



Scale 6 – *Angry-Peaceful*



Scale 7 – *Sleepy-Alert*



Mean^{-sa}

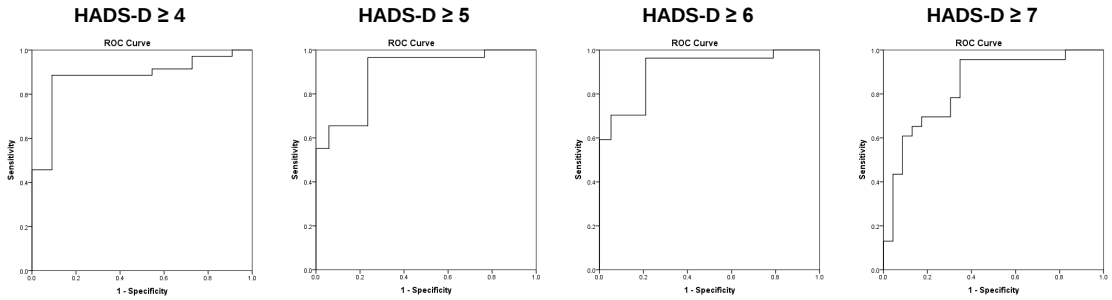


Figure 7.13. *ROC Curves for D-VAMS Scales and Mean^{-SA}*

7.2.5.6 Content Validity. Another key aspect of validity is the extent to which a psychometric instrument covers all the facets of the construct being measured. This is not something that can be calculated, evaluation of this entails a brief, qualitative examination with respect to the theoretical issues covered at length in Chapters 4 & 5.

An evaluation of the content of the D-VAMS should be made with respect to its value both as a mood assessment instrument and as a depression screening instrument. As a mood assessment instrument, content validity is strong, as it covers all the most important emotions and moods. From a categorical perspective, all of the basic emotions of (Ekman, 1970; Ekman & Friesen, 1971; Ekman & Friesen, 1976; Ekman, 2005) except for disgust are covered, with *Happiness*, *Sadness*, *Fear*, *Anger*, and *Surprise* all having a place within the CMA. However, disgust is a uniquely physiological response rooted in a repulsion to a noxious stimulus. Though fear and anger also possess a strong behavioural component vis-à-vis fight-or-flight behaviour (in the form of prototypical emotional episodes), they also have analogues as persistent moods removed from any immediate threat (such as with hostility or anxiety). Disgust, however, when removed from this literal sense, becomes something more akin to anger or dislike, and does not fit as cleanly into this taxonomy.

From a dimensional perspective D-VAMS covers the key factors that define Core Affect. Though, necessarily, the underlying CMA model used lacks the third dimension of *Dominance-Submissiveness* which complex emotions such as *Fear* and *Anger* also embody, these emotions are too important to omit from even a basic assessment of mood and so must stand despite their arguably more complex factorial makeup. The selection of these scale items therefore offers a delineation that is – content-wise – compatible with a slightly more complex dimensional structure, where a two-dimensional circumplex structure is crossed in its negative valence, negative

activation quadrant by a *Dominance-Submissiveness* dimension spanning 'afraid', 'distressed' and 'angry'. This fusion of two and three-dimensional domains seems to represent a critical junction from which taxonomies of emotions are drawn, as is attested to by the strong overlap between the items included in D-VAMS and the 'basic' emotions of Ekman and others.

As a depression screening instrument, however – as with other instruments that assess depression in the context of physical illness – the D-VAMS is limited purely to its most salient phenomenology, that is, sadness or depressed mood. While such symptoms predominate in the DSM-V criteria for depression, its somatic components, including fatigue, sleep or appetite problems, psychomotor retardation and agitation are necessarily missing. Also missing are the nuanced cognitive elements, such as feelings of hopelessness, guilt and morbid preoccupation with death. A similar simplification was also adopted in devising the content of HADS, yet this instrument still performed very well when compared to more comprehensive, standard measures of depression (McDowell, 2006).

7.2.6 Discussion.

7.2.6.1 Psychometric qualities of D-VAMS. With respect to construct validity, the size of Cronbach's α, the consistently strong, cross-scale correlations, the absence of discriminant validity between the face scales scores (see Table 7.5), and the results of the PCA indicate that D-VAMS scores are made up of only one significant factor, corresponding to 'pleasantness/unpleasantness', and accounting for around 80% of the variance (see Table 7.7). This is disappointing regarding the value of the coordinate system tested as part of this study, but revealing in the extent to which valence dominates the way that participants use D-VAMS to report mood, and supports Feldman's (1995) conclusions that when making judgements of their mood, people

weigh the valence dimension much more than the arousal dimension than the geometry of the CMA suggests. The evidence here demonstrates clearly that D-VAMS score essentially reflected whether the respondent's mood was positive or negative. It is likely that using scales that tap different proportions of valence and activation simply resulted in participants responding to the more dominant factor, with cues for affective valence superceding or masking those for activation in determining their response. Construct validity for all scales improved markedly from the first run to the second, suggesting a marked practise effect

Turning to criterion validity with the HADS, we see this unidimensional structure at work in the HADS and the D-VAMS-D/A correlations (see Table 7.8a & Table 7.8b). Though DVAMS-D/HADS-D correlations were good, the DVAMS-A statistic was not successful as a specific correlate of the HADS-A. Cross correlations between the DVAMS-D and HADS-A, and the DVAMS-A and HADS-D, substantially exceeded the DVAMS-A/HADS-A correlations in Run 1, and HADS-A/HADS-D correlations with DVAMS-A were uniformly worse than those with DVAMS-D. This shows a failure of both convergent and divergent validity with respect to the D-VAMS anxiety metric, underlining the apparent absence of the aforementioned activation factor in scores returned using D-VAMS.

The pattern of correlations of D-VAMS scores with the HADS-D, individually and in combination, further supports this (see Table 7.9), with consistently significant and generally high correlations across all of the scales, and no pattern of correlations favouring valence-heavy scales. The fact that the Mean^{-SA} statistic – in which the *Sleepy-Alert* scale was omitted – was a slightly better correlate of the HADS-D than the DVAMS-D metric derived from the CMA coordinate system is also consistent with this. The correlation between the HADS-D and HADS-A in this experiment was

also typically high (r=.64), making the challenge of distinguishing between depression and anxiety difficult even if a clear activation dimension were to be present. These results, however, may be due to the fact that this sample consisted of people whose depression and anxiety were highly correlated. If a larger sample of participants were to be separated into those who had high HADS-D scores but low HADS-A scores, and those who had high HADS-A scores but low HADS-D scores, it a two-factor structure may have been more evident in their D-VAMS scores.

DVAMS-D and DVAMS Mean/ Mean^{-SA} correlations against HADS-D also improved from the first run to the second, again suggesting a practice effect. The fact that D-VAMS/HADS-D correlations are very similar between the face and word versions of the scales also attests to the construct validity in that they largely agree with one another. Test-retest reliability for the scales individually was good, and excellent for the total scores (see Table 7.10).

A number of suitable cut-offs were found that offered good to excellent sensitivity and specificity against the HADS-D (see Table 7.2b and Figure 7.12b). Bennett et al. (2006) recommend a sensitivity of at least 0.8 and a specificity of at least 0.6 as an acceptable cut-off for the purposes of a screening measure. In this respect, both the DVAMS-D and DVAMS Mean^{-SA} metrics performed very well, with all cut-offs except HADS-D≥10 satisfying these criteria. Sensitivity and specificity generally improved with higher HADS-D cut-offs with an optimal cut-off HADS-D≥13 yielding 100% and 89% respectively.

The high Cronbach's α indicates that the seven scales can be reduced in number. Since D-VAMS performs primarily as measure of valence, and Scale 7 (*Sleepy-Alert*) was charted as close to valence-neutral in the PCA plots from Experiment 3 (see Figure 6.3 and Figure 6.4), this activation based scale seems like

the prime candidate for exclusion from the final version. The evidence suggests that excluding Scale 7 – as with the Mean^{-SA} metric – improves the psychometric qualities of D-VAMS; the D-VAMS /HADS-D correlations using the Mean^{-SA} statistic were uniformly better than those using the total mean score in both runs, and much better for the first run (D-VAMS Mean/HADS-D, r=.59; Mean^{-SA} /HADS-D, r=.70). Construct validity for Scale 7 was also much poorer for Run 1 (r=.40) and reliability was the poorest of the scales (r=.64); Scale 7 was also the scale which, when omitted, resulted in the smallest reduction of Cronbach's α (Scale 1-6 α=.948, compared to α=.950 for all scales). The findings here therefore suggest that Scale 7 is not required and should be dropped, and the mean score of scales 1-6 (Mean^{-SA}) adopted as the *de facto* D-VAMS score total, giving a measure of pleasantness of mood in the range of 0 to 100.

7.2.6.2 Strengths of the study. The present study sample comprised a fairly representative cross-section of stroke survivors employing participants from three different sources, – from online, from stroke clubs and from an NHS rehabilitation setting. As such, it was a fairly representative sample of stroke survivors, reflecting the balance of genders, range of ages, and time elapsed since stroke (see Table 7.3).

The design was elegant, with a simple task structure which enabled a variety of psychometric properties to be tested and evaluated in a reasonably short sitting. It was also designed with ease and flexibility of administration in mind, with a custom, project portal developed to administer the tasks automatically. The task could thereby be conducted on any device with an internet browser, and administered in people's homes via a mobile internet service when necessary. The automated nature of the task allowed for robust implementation of experimental controls such as randomisation of item presentation, and logging allowed such information to be stored for later

examination if required. The results display allowed for data to be easily cut-and-pasted into applications such as Microsoft Excel and SPSS, minimising the need for manual data entry and reducing transcription errors.

Where the task was personally administered using a tablet (as with the 31 non-internet participants), great efforts were made to allow the participant to enter their own responses with a minimum of intervention from the experimenter. In six cases where a participant found the sliders on the scales too difficult to operate, particular care was taken to ensure that the slider position was based on instructions from the participant, and that the participant was completely happy with the final position of the slider before continuing to the next page.

This validation study employed versions of the faces scales in which verbal labels were completely absent. As such – unlike in the validation study by Arruda et al. (1999) – participants were forced to rely on the D-VAMS images alone to report their mood. This study therefore represented a particularly strong test of the utility of the scales, as there was little question of circumventing the images provided as these were the sole means of reporting mood state.

Ethics Committee approval was sought and granted both at university and NHS level, and this project was overseen by supervisors who are expert in this area. A partnership of stroke care professionals including stroke survivors were also consulted on the design of the study, and their recommendations taken on board; feedback was also sought via a number of oral and poster presentations at conferences and other events.

7.2.6.3 Limitations of the study. Though the study sample was a reasonably representative cross-section of stroke survivors, it contained relatively few people from a rehabilitative setting, limiting the ability of results to be generalisable to those

in early stages after stroke. Were these results only to apply to people after a significant period of recovery, it may call into the question of the suitability of D-VAMS as a screening measure for depression in the early stages after stroke, since screening measures are most often used in the acute stage. Since this study was not conducted in the context of an intervention, the responsiveness of D-VAMS to change could not be assessed.

Controlling for nonverbal content only had the benefit of providing a strong test of the utility of the pictures, however it also meant the results reflected the rare scenario in which a stroke survivor was totally unable to recognise even single words. Though this offers a tough test of D-VAMS, it does not reflect a use with a typical range of stroke survivors of varying degree of communication impairment, many of whom would be able to use the mood words included on the live version to support their responses. So while the present findings may reflect the performance of D-VAMS in extreme cases of communication impairment, it doesn't tell us much about how the 'live' version of D-VAMS – with the mood words present – would fare; we can only surmise that the psychometric qualities would be improved somewhat by the presence of the words, though by how much remains unknown.

As discussed earlier, the D-VAMS – like the HADS – is concerned with the self-report of experienced mood, and not behavioural or somatic symptoms that could be due to a medical conditions. The behavioural manifestations of depression however, can be assessed with existing observer rated scales such as the SADQ. With the D-VAMS allowing a cognitively intact stroke survivor with even the most profound loss of communication ability to communicate how they feel, the two might complement one another very well in allowing these dual facets of depression –

behavioural and self-reported – to give a fuller picture of depressive symptomatology in this population.

Like previous research, the findings of this study also assume that the results for language-intact stroke survivors can be generalised to people with aphasia. For the reasons outlined earlier, a validation study in the target population (people with aphasia) is not feasible because any criterion measure – including of course the 'gold standard' diagnostic interview – is impossible for people who cannot use language. It would therefore not be possible to establish whether poor findings in a sample of people with more profound aphasia was due to difficulty in using the D-VAMS or their impaired ability to use a language-based, criterion measure employed in the study. However some faith should be faced in the compelling neuropsychological evidence that recognition of facial expressions – being processed in the right hemisphere – is extremely unlikely to be impaired in people with the left hemisphere lesions most commonly associated with aphasia.

Even granting that D-VAMS is a valid and reliable mood measure suitable for people with aphasia, there is still the problem of instructing a severely aphasic person in the use of the scale. Great efforts were made to ensure that it was easily recognisable, and that its purpose as a self-report, mood assessment instrument was clear from the visual cues and the context of its use, However, though the majority of participants easily grasped the concept behind the instrument and were able to use it without difficulty, there was at least one study participant who had difficulty with the concept of using a 'proxy face' to express their own mood.

In circumstances where profound speech difficulties are present, it may therefore be necessary to support the use of D-VAMS by using gestures and other body language to indicate its purpose. Providing the respondent is cognitively intact,

it is hoped that showing and demonstrating the scales to them, plus the use of gestures to indicate that the face is meant to reflect their own feelings, should be sufficient to convey its purpose. It is also likely that a person in emotional distress who wishes to communicate this but has no other way to do so will to be more motivated to make use of the instrument.

The context in which the HADS questions were used was also potentially problematic. Some of the HADS questions were framed in a way that they could cause confusion. The statement "I still enjoy the things I used to enjoy", in the context of their participation as a stroke survivor was liable to be misinterpreted as meaning *since their stroke*, and not simply how they have been feeling during the last week. There are two questions in the HADS that may also be troublesome in that they may be conflated by stroke symptomatology. As Lincoln et al. observed, "I feel as if I am slowed down" and "I still enjoy the things I used to enjoy" may relate to motor, cognitive or language impairments resulting from stroke, rather than symptoms of depression (Lincoln et al., 2012b, p. 308). This may lead to HADS-D scores that are artificially inflated as a result of responses that reflect stroke symptoms rather than mood *per se*.

It is inevitable that participants recruited online and performing the study task via the web portal unsupervised may also perform differently in other ways to those under supervision. One possible factor is social rules governing self-disclosure. Personal feelings can be a sensitive issue, and there is a social expectation that we should put on a positive face even when our mood is very low. Inhibitions about disclosing the extent of one's low mood may have biased supervised participants in favour of more positive responses on the mood scales or the HADS, as their responses could be seen by the experimenter. This may go some way to explaining the fact that

the Online subgroup had HADS-D scores (and D-VAMS scores) that were around 67% higher than those of the Stroke Club and Rehabilitation subgroup as a whole, however this may be at least in part due to Stroke Club participants being generally happier than the other groups. Whether this was because they were attending a Stroke Club, or whether it is because more depressed people are less likely to attend Stroke Club meetings, however, is debatable.

There is also the matter of sample size as it relates to the PCA and ROC analyses. The sample size was calculated using power calculations based on Pearson correlations, as these tests comprised almost all of the statistical analyses. This sample size, however, may be inadequate to accurately reveal factor structure through the use of PCA or allow for acceptable precision of ROC statistics, and so these results should be viewed with caution.

Finally, there is the question of an experimenter effect. Aside from the aforementioned inhibitions some may have about reporting sensitive information in the presence of a stranger, there is also the question of the influence of supervision. People frequently needed assistance using the tablet, and in extreme circumstances the experimenter would operate the slider on the participant's behalf, asking, "Up or down?" and calibrating it accordingly. Though every precaution was made to ensure that the slider positions represented as best as possible the wishes of the participant, there is always the possibility that subtle biases were introduced that might have favoured an expected pattern of responses.

7.2.4.4 Summary. This validation study demonstrated that an image-only version of D-VAMS has excellent construct validity, excellent test retest reliability and excellent convergent validity with the HADS-D for both the DVAMS-D and DVAMS Mean^{-SA} statistics. However the poor discriminant validity between scales,

the high Cronbach's α , the clear, single factor structure of D-VAMS scores strongly indicate that D-VAMS measures one construct, affective *valence*, and not some combination of *valence* and *activation*. The findings therefore suggest that D-VAMS can be improved by leaving out the seventh, *Sleepy-Alert* scale.

Because of this single-factor structure, the DVAMS-D/DVAMS-A coefficients derived from the two-factor geometry (see 7.1.1) are flawed and do not form differential metrics that can usefully distinguish anxious from depressed mood. However, it is possible that patterns of scores for the individual scales might enable levels of anxiety to be inferred, for example, by unusually high 'afraid' scores relative to those of other scales. The findings, though, also indicate that the D-VAMS Mean^{-SA} score gives a very good indicator of pleasantness of mood on a scale of 0–100, with a faces-only version of the scales giving good sensitivity and specificity at a range of HADS-D cut-offs.

While it is safe to conclude that D-VAMS is probably suitable as a general outcome measure for depression or low mood in people with communication problems following stroke, its utility as a screening measure (which is typically employed in the acute stage poststroke) is less certain due to the sample characteristics. Further study, employing a hospital-based sample in the acute, post-stroke stage, is required to confirm whether the present findings extend to individuals at the stage when they are usually screened.

8. Conclusions

8.1 Important Considerations

Before drawing conclusions as to the suitability of D-VAMS as an outcome measure or screening instrument, it is important to bear in mind the strict constraints that were imposed on the validation study to control for language-based cues. To simulate as best as possible the conditions for a profoundly aphasic respondent, no verbal cues whatsoever were presented in order to ensure total reliance on images of facial expressions alone. It is reasonable to expect that in a real-world implementation – in which the scales are also accompanied by their corresponding mood words – the D-VAMS would only perform better, as it is relatively rare for people with aphasia to be totally lacking in the comprehension of both written and spoken, single words.

It is therefore important to understand that these results probably represent the lower boundaries of the instrument's performance. The use of the live, 'word' version – in which the top and bottom ends of the face scales are also accompanied by the respective mood words – are more suitable for real world use, as in most people with aphasia language is still present to some degree, with single-word recognition frequently left intact.

8.2 Utility of D-VAMS as a Measure of Mood

The validation study findings indicate that D-VAMS is suitable for use in people after stroke. Measures of construct validity, internal consistency and reliability suggests that D-VAMS is a good, general measure of mood covering key domains of affect,

with the total score (Mean^{-SA}) providing a valid and reliable measure of pleasantness of mood on a scale of 0–100. High correlations with the HADS-D suggests that it is a suitable substitute measure for depressed mood where language impairment may otherwise make the use of conventional measures difficult. As such, D-VAMS should be useful as a general outcome measure for this population.

It does however require that a respondent is cognitively intact enough to comprehend the nature of the instrument, and communication support may be required in people whose aphasia is severe. The present study, furthermore, could not allow responsiveness to changes in mood to be assessed.

8.3 Utility of D-VAMS as a Screening Measure for Depression after Stroke

Likewise, sensitivity and specificity suggest good cut-offs for a range of HADS scores, with an optimal cut-off of HADS-D>4. This suggests that D-VAMS is acceptable for use as a screening measure for low mood after stroke, even for stroke patients with the most profound loss of communication. However, this instrument can only be considered a screening measure for depression *per se* in the imprecise sense in which it reflects *depressed mood*; as discussed earlier, mood is only one facet of depression as a clinical construct. D-VAMS is quick to administer (<5 mins) and can be run on most multi-function devices with an internet browser. It is freely available online, and an offline version is also available, making it easy to access from a typical medical or research environment.

A note of caution must be sounded, however, on the use of D-VAMS in the early stages poststroke. Stroke survivors from a rehabilitation setting comprised only 25% of the validation study sample, and none of these were within 2 months of their

stroke, so it remains to be seen whether similar findings would be observed in the early days or weeks following stroke. This is a matter for further investigation.

8.4 Recommendations for Use

It must be remembered that though the present study indicates that D-VAMS is a valid and reliable measure of mood suitable for use as a screening instrument for depression after stroke, it is not intended as a diagnostic instrument to be used in isolation. D-VAMS is intended to provide information that can help guide and inform medical staff and other carers in evaluating the mood of people with communication difficulties following stroke. As with all evaluations, it is prudent to be guided by a number of sources, such as proxy/caregiver ratings based on informal or structured observation. The combination of D-VAMS and SADQ-H10, for example, might usefully offer a fuller and more accurate profile of depressive symptomatology, with the self-report information of the former complementing the behavioural account of the latter.

D-VAMS should also be used under supervision. The interface should first be demonstrated to the respondent, showing the various ways in which the scales can be adjusted depending on the device being used. Depending on the type and severity of aphasia present in the respondent, the purpose of the scale should be communicated with the support of verbal or written communication or sign language. Where a respondent has problems reading but can still comprehend some spoken language, mood words accompanying the scales can also be indicated and read out loud on the scale pages.

Prior to providing a response on each of the scales, the endpoints and transitions between them should be shown to the respondent in order to allow as clear

as possible an understanding of what each scale represents, by way of their trajectory across affect space, and intensity and direction of a particular mood or emotion. If the respondent has a problem physically activating or adjusting controls on the D-VAMS interface, the supervisor can assist with this also. Since the evidence suggests a pronounced practise effect, it would be advisable for respondents new to D-VAMS to first begin with a practise run in order to familiarise themselves with the interface, and then to repeat the exercise as their 'live' run.

D-VAMS is free to use and available on the internet at DVAMS.COM. A downloadable version that can be run offline is also available via the Project Portal link, along with information and other resources.

Appendix I – Study 1 Information Sheet

STUDY INFORMATION SHEET

Developing a Pictographic Mood Scale Suitable for use with Aphasic Stroke Patients

A Pilot Study

It is estimated that 30% of stroke sufferers develop *aphasia*, a condition which results in impairment of a person's ability to speak or to understand language. A person with this condition may also have difficulty reading or writing, making communication of any kind extremely difficult. Clinical guidelines recommend the screening of stroke patients and the assessment of those at high risk of depression. There is therefore a need for screening measures that assess mood in people whose ability to communicate is severely impaired or absent.

Though an existing set of pictographic scales exists for this purpose, it has proved to be of limited clinical use. The objective of this study is to produce a nonverbal mood scale that can more sensitively identify whether a person may be at risk of depression or other mood disorders.

The study will be held at locations to be arranged in advance, normally at venues on the University Campus. The study is in two parts:

Part 1: *Creation of a set of photographs*

20-30 participants will be required for this part of the study. Each will be photographed in separate sessions. Each participant will be given a consent form to read and sign. They will also be given an instruction sheet. Once the participant has read and understood the instructions, they will be seated at a table and asked to pose facial expressions based on 26 mood descriptor words (such as 'happy', 'sad' etc.) that will be presented to them in turn. The experimenter will photograph each of the posed expressions until all have been completed.

This session will take approximately one hour.

Part 2: *Validation of photographs*

Phase 1: 60-100 participants will be required for this phase of the study. Single participants, or participants in groups of up to 5 will be asked to judge one set of 26 photographs in terms of the extent to which the facial expressions depicted are perceived to correspond to a set of listed mood words. The photographs will be displayed as a Powerpoint presentation, and the participants will be asked to mark their responses on a form provided, on a Likert scale of 1-7.

This task will take approximately one hour, inclusive of a ten minute break half way through.

Phase 2: 150-300 participants will be required for this phase of the study. Groups of participants be asked to judge one set of 26 photographs in terms in terms of selecting or providing a suitable mood word. The photographs will be displayed as a Powerpoint presentation, and the participants will be asked to mark their responses on a form provided.

This task will take approximately 5-10 minutes.

Participating in the study

- Participants should read the this information sheet and understand the nature and purpose of the research project.

- While information and material collected during the study may be published, participants will not be identified and their personal results will remain confidential.

- Data will be stored on password protected computers used by the researcher. Paper forms will be kept locked in a secure place. Data collected may be shared with other researchers, but personal details will not be disclosed without express permission of the participant.

- Participants taking part in the first part of the study will be asked to sign a consent form and may have their photographs used in further stages of the study and as part of the final scale.

- Participants may withdraw from the study at any stage without having to give a reason.

- Participants may contact the researcher or supervisor if they require further information about the research, and may contact the Research Ethics Coordinator of the Institute of Work Health and Organisations, University of Nottingham, if they wish to make a complaint relating to their involvement in the research.

Contact details

Researcher: *Paul Barrows - lpxpb4@nottingham.ac.uk*
Tel: 01623 655174

Supervisors: *Dr Shirley Thomas - shirley.thomas@nottingham.ac.uk*
Prof Nadina Lincoln - nadina.lincoln@nottingham.ac.uk

Appendix II – Study 1: Instructions for Actors

Developing a Pictographic Mood Scale Suitable for use with Aphasic Stroke Patients

Pilot Study - Part I

Instructions for 'Actors'

The purpose of this part of the study is to identify facial expressions which are most clearly recognisable, and which can most accurately convey a given mood state. Below are a list of mood descriptors:

Energetic	*Disappointed*	*Calm*	*Confused*
Pleased	*Tense*	*Bored*	*Sleepy*
Excited	*Angry*	*Tired*	*Content*
Happy	*Nervous*	*Miserable*	*Peaceful*
Enthusiastic	*Distressed*	*Depressed*	*Neutral*
Sad	*Anxious*	*Satisfied*	
Afraid	*Aroused*	*Relaxed*	

These are the mood states that you will be asked to portray using facial expressions. For posing the expressions, there are two methods that may be useful in helping to invoke a convincing expression.

❑ *Method 1 - "Remembered moments"*

To help 'summon up' the appropriate emotion when posing an expression, it is sometimes helpful to remember a moment when you vividly recall experiencing the feeling described. Try to think of a particularly clear memory when you felt that way.

❑ *Method 2 - "The Mime"*

An alternative is to imagine that you are trying to communicate with somebody behind a window pane where they can see but cannot hear you. Use your facial expression to try and communicate your mood state so that an imaginary viewer could understand as best as possible how you are feeling.

When you are ready for the sitting, the photographic session will begin. *Please note that for the sitting, all makeup, visible facial jewellery and eyeglasses must be removed*.

The experimenter will read the descriptors to you in random order, taking three photographs for each posed expression, preceded by one photograph with the descriptor card held up before you. You may adjust your posture and tilt your head up or down a little as part of your portrayal of mood states, but please remain central and do not turn your head to either side.

Appendix III – Study 1/2: Consent Form for Actors

PARTICIPANT CONSENT FORM

Project title Developing a pictographic mood scale suitable for use with aphasic stroke patient

Researcher's name Paul Barrows

Supervisors' names Dr Shirley Thomas & Prof Nadina Lincoln

	Please check
• I have read the Participant Information Sheet and the nature and purpose of the research project has been explained to me.	❑
• I understand the purpose of the research.	❑
• I understand that I may withdraw from the research project at any stage without having to give a reason.	❑
• I understand that I will be photographed for the study session.	❑
• I understand that while information and material collected during the study may be published, I will not be identified and my personal results will remain confidential. I understand that the photographs collected will be used in further stages of the study and as part of the final scale.	❑
• I understand that data will be stored on password protected computers used by the researcher. Paper forms will be kept locked in a secure place. Data collected may be shared with other researchers, but personal details will not be disclosed without express permission of the participant.	❑
• I understand that I may contact the researcher or supervisor if I require further information about the research, and that I may contact the Research Ethics Coordinator of the Institute of Work Health and Organisations, University of Nottingham, if I wish to make a complaint relating to my involvement in the research.	❑
• I agree to take part.	❑

Signed …… (research participant)

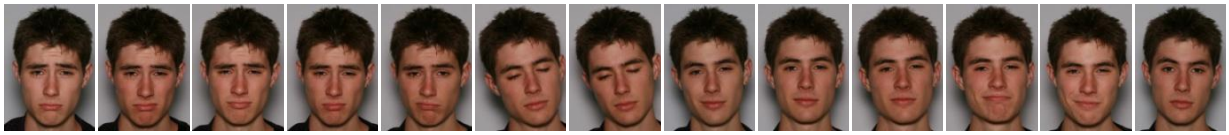
Print name …… **Age**…… **Date** ……

£___ received (initial) ……

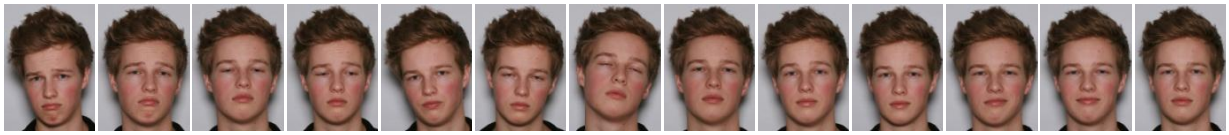
Appendix IV – 20 Actors Each Pose Expressions for 26 Mood Words

Study 1: Actors 1-4 Pose Expressions for 26 Mood Words

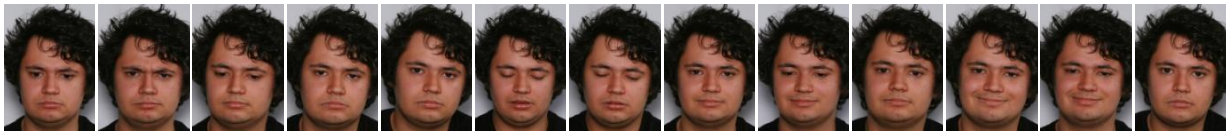
Actor A01



Actor A02



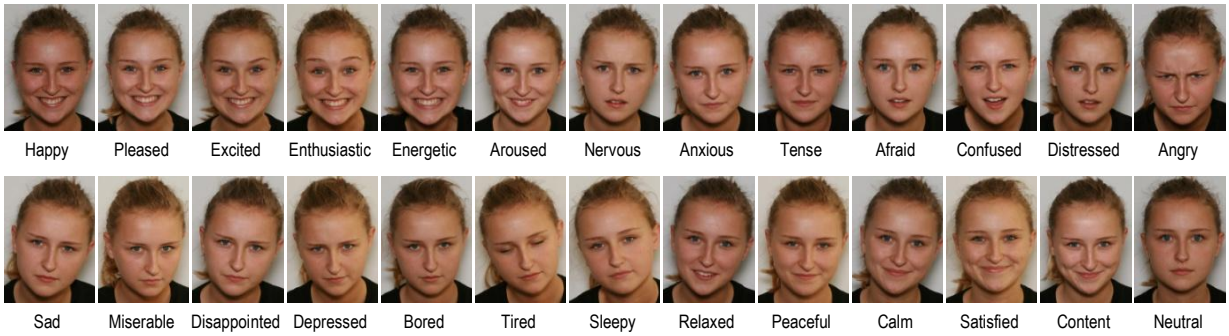
Actor A03



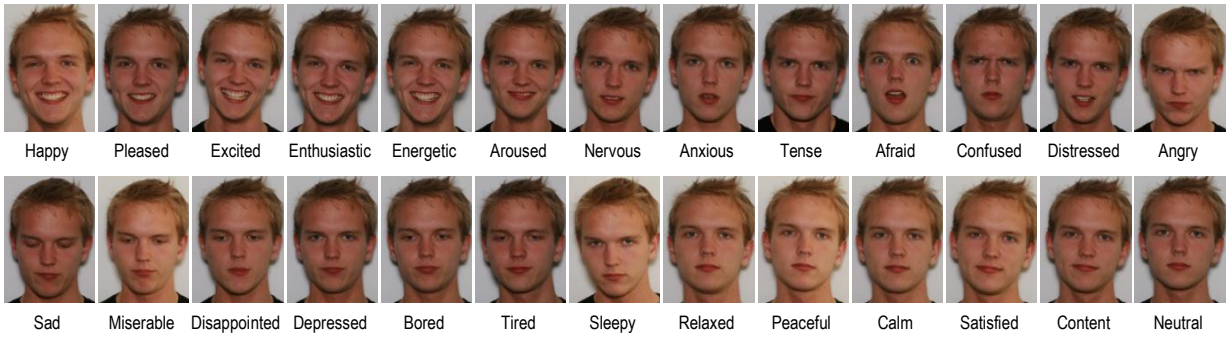
Actor A04

Study 1: Actors 5-8 Pose Expressions for 26 Mood Words

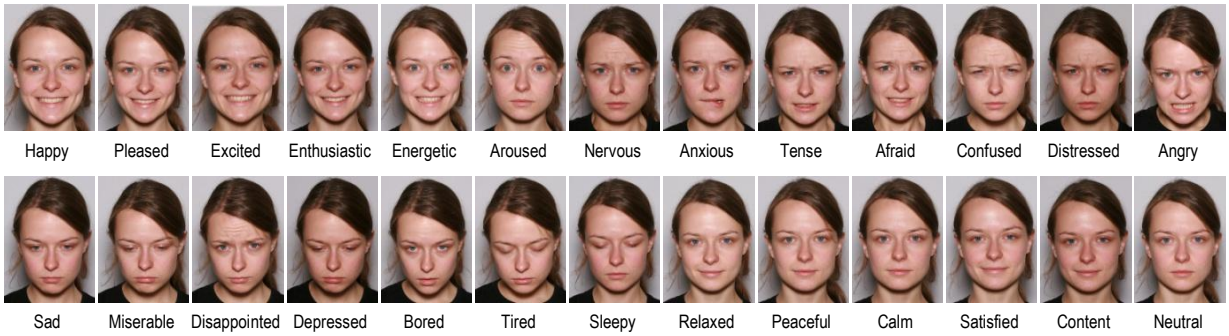
Actor A05



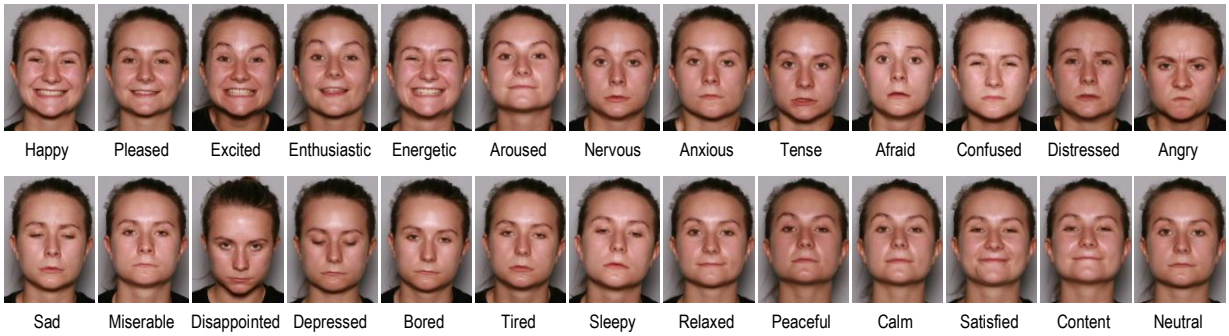
Actor A06



Actor A07

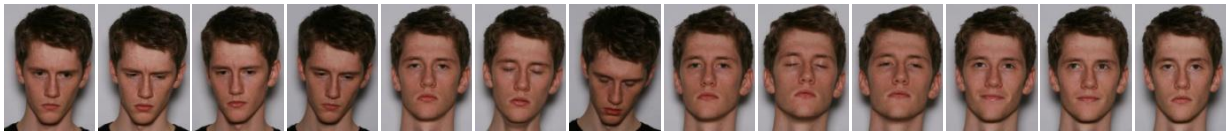


Actor A08

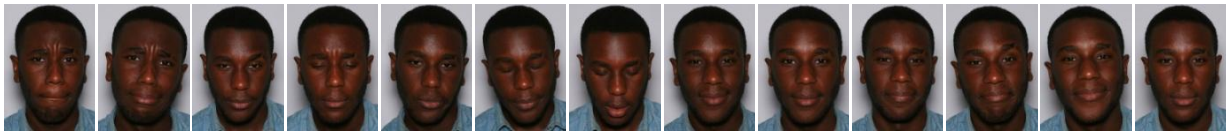


Study 1: Actors 9-12 Pose Expressions for 26 Mood Words

Actor A09

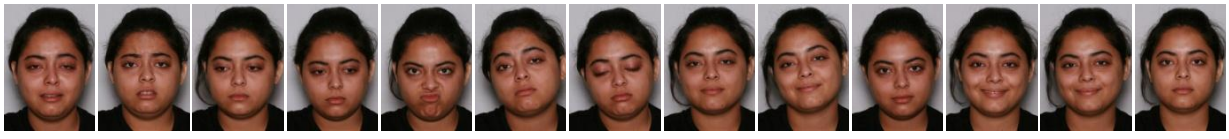


Actor A10



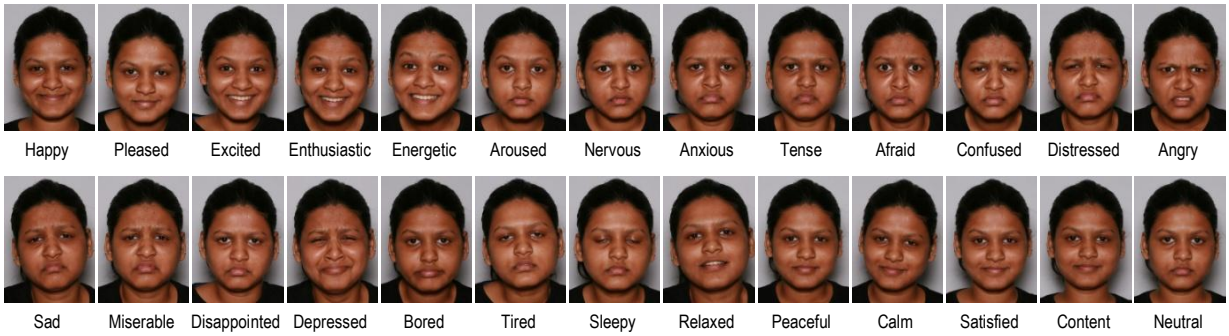
Actor A11

Actor A12

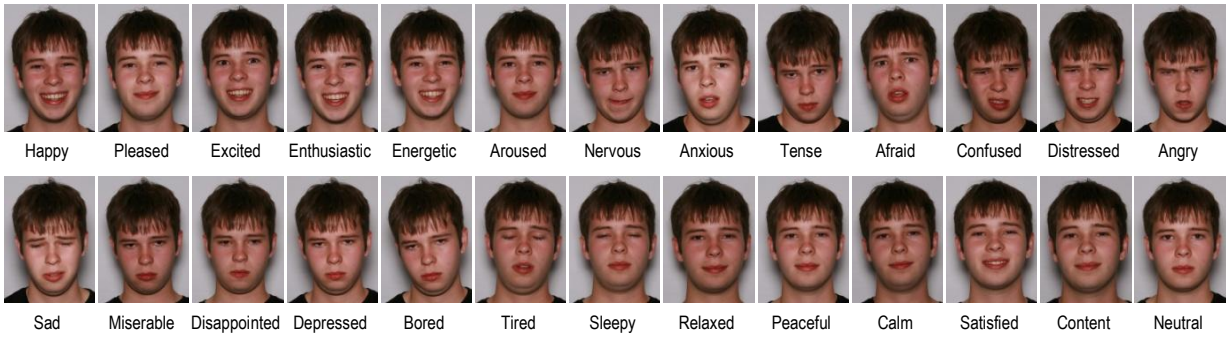


Study 1: Actors 13-16 Pose Expressions for 26 Mood Words

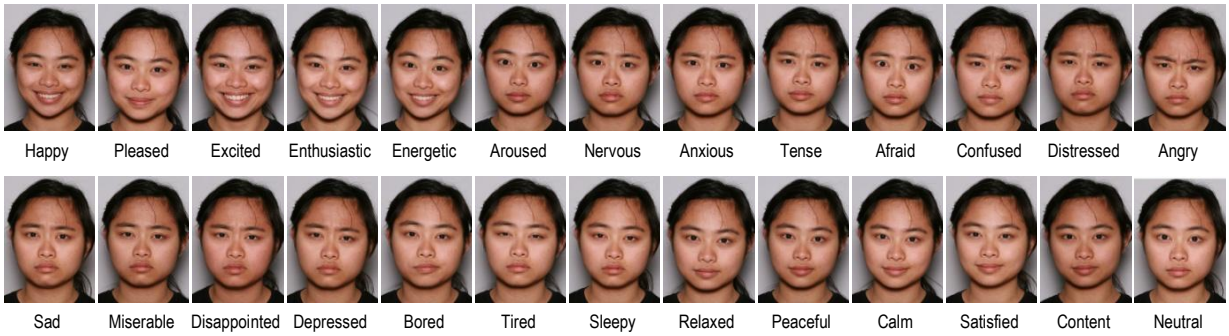
Actor A13



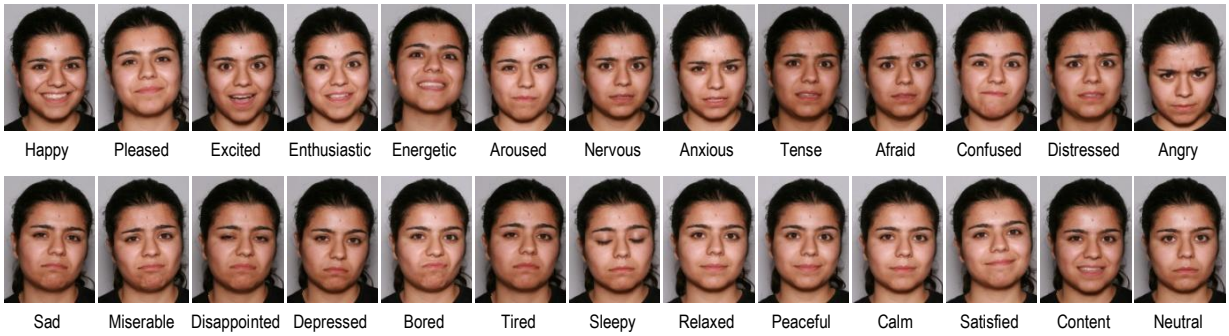
Actor A14



Actor A15

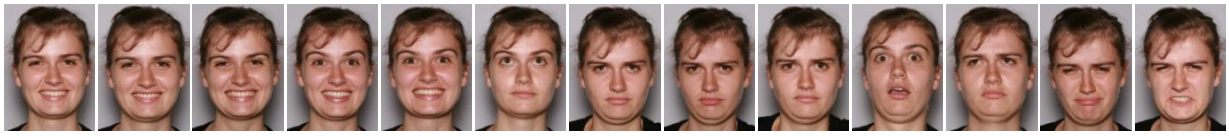


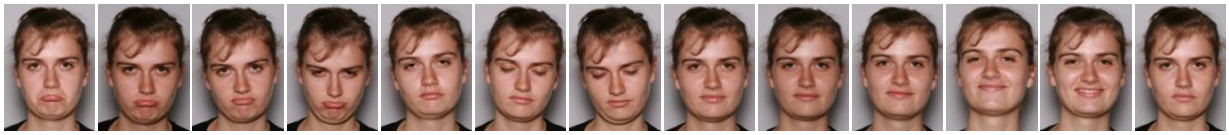
Actor A16



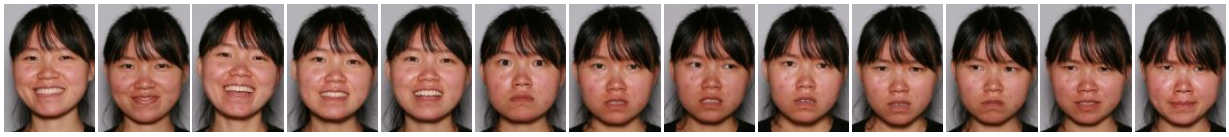
Experiment 1: 20 Actors Each Pose Expressions in Response to 26 Mood Words

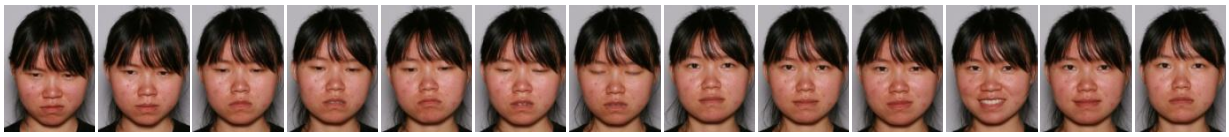
Actor A17



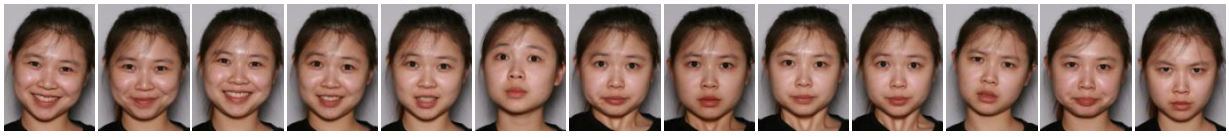


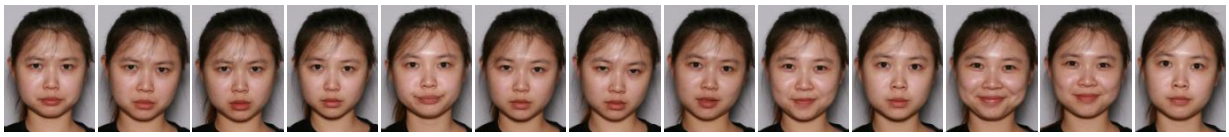
Actor A18



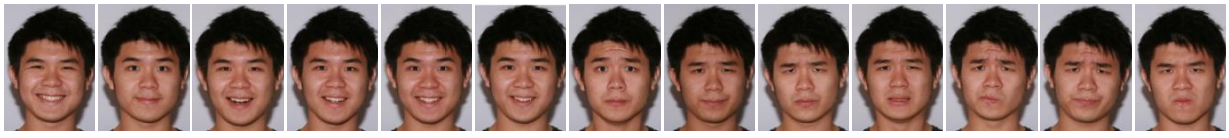


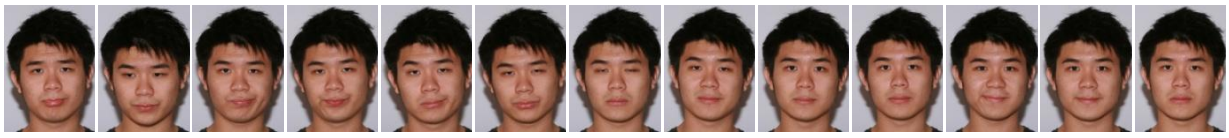
Actor A19





Actor A20





Appendix V – Consent Form for Online Experiments

PARTICIPANT CONSENT FORM

Project title Developing mood scales suitable for use with aphasic stroke patients

Researcher's name Paul Barrows

Supervisors' names Dr Shirley Thomas & Prof Nadina Lincoln

	Please check
• I have read the study description under the About tab, and understand the purpose of the research. I understand that I may withdraw from the research project at any stage without having to give a reason.	❑
• I understand that results from data collected in this study may be used in published reports, but I will not be identified and my personal data will remain confidential.	❑
• I understand that data will be stored on password protected computers used by the researcher. Data collected may be shared with other researchers, but personal details will not be disclosed without your permission. I understand that I have the right to withdraw my data up to 7 days after the experiment. After 7 days it cannot be guaranteed that the data has not been included in any analysis and/or write up. Personal data will be held for 7 years after the study has ended and then destroyed as applicable.	❑
• I understand that I may contact the researcher or supervisor if I require further information about the research, and that I may contact the Research Ethics Coordinator of the Faculty of Medicine and Health Sciences: Louise Sabir, E-mail: louise.sabir@nottingham.ac.uk., if I wish to make a complaint relating to my involvement in the research.	❑
• I agree to take part.	❑

Appendix VI – Study 2/3: Information Sheet

STUDY INFORMATION SHEET

Developing a Mood Scale Suitable for use with Aphasic Stroke Patients

It is estimated that 30% of stroke sufferers develop *aphasia*, a condition which results in impairment of a person's ability to speak or to understand language. A person with this condition may also have difficulty reading or writing, making communication of any kind extremely difficult. Clinical guidelines recommend the screening of stroke patients and the assessment of those at high risk of depression. There is therefore a need for screening measures that assess mood in people whose ability to communicate is severely impaired or absent.

Though an existing set of pictographic scales exists for this purpose, it has proved to be of limited clinical use. The objective of this study is to produce a nonverbal mood scale that can identify whether a person may be at risk of depression or other mood disorders.

Study 1 of this project was conducted during the academic year of 2012/2013. It involved the creation and validation of a set of mood photographs and the development of a coordinate system within which these mood images could be located. This information sheet details the studies that will occupy the remainder of the project up to its conclusion in the construction and validation of the final scale. Sessions involving photographic sittings will be conducted at locations to be arranged in advance, normally at venues on the University Campus. The remainder of the study will be conducted via an internet based experimental portal (xvams.com).

Photographic Sittings

Participants for the photographic sittings will sign a detailed consent form. They will retain the right to withdraw from the study at any point, and to withdraw their data from the study within 7 days of the sitting. After 7 days it cannot be guaranteed that their data has not been included in analysis and/or write up.

Online Tasks

Participants in the online part of the study will be require to create a login with a username and password. The participant will be asked to provide basic demographic information including gender, year of birth and ethnicity, and will be required to check a checkbox agreeing to the terms and conditions of the study, including arrangements for payment of any inconvenience allowance if applicable. It will be emphasised that the email address provided by the participant will only be used for correspondence relating to the study and not disclosed to any third party.

Study 2: Creating Reference Image Sets For the Scales

Part 1: Collecting Supplemental and Transitional Images

For the first part of this study, highest scoring actors from the first study will be recalled to pose the additional expressions representing varying intensities of a subset of 12 of the expressions originally posed.

On six bipolar scales, the participant will then be instructed – using their existing photographs for reference – to reproduce the end-point expression, and then to gradually modify their expression, making it less intense until it has reached a 'neutral' point. They will then be asked to do the same for the expression at the opposite end of the scale. As the expressions are posed, many photographs will be taken in order to capture a range of expression intensities from which candidate interval images will be selected. This process will be repeated until enough images of sufficient quality have been collected for the scale. Once this process is complete for one scale, the process will be repeated for the next scale in the list, until photographs for all of the scales have been completed. A further pose will be requested to create an 'alert' item that is required to complete one of the bipolar scales.

(Experimental session duration: approx. 60-90 min)

Part 2: Screening and Selection of Candidate Images for the Final Scale

The photographs will be separated into pools for each scale, and divided into three groups, those which belong to one end of the scale (between 'neutral' and the first pole), another will be those from the other end of the scale (between 'neutral' and the second pole), and a third group of the 'neutral' poses which appear close to the midpoint. From these groups, candidate images representing one-third and two-thirds expression

intensities will be selected, along with candidates for the 'neutral' position. This stage of the selection process will be based on inspection of the photographs by the experimenter.

Once this process is complete for all of the scales, each scale will have two sets of transitional image candidates, one for each pole of the scale. Candidate 'neutral' images from each of the scales will be pooled and further refined, eliminating any images that are clearly weaker in terms of face validity. Photographs for the 'alert' pose will also be sorted and shortlisted accordingly.

Part 3: Judgement Task for 'Alert' and 'Neutral' Items

In order to complete the scales, one more mood item, provisionally titled 'alert', and a 'neutral' item – representing the origin of the circumplex – will need to be validated. This will occur by means of a judgement task much like those completed in Study 1, where a number of images will be presented to participants, who will then score them on Likert scales in order to assess their location in two-factor space. These candidate images will be those posed by actors recalled for part 1 of this study.

(Experimental session duration: approx. 10-15 min)

Part 4: Task to Locate Scale Positions of Transitional Items

Finally, a further judgement task will be implemented to empirically locate the position of the transitional images comprising the one-third and two-third intensity images between the neutral point and the end-points of the scales to which they belong. Again, this will be accomplished by use of an online judgement task in which participants will be asked to scale the images along Likert scales between pairs of reference images.

(Experimental session duration: approx. 10-15 min)

Study 3: Construction and Validation of the Prototype Scales

Part 1: Charting the Scales and Generating Second-Order, Morphed Transitions

Once the transitional images have been empirically located on their respective scales, further transitional images will be generated by morphing software between neighbouring items, such that each final scale comprises a linear array of 41 images from one end to the other. This series of images will enable a smooth transition of expressions to be represented from one end of the scale to the other. A set of these scales will be

generated for each of the actors recalled for further study in study 2 part 1.

The final scales will then be incorporated into software accessible via an internet website, or which can operate in standalone mode on a suitably equipped device.

Part 2: Validating The Scales in a Non-Clinical Sample

Finally, the scales will be validated by having participants in them concurrently with language dependant versions of the scales and an electronic form of at least two other measures commonly used to indicate severity of depression symptoms. The resulting correlation matrix between the scores

(Experimental task duration: approx. 20-30 min)

Participating in the study

- Participants should read this information sheet and understand the nature and purpose of the research project

- While information and material collected during the study may be published, participants will not be identified and their personal results will remain confidential

- Data will be stored on password protected computers used by the researcher. Paper forms will be kept locked in a secure place. Data collected may be shared with other researchers, but personal details will not be disclosed without express permission of the participant. Participants will be informed that they have the right to withdraw their data up to 7 days after the sitting. After 7 days it cannot be guaranteed that their data has not been included in analysis and/or write up. Data will be held for 7 years after the study has ended and then destroyed as applicable

- Participants taking part in a photographic sitting will be asked to sign a consent form and may have their photographs used in further stages of the study and as part of the final scale

- Participants may withdraw from the study at any stage without having to give a reason

- Participants may contact the researcher or supervisor if they require further information about the research, and may contact the Research Ethics Coordinator of the Faculty of Medicine and Health Sciences: Louise Sabir, E-mail:

louise.sabir@nottingham.ac.uk., if they wish to make a complaint relating to their involvement in the research.

Contact details

Researcher: *Paul Barrows - lpxpb4@nottingham.ac.uk*
Tel: 01623 655174

Supervisors: *Dr Shirley Thomas - shirley.thomas@nottingham.ac.uk*
Prof Nadina Lincoln - nadina.lincoln@nottingham.ac.uk

Appendix VII – Study 2: Instructions for Actors

Developing a Mood Scale Suitable for use with Aphasic Stroke Patients

Study 2

Instructions for 'Actors'

For this part of the study, you will first be given an information sheet and consent form. When you have read the information sheet, please fill in and sign the consent form (with the exception of the payment details fields, which are to be completed at the end of the experiment). You will then be given a short talk explaining the theory behind the scales being developed and the need for additional images representing transitional stages between facial expressions.

Currently, 7 bipolar scales are to be constructed, with endpoints as below. For each of these scales, you will be asked to pose the corresponding facial expressions, but this time with varying degrees of intensity. You will also be coached in posing one more expression not included in the original series.

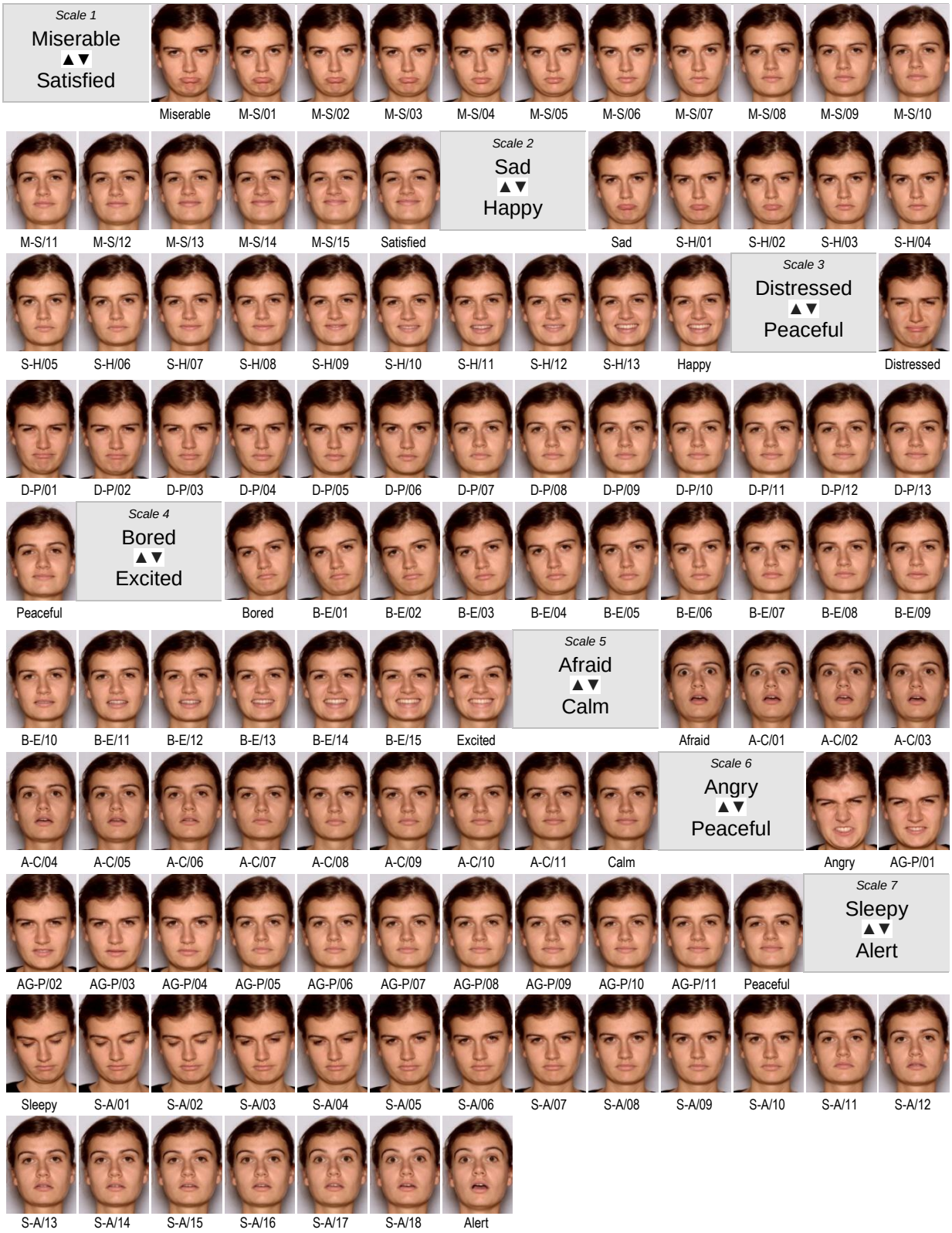
Scale	Pole 1	Pole 2
1 (M-S)	Miserable	Satisfied
2 (S-H)	Sad	Happy
3 (D-P)	Distressed	Peaceful
4 (B-E)	Bored	Excited
5 (A-C)	Afraid	Calm
6 (A-P)	Angry	Peaceful
7 (T-A)	Tired	Alert

As with the previous sitting, *please make sure that all makeup, visible facial jewellery and eyeglasses are removed before the sitting begins*. Please also ensure that long hair is tied back, and secured with hair gel if required to prevent loose strands from falling in front of your face during the sitting.

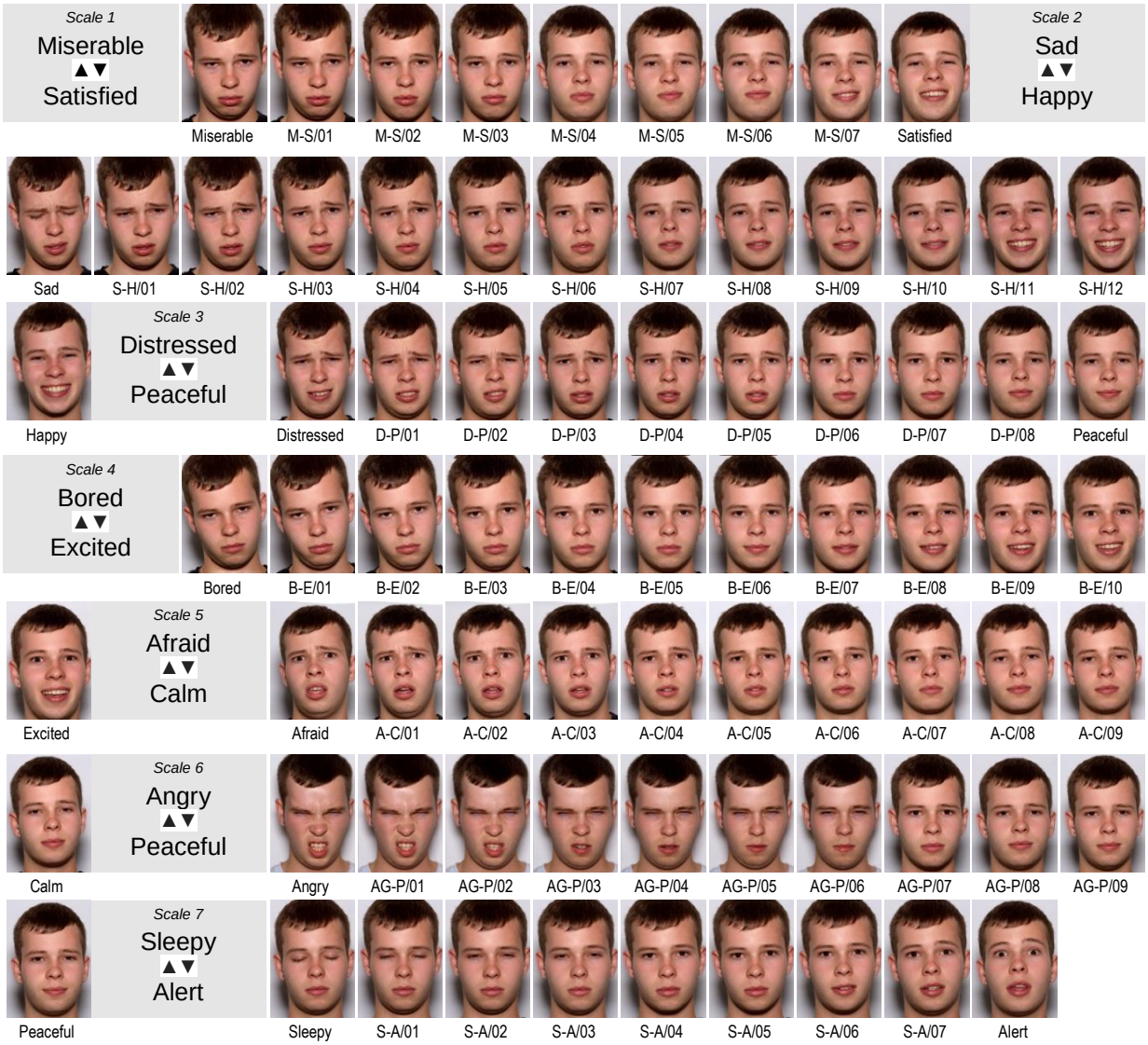
When you are ready for the sitting, the photographic session will begin. The experimenter may instruct you to make more than one pass along each of the listed scales as photographs are taken, and you may also be given more specific instructions relating to the tilt of your head, or asked to modulate specific elements of your facial expression in order to maintain continuity with other images in a series.

Once the experimenter has collected the required images, you will receive the agreed allowance payment and asked to initial the payment section of the consent form as acknowledgement.

Appendix VIII - Scale Keyframe Images for Actor #1

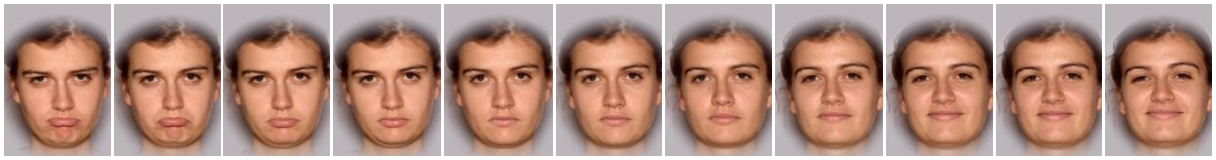


Appendix IX- Scale Keyframe Images for Actor #2



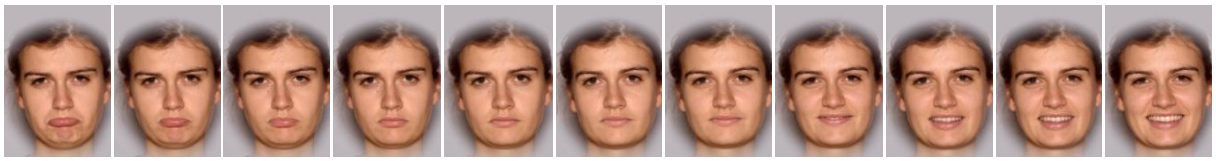
Appendix X – Morphed Scale Continua for Actor #1 at 10% Increments

Scale 1: Miserable – Satisfied



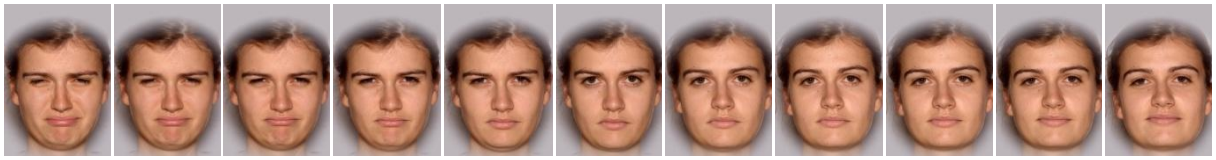
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 2: Sad – Happy



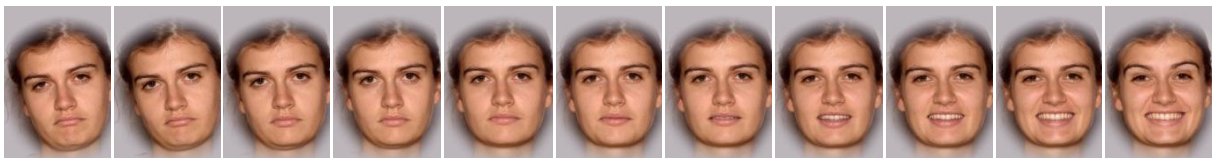
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 3: Distressed – Peaceful



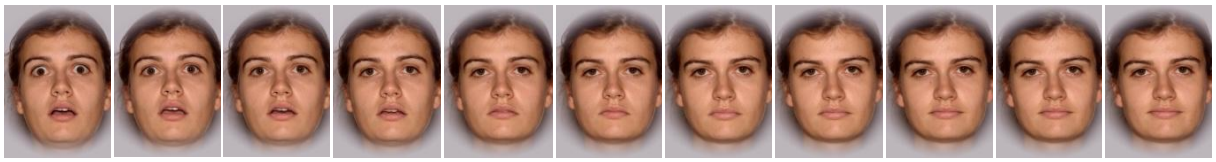
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 4: Bored – Excited



0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 5: Afraid – Calm

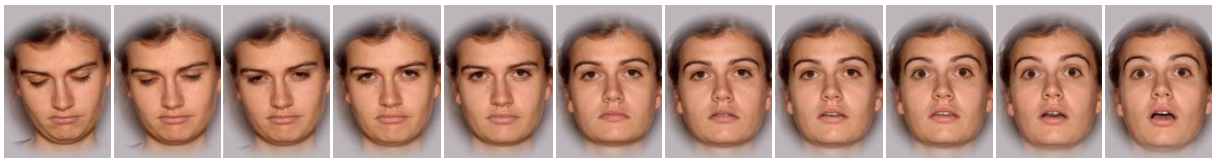


0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 6: Angry – Peaceful

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 7: Sleepy – Alert



0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Appendix XI – Morphed Scale Continua for Actor #2 at 10% Increments

Scale 1: Miserable – Satisfied

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 2: Sad – Happy

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 3: Distressed – Peaceful

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 4: Bored – Excited

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 5: Afraid – Calm

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 6: Angry – Peaceful

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Scale 7: Sleepy – Alert

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Appendix XII – Study 3: Participant Information Sheet

'Emotiscope' Study (D-VAMS)

Stroke survivor volunteers needed for short (10-15 min approx) online study (xvams.com)

Brief Description:

A new, tablet/computer-based tool has been developed to assess mood in stroke patients with communication problems; the tool uses sliders to animate facial expressions between mood states. The scales (Dynamic Visual Analogue Mood Scales - DVAMS) now need validating in a small sample of stroke survivors without significant aphasia.

For this essential study, we need volunteer stroke survivors to complete a task of approximately 10-15 minutes. Participants should:

- be English-speakers who are over 18 years of age
- have had a stroke
- be capable of giving informed consent.

People with some aphasia may take part if they are able to understand written or spoken English, as this is required to perform the task.

If you fit these criteria – or know somebody who does – you/ they may take part in the study via the DVAMS project portal at: http://www.xvams.com. If a volunteer does not have an internet connection, a home visit with a tablet or laptop can be arranged by our researchers if they are within the Nottingham area, UK. In order to take part you will need to create a username and password on the portal website. A valid email address is required, but it will be used only by the project experimenter or supervisor, and only to contact you regarding the experiment if necessary, in line with ethics committee requirements.

Further details:

Stroke survivors are at particularly high risk of depression or other mood problems, and this may substantially affect a person's

recovery. Evidence suggests that people with communication problems due to aphasia are particularly at risk, yet there are few instruments to assess mood in people with severe communication difficulties. There is therefore a need to develop screening measures to assess mood in people whose ability to communicate is severely impaired or absent.

Dynamic Visual Analogue Mood Scales (D-VAMS or 'Emotiscope') are the end result of a 3 year project to develop such an assessment instrument. It is based on facial expressions of a number of emotions, which can be changed using a slider.

This study has been approved by the University of Nottingham, Faculty of Medicine & Health Sciences Ethics Committee. It is hoped that this tool will be useful in screening for depression and other mood problems after stroke in patients who have serious communication problems as a result of aphasia. Further information about the experiment and a downloadable information sheet is available here:
http://www.xvams.com/about.aspx

Paul Barrows
PhD Candidate

Division of Rehabilitation & Ageing
School of Medicine
The University of Nottingham
NG7 2UH
t: +44 (0)1623 655174
Email: Paul Barrows at lpxpb4@nottingham.ac.uk

Appendix XIII – Study 3: NHS REC Participant Information Sheet

Participant Information Sheet
(Version 1.1: 2015-10-14)

Title of Study: Developing mood scales suitable for use with stroke patients with communication difficulties

Name of Researcher(s): Paul Barrows, Shirley Thomas

We would like to invite you to take part in our research study. Before you decide we would like you to understand why the research is being done and what it would involve for you. One of our team will go through the information sheet with you and answer any questions you have. Talk to others about the study if you wish. Ask us if there is anything that is not clear.

What is the purpose of the study?

A new, tablet/computer-based tool has been developed to assess mood in stroke patients with communication problems. The tool uses sliders to animate facial expressions between different mood states (such as happy–sad). The scales now need validating in a small sample of stroke survivors without significant aphasia.

Why have I been invited?

You are being invited to take part because you have had a stroke. We are inviting 20 participants like you to take part.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent-to-contact form. You will be asked to provide informed consent in an online form provided prior to taking part. If you decide to take part you are still free to withdraw at any time and without giving a reason. This would not affect your legal rights.

What will happen to me if I take part?

You will first be asked to sign a consent-to-contact form, which will be provided to you by a member of your care team. For this, we only require your name, a contact phone number (or email) and your signature. This gives permission for an investigator to contact you to arrange a suitable time and place to take part. A home visit can be arranged if you like.

The study involves taking part in a short task using a home computer, laptop computer or tablet computer, which will be provided for you by the experimenter. The task usually takes 10-15 minutes, and involves responding to a number of

questions about your mood. You respond to these questions by selecting answers in multiple-choice format, or by moving sliders along a rating scale.

Expenses and payments

Participants will not be paid to participate in the study. Travel expenses will be offered for any visits incurred as a result of participation.

What are the possible disadvantages and risks of taking part?

If you are visited at home, wireless internet will be provided if you don't have your own internet connection. However, these services are slower and less reliable than ordinary connections. This may result in the task being longer than usual, sometimes taking half an hour or more.

What are the possible benefits of taking part?

We cannot promise the study will help you but the information we get from this study may help us to better assess the mood of stroke survivors who have severe communication difficulties.

What happens when the research study stops?

After the analysis, all participants will be contacted and offered an update on the study. A study bulletin sheet will be made available which will explain the study results and their implications for stroke survivors.

What if there is a problem?

If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions. The researchers contact details are given at the end of this information sheet. If you remain unhappy and wish to complain formally, you can do this by contacting Nottingham CityCare Partnership Customer Care Team: 0115 883 9654.

Will my taking part in the study be kept confidential?

Any personal details held about you, such as your contact details or address, will be held securely on password protected computers in password protected files. This information will be deleted 3 to 6 months after the end of the study. We will follow ethical and legal practice and all information about you will be handled in confidence.

If you join the study, some parts of the data collected for the study will be looked at by authorised persons from the University of Nottingham who are organising the research. They may also be looked at by authorised people to check that the study is being carried out correctly. All will have a duty of confidentiality to you as a research participant and we will do our best to meet this duty.

All information which is collected about you during the course of the research will be kept strictly confidential, stored in a secure and locked office, and on a password protected database. Any information about you which leaves the hospital will have your name and address removed (anonymised) and a unique code will be used so that you cannot be recognised from it.

Your personal data (e.g: address, telephone number) will be kept for 3-6 months after the end of the study so that we are able to contact you about the findings of the study (unless you advise us that you do not wish to be contacted). All research data will be kept securely for 7 years. After this time your data will be disposed of securely. During this time all precautions will be taken by all those involved to maintain your confidentiality, only members of the research team will have access to your personal data.

What will happen if I don't want to carry on with the study?

Your participation is voluntary and you are free to withdraw at any time, without giving any reason, and without your legal rights being affected. If you withdraw then the information collected so far cannot be erased and this information may still be used in the project analysis.

Involvement of the General Practitioner/Family doctor (GP)

In the event that your scores on the mood questionnaire raise concerns about your well-being, the researcher may contact you to suggest that you contact your GP. The researcher may offer to notify your GP of any concerns raised, but only with your consent. In this case the GP would be notified of your Hospital Anxiety and Depression Scale (HADS) scores and the cut-off employed in the study to flag areas of concern.

What will happen to the results of the research study?

The results will be written up as part of an academic thesis. They will also be written up as one or more papers and submitted to suitable journals during 2016. A study bulletin sheet will also be provided and made available to participants following the study.

Who is organising and funding the research?

This research is being organised by the University of Nottingham and is being funded by the principal investigator Paul Barrows.

Who has reviewed the study?

All research in the NHS is looked at by independent group of people, called a Research Ethics Committee, to protect your interests. This study has been reviewed and given favourable opinion by West of Scotland Research Ethics Service (WoSRES) Research Ethics Committee.

Further information and contact details

Chief investigator: Dr Shirley Thomas
Division of Rehabilitation and Ageing
School of Medicine, Room B105, Medical School
Queens Medical Centre Nottingham NG7 2UH
Phone: +44 (0) 115 84 67484
Email: shirley.thomas@nottingham.ac.uk

Principal investigator: Paul Barrows
School of Medicine, Room B105, Medical School
Queens Medical Centre Nottingham NG7 2UH
Email: lpxpb4@nottingham.ac.uk

Appendix XIV – Study 3: NHS REC Consent to Contact Form

Consent to contact form

Background:
Where a member of staff is recruiting participants on behalf of an investigator, this form may be used to give the investigator permission to contact a prospective participant.

To be filled in and signed by a prospective participant for the study described in an accompanying information sheet, and then passed on to the principle investigator.

Contact Details for Prospective Study Participant

Name	
Phone number or email	

I confirm that I have read the study information sheet and give consent for the principal investigator to contact me about this study. I understand that my contact information will remain confidential.

Signature ……

Appendix XV – Study 3: NHS REC Online Consent Form

CONSENT FORM
(Final version 1.0: 2015-09-24)

Title of Study: Developing mood scales suitable for use with aphasic stroke patients

REC ref: 15/WS/0239

Name of Researcher: Paul Barrows

Name of Participant:

1. I confirm that I have read and understand the information sheet version number 1.1 dated 14/10/2015 for the above study and understand the purpose of the research. I understand that I may withdraw from the research project at any stage without having to give a reason. ☐

2. I understand that results from data collected in this study may be used in published reports and may be shared anonymously with other researchers, but I will not be identified and my personal data will remain confidential. ☐

3. I understand that data collected in the study may be looked at by authorised individuals from the University of Nottingham, the research group and regulatory authorities where it is relevant to my taking part in this study. I give permission for these individuals to have access to these records and to collect, store, analyse and publish information obtained from my participation in this study. I understand that my personal details will be kept confidential. I understand that I have the right to withdraw my data up to 7 days after the experiment. After 7 days it cannot be guaranteed that the data has not been included in any analysis and/or write up. Data will be held for 7 years after the study has ended. ☐

4. I understand that the information held and maintained by University of Nottingham and other central UK NHS bodies may be used to help contact me or provide information about my health status. ☐

5. I understand that if scores returned on the questionnaire raise concerns about my mental wellbeing, a researcher may contact me and offer to contact my GP on my behalf. ☐

6. I agree to take part in the above study. ☐

Appendix XVI – Study 3: D-VAMS Working Practise Document

Dynamic Visual Analogue Mood Scales (D-VAMS) Validation Study

Working Practice Document: Concerns about participant mental health and wellbeing during the Experiment

This document provides guidance to researchers noting concerns about a participant's mental health and wellbeing at the time of the experimental task.

Where a Research Assistant (RA) has concerns about suicidal risk they should inform the Principal Investigator (PI). They should not be expected to make a decision in isolation and should discuss what actions are required with the PI, a supervisor, or a senior member of staff.

After the Experimental Task

1. The HADS and D-VAMS scores will be reviewed in a timely manner. If there is evidence that the participant has a significant mental health problem (HADS >= 11) then the participant will be contacted to discuss their results. If the researcher believes that the participant's mood is of concern, or suspects a suicidal risk, then they should check whether the participant has spoken to anyone about their feelings, e.g. their GP, and what action has been taken. If the researcher is reassured that any issues are being actively addressed, then no further action is required. However if the participant has not shared their feelings prior to this, advise them to seek support through their GP or to contact the Samaritans helpline.

2. The researcher should also offer to contact the GP on their behalf. It may be necessary for the researcher to share information with the GP; the researcher should always ask the participant's permission before doing this, documenting their actions and informing the PI.

3. The RA should inform the PI and document any actions that they have taken, including verbal consent to share information.

What to do if the participant does not give verbal consent to share information and/or you have immediate high level concerns

It is best practice wherever possible to agree the sharing of medical information with the participant before taking action; however in some cases this may not be possible.

1. If the researcher has immediate high level concerns, for example the participant is at risk of suicide/self-harm, they should consult the PI, a supervisor, or a senior member of staff to determine the next steps. It may be the case that the PI needs to contact the participant to further assess the level of risk prior to making decisions about the way forward. Document your discussions with all parties.

2. If the level of immediate risk to the participant is lower (e.g. evidence of low mood but no indication of suicidal intent) and the participant does not want to take any further action or share information, discuss the way forward with the PI. It may be the case that once advice has been provided, no further action is needed. Document your discussions with all parties.

Further information and contacts

See "Suicide mitigation in Primary Care" leaflet

Appendix XVII – D-VAMS Interface

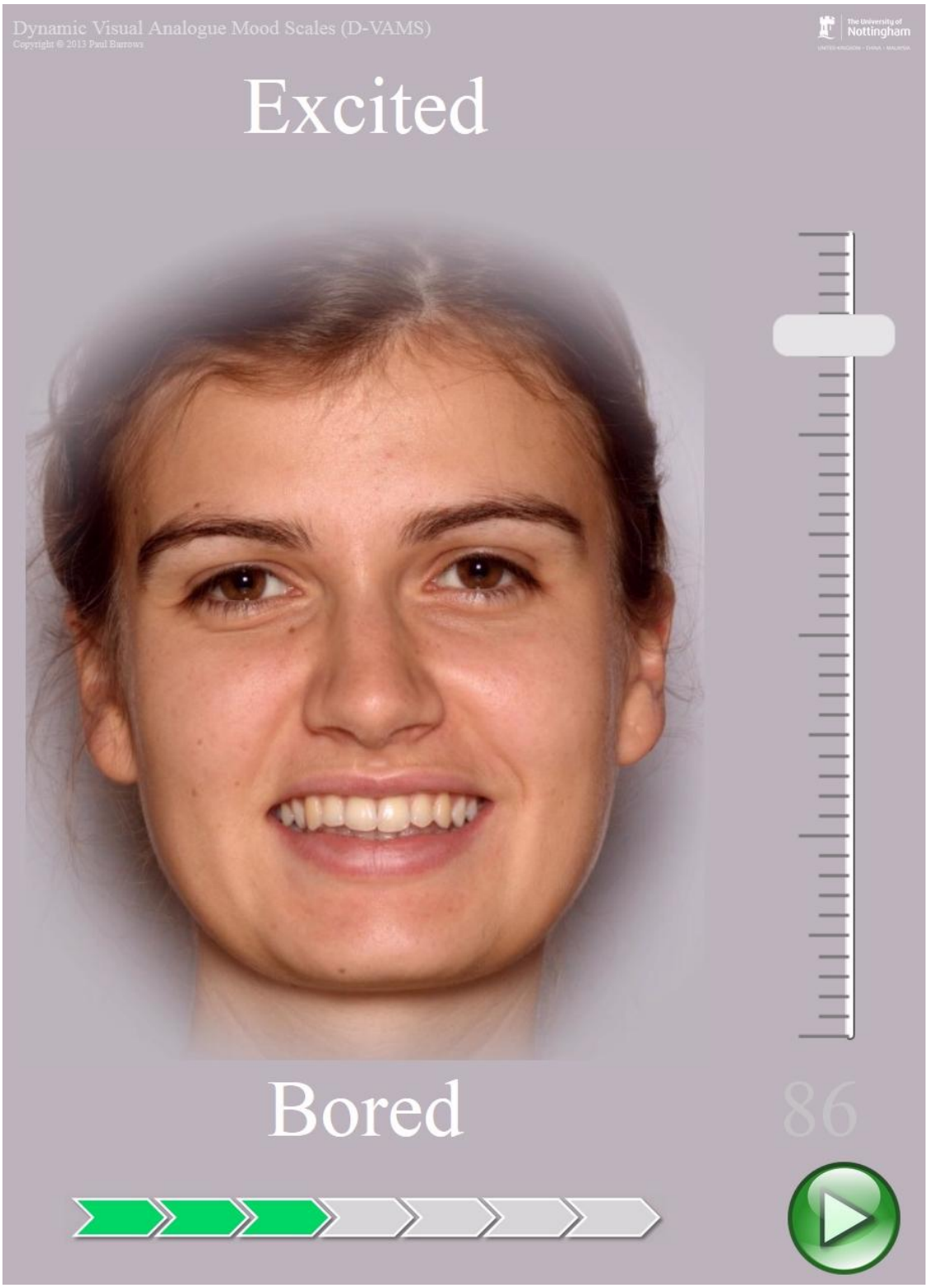


Figure 1. *D-VAMS Scale Response Page*

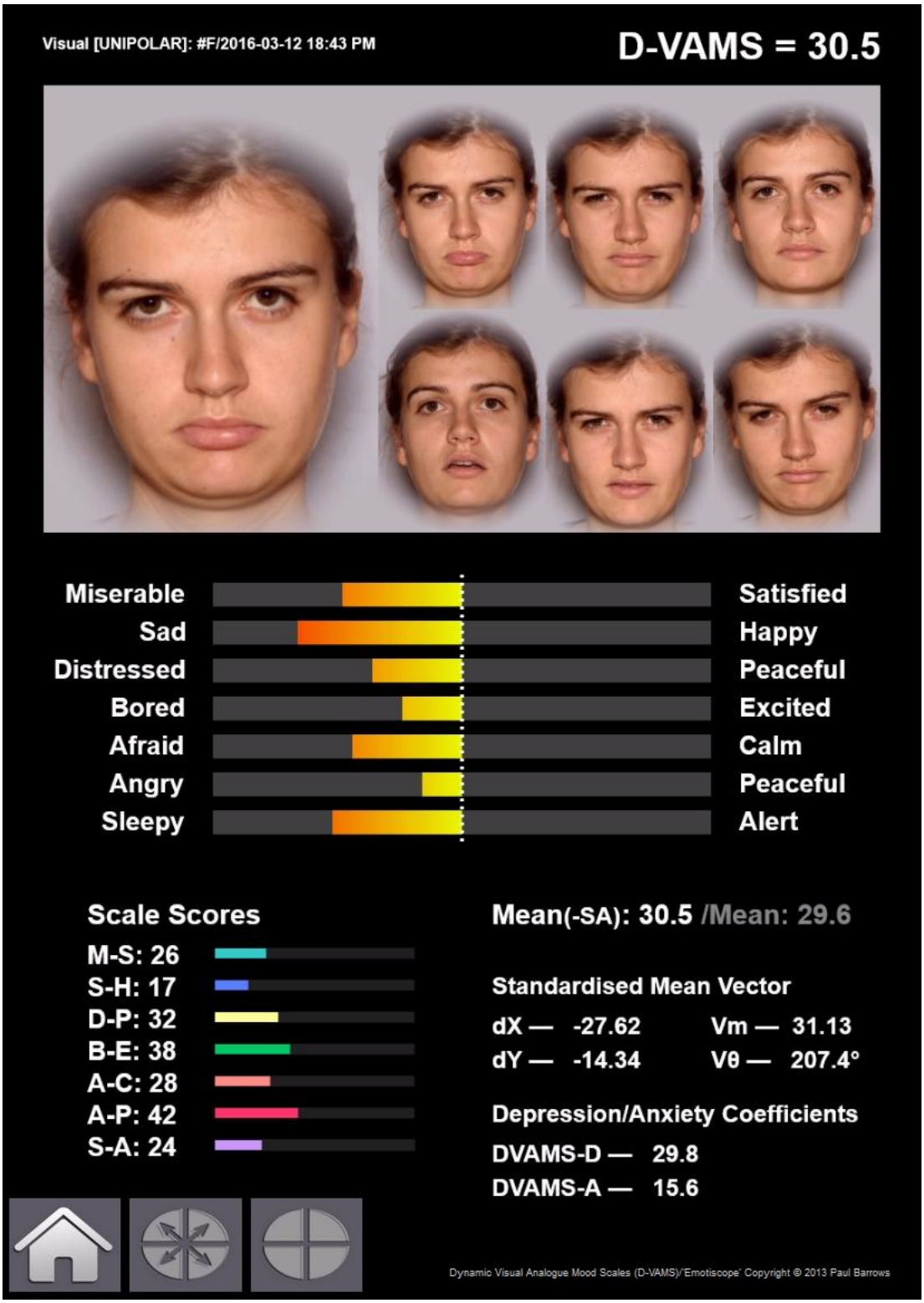


Figure 2. *D-VAMS Results Page - Faces View*

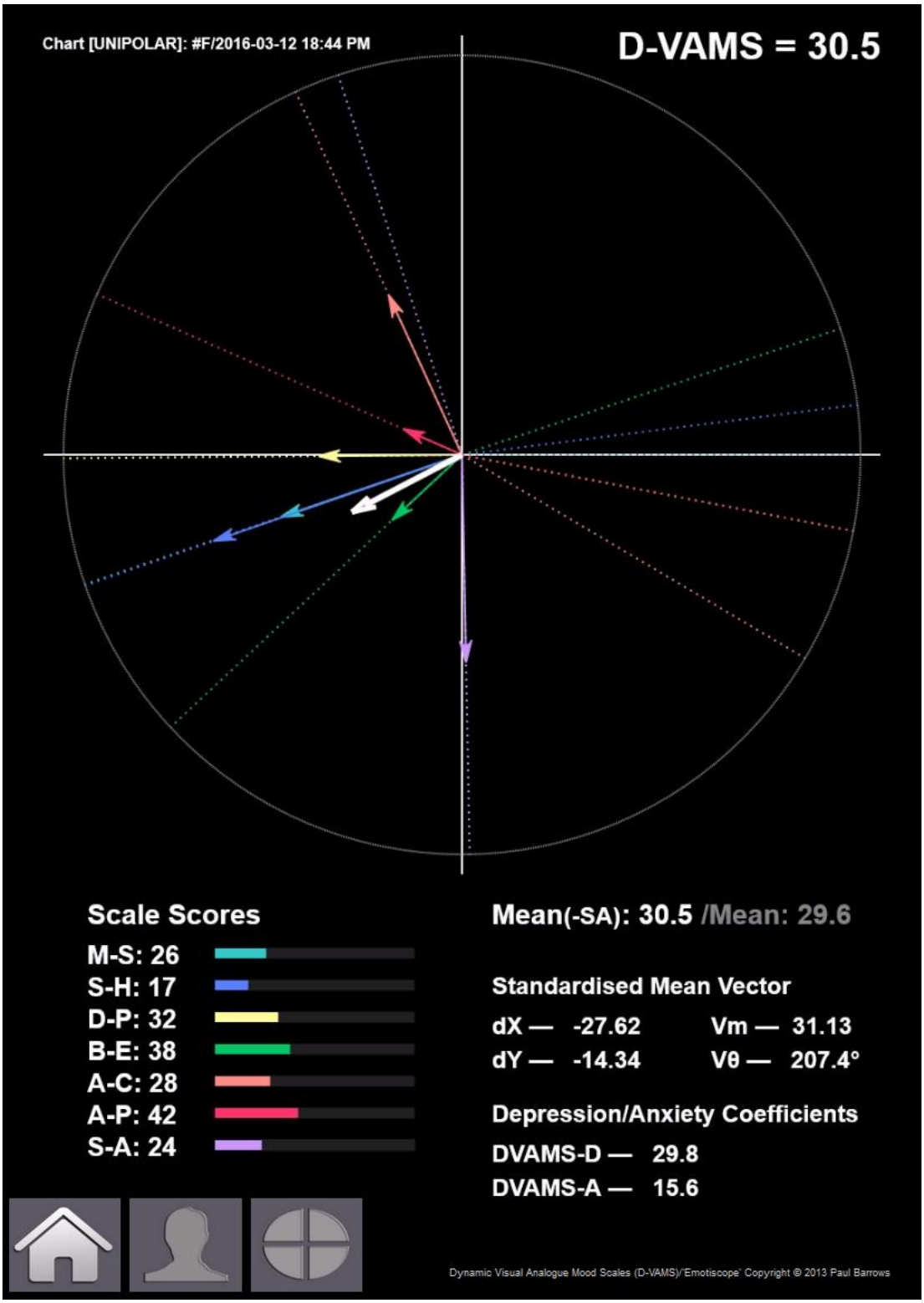


Figure 3. *D-VAMS Results Page - Chart View(Unipolar Mode)*

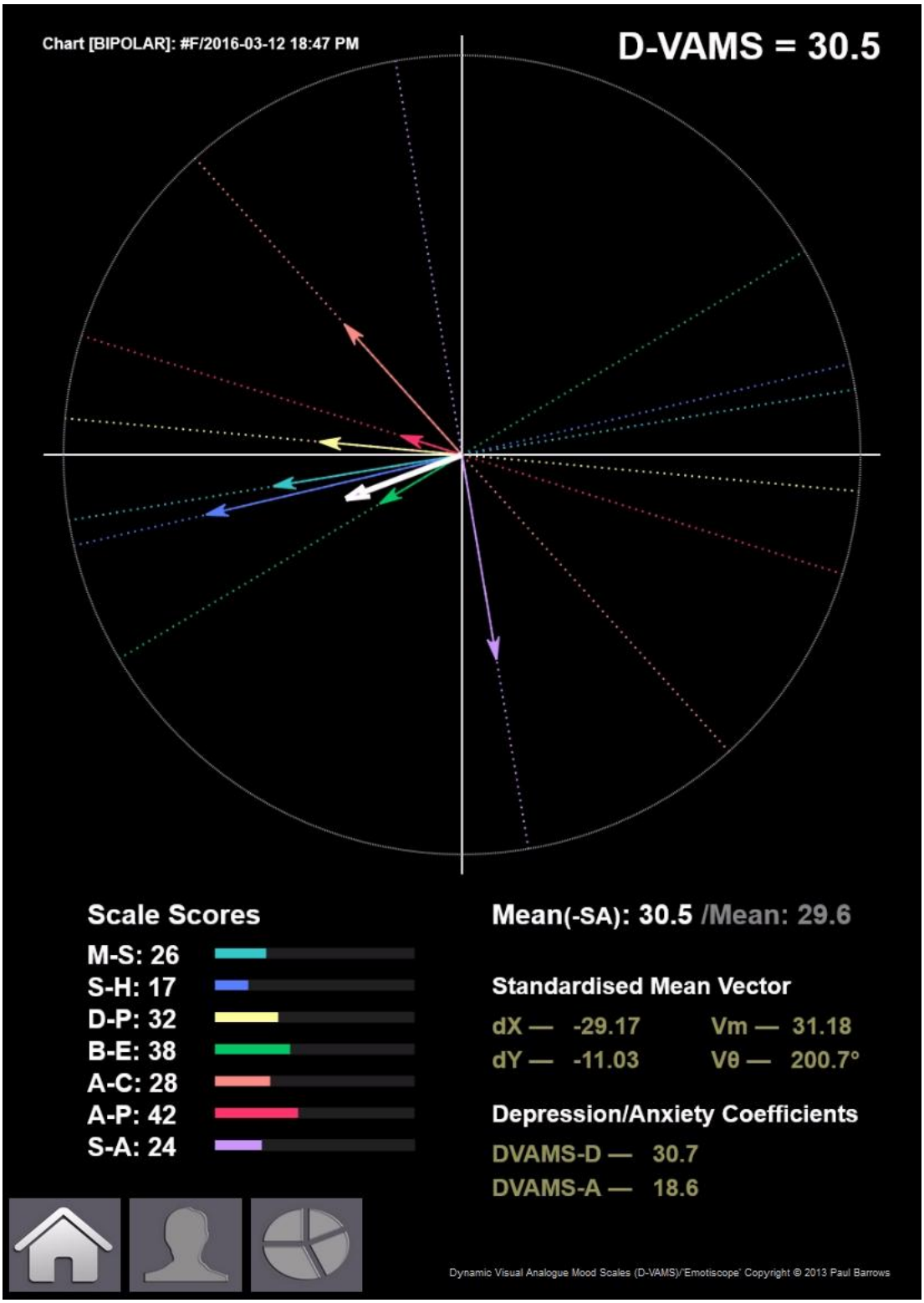


Figure 4. *D-VAMS Results Page - Chart View(Bipolar Mode)*

References

Abelson, R. P., & Sermat, V. (1962). Multidimensional scaling of facial expressions. *Journal of Experimental Psychology, 64*(6), 546-554. doi: 10.1037/h0042280

Aben, I., Verhey, F., lousberg, R., Lodder, J., & Honig, A. (2002). Validity of the beck depression inventory, hospital anxiety and depression scale, SCL-90, and hamilton depression rating scale as screening instruments for depression in stroke patients. *Psychosomatics, 43*(5), 386-393. doi: 10.1176/appi.psy.43.5.386

Acton, G. S., & Revelle, W. (2002). Interpersonal personality measures show circumplex structure based on new psychometric criteria. *Journal of Personality Assessment, 79*(3), 446-471. doi: 10.1207/S15327752JPA7903_04

Acton, G. S., & Revelle, W. (2004). Evaluation of ten psychometric criteria for circumplex structure. *Methods of Psychological Research - Online, 9*(1), 1-27.

Adolphs, R., Damasio, H., Tranel, D., & Damasio, R. (1996). Cortical systems for the recognition of emotion in facial expressions. *The Journal of Neuroscience, 16*(23), 7678-7687. doi: http://www.jneurosci.org/content/16/23/7678.long

Adshead, F., Cody, D. D., & Pitt, B. (1992). BASDEC: a novel screening instrument for depression in elderly medical inpatients. *BMJ, 305*(6850), 397. doi: 10.1136/bmj.305.6850.397

Ahearn, E. P. (1997). The use of visual analog scales in mood disorders: a critical review. *Journal of Psychiatric Research, 31*(5), 569-579. doi: 10.1016/S0022-3956(97)00029-0

Altieri, M., Maestrini, I., Mercurio, A., Troisi, P., Sgarlata, E., Rea, V., Di Piero, V., & Lenzi, G. L. (2012). Depression after minor stroke: prevalence and predictors. *European Journal of Neurology, 19*(3), 517-521. doi: 10.1111/j.1468-1331.2011.03583.x

American Psychiatric Association. (1980). Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders (3rd ed., revised). Washington, DC: Author.

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., text rev.). Washington, DC: Author.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: Author. doi: 10.1176/appi.books.9780890425596

Andersen, G. (1997). Post-stroke depression: Diagnosis and incidence. *European Psychiatry, 12 Suppl 3*, 255s-260s. doi: 10.1016/S0924-9338(97)89093-1

Andersen, G., Vestergaard, K., Riis, J., & Lauritzen, L. (1994). Incidence of post-stroke depression during the first year in a large unselected stroke population determined using a valid standardized rating scale. *Acta Psychiatrica Scandinavica, 90*(3), 190-195. doi: 10.1111/j.1600-0447.1994.tb01576.x

Arruda, J. E., Stern, R. A., & Legendre, S. A. (1997). Assessment of mood state in patients undergoing electroconvulsive therapy: the utility of visual analog mood scales developed for cognitively impaired patients. *Convulsive Therapy, 12*(4), 207-212.

Arruda, J. E., Stern, R. A., & Somerville, J. A. (1999). Measurement of mood states in stroke patients: validation of the visual analog mood scales. *Archives of Physical Medicine and Rehabilitation, 80*(6), 676-680. doi: 10.1016/S0003-9993(99)90171-5

Astrom, M., Adolfsson, R., & Asplund, K. (1993). Major depression in stroke patients. A 3-year longitudinal study. *Stroke, 24*(7), 976-982. doi: 10.1161/01.STR.24.7.976

Averill, J. R. (1975). *A semantic atlas of emotional concepts*: American Psychological Association.

Ayerbe, L., Ayis, S., Rudd, A. G., Heuschmann, P. U., & Wolfe, C. D. (2011). Natural history, predictors, and associations of depression 5 years after stroke: the South London Stroke Register. *Stroke, 42*(7), 1907-1911. doi: 10.1161/STROKEAHA.110.605808

Bacher, Y., Korner-Bitensky, N., Mayo, N., Becker, R., & Coopersmith, H. (1990). A longitudinal study of depression among stroke patients participating in a rehabilitation program. *Canadian Journal of Rehabilitation, 4*(1), 27-37.

Barker-Collo, S. L. (2007). Depression and anxiety 3 months post stroke: prevalence and correlates. *Archives of Clinical Neuropsychology, 22*(4), 519-531. doi: 10.1016/j.acn.2007.03.002

Bartoli, F. L., A., Annamaria, L., Crocamo, C., Mantero, V., Carra, G., Agostoni, E., & Clerici, M. (2013). Depression after Stroke and Risk of Mortality: A Systematic Review and Meta-Analysis. *Stroke Research and Treatment, 2013*, 1-11. doi: 10.1155/2013/862978

Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., & Shinkareva, S. V. (2012). Decoding the neural representation of affective states. *Neuroimage, 59*(1), 718-727. doi: 10.1016/j.neuroimage.2011.07.037

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment, 67*(3), 588-597. doi: 10.1207/s15327752jpa6703_13

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571. doi: 10.1001/archpsyc.1961.01710120031004

Benaim, C., Cailly, B., Perennou, D., & Pelissier, J. (2004). Validation of the aphasic depression rating scale. *Stroke, 35*(7), 1692-1696. doi: 10.1161/01.STR.0000130591.95710.20

Benaim, C., Decavel, P., Bentabet, M., Froger, J., Pelissier, J., & Perennou, D. (2010). Sensitivity to change of two depression rating scales for stroke patients. *Clinical Rehabilitation, 24*(3), 251-257. doi: 10.1177/0269215509346087

Bennett, H. E., Thomas, S. A., Austen, R., Morris, A. M., & Lincoln, N. B. (2006). Validation of screening measures for assessing mood in stroke patients. *British Journal of Clinical Psychology, 45*(Pt 3), 367-376. doi: 10.1348/014466505X58277

Benson, D. F. (1973). Psychiatric aspects of aphasia. *Br J Psychiatry, 123*(576), 555-566. doi: 10.1192/bjp.123.5.555

Bentler, P. M. (1969). Semantic space is (approximately) bipolar. *Journal of Psychology, 71*, 33-40. doi: 10.1080/00223980.1969.10543067

Berg, A., Lonnqvist, J., Palomaki, H., & Kaste, M. (2009). Assessment of depression after stroke: a comparison of different screening instruments. *Stroke, 40*(2), 523-529. doi: 10.1161/STROKEAHA.108.527705

Berg, A., Palomaki, H., Lehtihalmes, M., Lonnqvist, J., & Kaste, M. (2003). Poststroke depression: an 18-month follow-up. *Stroke, 34*(1), 138-143. doi: 10.1161/01.STR.0000048149.84268.07

Bergersen, H., Froslie, K. F., Stibrant Sunnerhagen, K., & Schanke, A. K. (2010). Anxiety, depression, and psychological well-being 2 to 5 years poststroke. *Journal of Stroke and Cerebrovascular Diseases, 19*(5), 364-369. doi: 10.1016/j.jstrokecerebrovasdis.2009.06.005

Bhogal, S. K., Teasell, R., Foley, N., & Speechley, M. (2004). Lesion location and poststroke depression: systematic review of the methodological limitations in the literature. *Stroke, 35*(3), 794-802. doi: 10.1161/01.STR.0000117237.98749.26

Borgatta, E. F. (1961). Mood, personality and interaction. *Journal of General Psychology, 64*, 105-137. doi: 10.1080/00221309.1961.9920431

Borod, J. C., Cicero, B. A., Obler, L. K., & Whalen, J. R. (1998). Right hemisphere emotional perception: evidence across multiple channels. *Neuropsychology, 12*(3), 446-458. doi: 10.1037/0894-4105.12.3.446

Borod, J. C., Koff, E., Perlman Lorch, M., & Nicholas, M. (1986). The expression and perception of facial emotion in brain-damaged patients. *Neuropsychologia, 24*(2), 169-180. doi: 10.1016/0028-3932(86)90050-3

Bosanquet, N., & Franks, P. J. (1998). *Stroke care: reducing the burden of disease.* London: The Stroke Association.

Bourke, C., Douglas, K., & Porter, R. (2010). Processing of facial emotion expression in major depression: a review. *Australian and New Zealand Journal of Psychiatry, 44*(8), 681-696. doi: 10.3109/00048674.2010.496359

Bradburn, N. M. (1969). *The Structure of Psychological Well-Being*. Chicago: Aldine Publishing Company.

Brumfitt, S. M., & Sheeran, P. (1999). The development and validation of the Visual Analogue Self-Esteem Scale (VASES). *British Journal of Clinical Psychology, 38*, 387-400. doi: 10.1348/014466599162980

Bush, L. E. (1973). Individual differences in multidimensional scaling of adjectives denoting feelings. *Journal of Personality and Social Psychology, 25*(1), 50-57. doi: 10.1037/h0034274

Butters, M. A., Becker, J. T., Nebes, R. D., Zmuda, M. D., Mulsant, B. H., Pollock, B. G., & Reynolds, C. F., 3rd. (2000). Changes in cognitive functioning following treatment of late-life depression. *American Journal of Psychiatry, 157*(12), 1949-1954. doi: 10.1176/appi.ajp.157.12.1949

Carota, A., & Bogousslavsky, J. (2003). Poststroke depression. *Advances in Neurology, 92*, 435-445.

Carson, A. J., MacHale, S., Allen, K., Lawrie, S. M., Dennis, M., House, A., & Sharpe, M. (2000). Depression after stroke and lesion location: a systematic review. *Lancet, 356*(9224), 122-126. doi: 10.1016/S0140-6736(00)02448-X

Chemerinski, E., Robinson, R. G., & Kosier, J. T. (2001). Improved recovery in activities of daily living associated with remission of poststroke depression. *Stroke, 32*(1), 113-117. doi: 10.1161/01.STR.32.1.113

Clyde, D. J. (1963). *Manual for the Clyde Mood Scale*. Coral Gables, Florida: University of Miami Biometrics Laboratory.

Cobley, C. S., Thomas, S. A., Lincoln, N. B., & Walker, M. F. (2012). The assessment of low mood in stroke patients with aphasia: reliability and validity of the 10-item Hospital version of the Stroke Aphasic Depression Questionnaire (SADQH-10). *Clinical Rehabilitation, 26*(4), 372-381. doi: 10.1177/0269215511422388

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., Zhu, H., Kangarlu, A., Duan, Y., Russell, J. A., & Peterson, B. S. (2010). Neural systems subserving valence and arousal during the experience of induced emotions. *Emotion, 10*(3), 377-389. doi: 10.1037/a0018484

Crawford, J. R., & Henry, J. D. (2004). The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43*(Pt 3), 245-265. doi: 10.1348/0144665031752934

Cully, J. A., Gfeller, J. D., Heise, R. A., Ross, M. J., Teal, C. R., & Kunik, M. E. (2005). Geriatric depression, medical diagnosis, and functional recovery during acute rehabilitation. *Archives of Physical Medicine and Rehabilitation, 86*(12), 2256-2260. doi: 10.1016/j.apmr.2005.07.292

Dieguez, S., Staub, F., Bruggimann, L., & Bogousslavsky, J. (2004). Is poststroke depression a vascular depression? *Journal of the Neurological Sciences, 226*(1-2), 53-58. doi: 10.1016/j.jns.2004.09.012

Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology, 47*(5), 1105-1117. doi: 10.1037/0022-3514.47.5.1105

Dixon, J. S. (1986). Agreement between horizontal and vertical visual analogue scales. *Br J Rheumatol, 25*(4), 415-416. doi: 10.1093/rheumatology/25.4.415-a

Dixon, J. S., & Bird, H. A. (1981). Reproducibility along a 10 cm vertical visual analogue scale. *Annals of the Rheumatic Diseases, 40*(1), 87-89. doi: 10.1136/ard.40.1.87

Ebner, N. C., & Johnson, M. K. (2009). Young and older emotional faces: are there age-group differences in expression identification and memory? *Emotion, 9*(3), 329-339. doi: 10.1037/a0015179

Ekman, P. (1970). Universal facial expressions of emotions. *California Mental Health Digest, 8*(4), 151-158.

Ekman, P. (1992). Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of The Royal Society B Biological Sciences, 335*(1273), 63-69. doi: 10.1098/rstb.1992.0008

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384-392. doi: 10.1037/0003-066X.48.4.384

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin, 115*(2), 268-287.

Ekman, P. (2005). Facial Expressions. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* Chichester, UK: John Wiley & Sons, Ltd.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124-129. doi: 10.1037/h0030377

Ekman, P., & Friesen, W. V. (1976). *Pictures of Facial Affect*: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., & et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *J Pers Soc Psychol, 53*(4), 712-717. doi: 10.1037/0022-3514.53.4.712

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin, 128*(2), 203-235. doi: 10.1037/0033-2909.128.2.203

Ellis, C., Zhao, Y., & Egede, L. E. (2010). Depression and increased risk of death in adults with stroke. *Journal of Psychosomatic Research, 68*(6), 545-551. doi: 10.1016/j.jpsychores.2009.11.006

Engen, T., & Levy, N. (1956). Constant-sum judgments of facial expressions. *Journal of Experimental Psychology, 6*, 396-398.

Engen, T., Levy, N., & Schlosberg, H. (1957). A new series of facial expressions. *American Psychologist, 12*(5), 264-266. doi: 10.1037/h0049283

Engen, T., Levy, N., & Schlosberg, H. (1958). The dimensional analysis of a new series of facial expressions. *Journal of Experimental Psychology, 55*(5), 454-458. doi: 10.1037/h0047240

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi: 10.3758/BF03193146

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74*(4), 967-984. doi: 10.1037/0022-3514.74.4.967

Feldman, L. A. (1995). Variations in the Circumplex Structure of Mood. *Personality and Social Psychology Bulletin, 21*(1), 806-817. doi: 10.1177/0146167295218003

Feleky, A. M. (1914). The expression of the emotions. *Psychological Review, 21*, 33-41. doi: 10.1037/h0076058

Folstein, M. F., & Luria, R. (1973). Reliability, validity, and clinical application of the Visual Analogue Mood Scale. *Psychological Medicine, 3*(4), 479-486. doi: 10.1017/S0033291700054283

Folstein, M. F., Maiberger, R., & McHugh, P. R. (1977). Mood disorder as a specific complication of stroke. *Journal of Neurology, Neurosurgery and Psychiatry, 40*(10), 1018-1020. doi: 10.1136/jnnp.40.10.1018

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology, 14*, 83-102. doi: 10.1037/h0074329

Frois-Wittman, J. (1930). The judgment of facial expression. *Journal of Experimental Psychology, 1935*(13), 113-151. doi: 10.1037/h0070158

Gainotti, G., Azzoni, A., Gasparini, F., Marra, C., & Razzano, C. (1997a). Relation of lesion location to verbal and nonverbal mood measures in stroke patients. *Stroke, 28*(11), 2145-2149. doi: 10.1161/01.STR.28.11.2145

Gainotti, G., Azzoni, A., Razzano, C., lanzillotta, M., Marra, C., & Gasparini, F. (1997b). The post-stroke depression rating scale: a test specifically devised to investigate affective disorders of stroke patients. *Journa of Clinical and Experimental Neuropsychology, 19*(3), 340-356. doi: 10.1080/01688639708403863

Garra, G., Singer, A. J., Taira, B. R., Chohan, J., Cardoz, H., Chisena, E., & Thode, H. C., Jr. (2010). Validation of the Wong-Baker FACES Pain Rating Scale in

pediatric emergency department patients. *Academic Emergency Medicine*, *17*(1), 50-54. doi: 10.1111/j.1553-2712.2009.00620.x

Gift, A. G. (1989a). Validation of a vertical visual analogue scale as a measure of clinical dyspnea. *Rehabilitation Nursing, 14*(6), 323-325. doi: 10.1002/j.2048-7940.1989.tb01129.x

Gift, A. G. (1989b). Visual analogue scales: measurement of subjective phenomena. *Nursing Research, 38*(5), 286-288. doi: 10.1097/00006199-198909000-00006

Goodwin, R. D., & Devanand, D. P. (2008). Stroke, depression, and functional health outcome among adults in the community *Journal of Geriatric Psychiatry and Neurology, 21*(1), 41-46. doi: 10.1177/0891988707311041

Gordon, W. A., & Hibbard, M. A. (1997). Post-stroke depression: An examination of the literature. *Archives of Physical Medicine and Rehabilitation*, *78*, 658-663. doi: 10.1016/S0003-9993(97)90433-0

Gorsuch, R. L. (1983). *Factor analysis* (2 ed.). Hillsdale, NJ: Erbaum.

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology, 64*, 1029-1041. doi: 10.1037/0022-3514.64.6.1029

Guttman, L. A. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. New York: Columbia University Press.

Hacker, V. L., Stark, D., & Thomas, S. (2010). Validation of the stroke aphasic depression questionnaire using the brief assessment schedule depression cards in an acute stroke sample. *British Journal of Clinical Psychology, 49*(Pt 1), 123-127. doi: 10.1348/014466509X467440

Hackett, M. L., & Anderson, C. S. (2005). Predictors of depression after stroke: a systematic review of observational studies. *Stroke, 36*(10), 2296-2301. doi: 10.1161/01.STR.0000183622.75135.a4

Hackett, M. L., & Pickles, K. (2014). Part I: frequency of depression after stroke: an updated systematic review and meta-analysis of observational studies. *International Journal of Stroke, 9*(8), 1017-1025. doi: 10.1111/ijs.12357

Hackett, M. L., Yapa, C., Parag, V., & Anderson, C. S. (2005). Frequency of depression after stroke: a systematic review of observational studies. *Stroke, 36*(6), 1330-1340. doi: 10.1161/01.STR.0000165928.19135.35

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry, 23*, 56-62. doi: 10.1136/jnnp.23.1.56

Hammond, M. F., O'Keeffe, S. T., & Barer, D. H. (2000). Development and validation of a brief observer-rated screening scale for depression in elderly medical patients. *Age and Ageing, 29*(6), 511-515. doi: 10.1093/ageing/29.6.511

Hatano, S. (1976). Experience from a multicentre stroke register: a preliminary report. *Bulletin of World Health Organisation, 54*, 541.

Hayes, M. H. S., & Paterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin, 18*, 98-99.

Higgins, E. T., Klein, R., & Strauman, T. (1985). Self-concept discrepancy theory: A psychological model for distinguishing among different aspects of anxiety and depression. *Social Cognition, 3*, 51-76. doi: 10.1521/soco.1985.3.1.51

Hommel, M., Carey, L., & Jaillard, A. (2015). Depression: Cognition relations after stroke. *International Journal of Stroke, 10*(6), 893-896. doi: 10.1111/ijs.12057

House, A., Dennis, M., Hawton, K., & Warlow, C. (1989). Methods of identifying mood disorders in stroke patients: experience in the Oxfordshire Community Stroke Project. *Age and Ageing, 18*(6), 371-379. doi: 10.1093/ageing/18.6.371

Hulin, W. S., & Katz, D. (1935). The Frois-Wittman pictures of facial expression. *Journal of Experimental Psychology, 18*, 482-498. doi: 10.1037/h0056770

Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*(2), 288-299. doi: 10.1037/0033-2909.115.2.288

Kanner, L. (1931). Judging emotions from facial expressions. *Psychological Review Publications, 41*(186), 1-98. doi: 10.1086/216004

Kauhanen, M., Korpelainen, J. T., Hiltunen, P., Brusin, E., Mononen, H., Maatta, R., Nieminen, P., Sotaniemi, K. A., & Myllyla, V. V. (1999). Poststroke depression correlates with cognitive impairment and neurological deficits. *Stroke, 30*(9), 1875-1880. doi: 10.1161/01.STR.30.9.1875

Kauhanen, M. L., Korpelainen, J. T., Hiltunen, P., Maatta, R., Mononen, H., Brusin, E., Sotaniemi, K. A., & Myllyla, V. V. (2000). Aphasia, depression, and non-verbal cognitive impairment in ischaemic stroke. *Cerebrovascular Diseases, 10*(6), 455-461. doi: 10.1159/000016107

Kontou, E., Thomas, S. A., & Lincoln, N. B. (2012). Psychometric properties of a revised version of the Visual Analog Mood Scales. *Clinical Rehabilitation, 26*(12), 1133-1140. doi: 10.1177/0269215512442670

Kucharska-Pietura, K., Phillips, M. L., Gernand, W., & David, A. S. (2003). Perception of emotions from faces and voices following unilateral brain damage. *Neuropsychologia, 41*(8), 1082-1090. doi: 10.1016/S0028-3932(02)00294-4

Kutlubaev, M., & Hackett, M. L. (2014). Part II: predictors of depression after stroke and impact of depression on stroke outcome: an updated systematic review of observational studies. *International Journal of Stroke, 9*(8), 1026-1036. doi: 10.1111/ijs.12356

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Emotion: Review of personality & social psychology* (pp. 25-29). Newbury Park: Sage.

Lee, A. C., Tang, S. W., Yu, G. K., & Cheung, R. T. (2008). The smiley as a simple screening tool for depression after stroke: a preliminary study. *Int J Nurs Stud, 45*(7), 1081-1089. doi: 10.1016/j.ijnurstu.2007.05.008

Leeds, L., Meara, R. J., & Hobson, J. P. (2004). The utility of the Stroke Aphasia Depression Questionnaire (SADQ) in a stroke rehabilitation unit. *Clinical Rehabilitation, 18*(2), 228-231. doi: 10.1191/0269215504cr685oa

Lenze, E. J., Rogers, J. C., Martire, L. M., Mulsant, B. H., Rollman, B. L., Dew, M. A., Schulz, R., & Reynolds, C. F., 3rd. (2001). The association of late-life depression and anxiety with physical disability: a review of the literature and prospectus for future research. *American Journal of Geriatric Psychiatry, 9*(2), 113-135. doi: 10.1097/00019442-200105000-00004

Lewis, S. C., Dennis, M. S., O'Rourke, S. J., & Sharpe, M. (2001). Negative attitudes among short-term stroke survivors predict worse long-term survival. *Stroke, 32*(7), 1640-1645. doi: 10.1161/01.STR.32.7.1640. 2001;32:1640-1645

Lincoln, N. B., Brinkmann, N., Cunningham, S., Dejaeger, E., De Weerdt, W., Jenni, W., Mahdzir, A., Putman, K., Schupp, W., Schuback, B., & De Wit, L. (2012a). Anxiety and depression after stroke: a 5 year follow-up. *Disabil Rehabil*. doi: 10.3109/09638288.2012.691939

Lincoln, N. B., Kneebone, I. I., Macniven, A. B., & Morris, R. C. (2012b). *Psychological Management of Stroke*. Chichester: John Wiley & Sons, Ltd. .

Lincoln, N. B., Nicholl, C. R., Flannaghan, T., Leonard, M., & Van der Gucht, E. (2003). The validity of questionnaire measures for assessing depression after stroke. *Clin Rehabil, 17*(8), 840-846. doi: 10.1191/0269215503cr687oa

Linden, T., Blomstrand, C., & Skoog, I. (2007). Depressive disorders after 20 months in elderly stroke patients: a case-control study. *Stroke, 38*(6), 1860-1863. doi: 10.1161/STROKEAHA.106.471805

Lipsey, J. R., Robinson, R. G., Pearlson, G. D., Rao, K., & Price, T. R. (1983). Mood change following bilateral hemisphere brain injury. *British Journal of Psychiatry, 143*, 266-273. doi: 10.1192/bjp.143.3.266

Lorish, C. D., & Maisiak, R. (1986). The face scale: A brief, nonverbal method for assessing patient mood *Arthritis and Rheumatism, 29*(7), 906-909. doi: 10.1002/art.1780290714

Lorr, M., Daston, P., & Smith, I. R. (1967). An analysis of mood states. *Educational and Psychological Measurement, 27*(1), 89-96. doi: 10.1177/001316446702700109

Luria, R. E. (1975). The validity and reliability of the visual analogue mood scale. *Journal of Psychiatric Research, 12*(1), 51-57. doi: 10.1016/0022-3956(75)90020-5

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149. doi: 10.1037/1082-989X.1.2.130

McDowell, I. (2006). *A Guide to Rating Scales and Questionnaires.* (3rd ed.). Oxford: Oxford University Press.

McNair, D. M., Lorr, M., & Droppleman, L. (1971). *Manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Service.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: The MIT Press.

Mill, A., Allik, J., Realo, A., & Valk, R. (2009). Age-related differences in emotion recognition ability: a cross-sectional study. *Emotion, 9*(5), 619-630. doi: 10.1037/a0016562

Mitchell, A. J., Baker-Glenn, E. A., Granger, L., & Symonds, P. (2010). Can the Distress Thermometer be improved by additional mood domains? Part I. Initial validation of the Emotion Thermometers tool. *Psychooncology, 19*(2), 125-133. doi: 10.1002/pon.1523

Montagne, B., Nys, G. M., Zandvoort, M. J. E., Kappelle, L. J., De Haan, E. H., & Kessels, R. P. (2007). The perception of emotional facial expressions in stroke patients with and without depression. *Acta Neuropsychiatrica, 19*(5), 279-283. doi: 10.1111/j.1601-5215.2007.00235.x

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry, 134*, 382-389. doi: 10.1192/bjp.134.4.382

Moreno, C. R., Borod, J. C., Welkowitz, J., & Alpert, M. (1990). Lateralization for the expression and perception of facial emotion as a function of age. *Neuropsychologia, 28*(2), 199-209. doi: 10.1016/0028-3932(90)90101-S

Morris, P. L., Raphael, B., & Robinson, R. G. (1992). Clinical depression was associated with impaired recovery from stroke. *Medical Journal of Australia, 157*(4), 239-242. doi: 10.7326/ACPJC-1993-118-3-090

Morris, P. L., Robinson, R. G., Raphael, B., & Hopwood, M. J. (1996). Lesion location and poststroke depression. *Journal of Neuropsychiatry and Clinical Neurosciences, 8*(4), 399-403. doi: 10.1176/jnp.8.4.399

Murata, Y., Kimura, M., & Robinson, R. G. (2000). Does cognitive impairment cause post-stroke depression? *American Journal of Geriatric Psychiatry, 8*(4), 310-317. doi: 10.1097/00019442-200011000-00007

Narushima, K., Kosier, J. T., & Robinson, R. G. (2003). A reappraisal of poststroke depression, intra- and inter-hemispheric lesion location using meta-analysis. *Journal of Neuropsychiatry and Clinical Neurosciences, 15*(4), 422-430. doi: 10.1176/jnp.15.4.422

Nowlis, V., & Nowlis, H. H. (1956). The description and analysis of mood. *Annals of the New York Academy of Sciences, 65*, 345-355. doi: 10.1111/j.1749-6632.1956.tb49644.x

Nyenhuis, D. L., Stern, R. A., Yamamoto, C., Luchetta, T., & Arruda, J. E. (1997). Standardization and Validation of the Visual Analogue Mood Scales. *The Clinical Neuropsychologist, 11*(4), 407-415. doi: 10.1080/13854049708400470

Nys, G. M., van Zandvoort, M. J., van der Worp, H. B., de Haan, E. H., de Kort, P. L., & Kappelle, L. J. (2005). Early depressive symptoms after stroke: neuropsychological correlates and lesion characteristics. *Journal of the Neurological Sciences, 228*(1), 27-33. doi: 10.1016/j.jns.2004.09.031

Office for National Statistics. (2010). Mortality Statistics: Deaths registered in 2010 (Series DR) Tables 1–4 and Tables 6–14. Retrieved from http://www.ons.gov.uk/ons/rel/vsob1/mortality-statistics--deaths-registered-in-england-and-wales--series-dr-/2010/dr-table5-2010.xls

Osgood, C. E. (1966). Dimensionality of the semantic space for communication via facial expressions. *Scandinavian Journal of Psychology, 7*(1), 1-30. doi: 10.1111/j.1467-9450.1966.tb01334.x

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, Illinois: University of Illinois Press.

Ostir, G. V., Berges, I. M., Ottenbacher, A., & Ottenbacher, K. J. (2011). Patterns of change in depression after stroke. *Journal of the American Geriatric Society, 59*(2), 314-320. doi: 10.1111/j.1532-5415.2010.03266.x

Paolucci, S., Antonucci, G., Pratesi, L., Traballesi, M., Grasso, M. G., & Lubich, S. (1999). Poststroke depression and its role in rehabilitation of inpatients. *Archives of Physical Medicine and Rehabilitation, 80*(9), 985-990. doi: 10.1016/S0003-9993(99)90048-5

Paolucci, S., Gandolfo, C., Provinciali, L., Torta, R., & Toso, V. (2006). The Italian multicenter observational study on post-stroke depression (DESTRO). *Journal of Neurology, 253*(5), 556-562. doi: 10.1007/s00415-006-0058-6

Parikh, R. M., Robinson, R. G., Lipsey, J. R., Starkstein, S. E., Fedoroff, J. P., & Price, T. R. (1990). The impact of poststroke depression on recovery in activities of daily living over a 2-year follow-up. *Archives of Neurology, 47*(7), 785-789. doi: 10.1001/archneur.1990.00530070083014

Patil, V. H., Singh, S. N., Mishra, S., & Donavan, D. T. (2007). "Parallel Analysis Engine to Aid Determining Number of Factors to Retain" [Computer software]. Available from http://smishra.faculty.ku.edu/parallelengine.htm; Utility developed as part of Patil, Vivek H., Surendra N. Singh, Sanjay Mishra, and Todd Donovan (2008), "Efficient Theory Development and

Factor Retention Criteria: A Case for Abandoning the 'Eigenvalue Greater Than One' Criterion," Journal of Business Research, 61 (2), 162-170.

Philippi, C. L., Mehta, S., Grabowski, T., Adolphs, R., & Rudrauf, D. (2009). Damage to association fiber tracts impairs recognition of the facial expression of emotion. *The Journal of Neuroscience, 29*(48), 15089-15099. doi: 10.1523/JNEUROSCI.0796-09.2009

Pinoit, J. M., Bejot, Y., Rouaud, O., Benatru, I., Osseby, G. V., & Bonin, B. (2006). Post-stroke depression, another handicap. English translation of "Dépression après un accident vasculaire cérébral, un handicap supplémentaire. Presse Med, 35(12), 1789-1793." Retrieved 9 July, 2016, from http://www.researchgate.net/publication/6639434_Post-stroke_depression_another_handicap

Pohjasvaara, T., Leppavuori, A., Siira, I., Vataja, R., Kaste, M., & Erkinjuntti, T. (1998). Frequency and clinical determinants of poststroke depression. *Stroke, 29*(11), 2311-2317. doi: 10.1161/01.STR.29.11.2311

Pohjasvaara, T., Vataja, R., Leppavuori, A., Kaste, M., & Erkinjuntti, T. (2001). Depression is an independent predictor of poor long-term functional outcome post-stroke. *European Journal of Neurology, 8*(4), 315-319. doi: 10.1046/j.1468-1331.2001.00182.x

Posner, J., Russell, J. A., Gerber, A., Gorman, D., Colibazzi, T., Yu, S., Wang, Z., Kangarlu, A., Zhu, H., & Peterson, B. S. (2009). The neurophysiological bases of emotion: An fMRI study of the affective circumplex using emotion-denoting words. *Hum Brain Mapp, 30*(3), 883-895. doi: 10.1002/hbm.20553

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and

psychopathology. *Dev Psychopathol*, *17*(3), 715-734. doi: 10.1017/S0954579405050340

Price, C. I., Curless, R. H., & Rodgers, H. (1999). Can stroke patients use visual analogue scales? *Stroke, 30*(7), 1357-1361. doi: 10.1161/01.STR.30.7.1357

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*(1), 45-56. doi: 10.1016/0304-3959(83)90126-4

Quaranta, D., Marra, C., & Gainotti, G. (2008). Mood disorders after stroke: diagnostic validation of the poststroke depression rating scale. *Cerebrovascular Diseases, 26*(3), 237-243. doi: 10.1159/000147450

Remington, N. A., Fabrigar, L. R., & Visser, P. S. (2000). Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology, 79*(2), 286-300. doi: 10.1037/0022-3514.79.2.286

Reynolds, S. L., Haley, W. E., & Kozlenko, N. (2008). The impact of depressive symptoms and chronic diseases on active life expectancy in older Americans. *American Journal of Geriatric Psychiatry*, *16*(5), 425-432. doi: 10.1097/JGP.0b013e31816ff32e

Robinson, R. G. (2003). Poststroke depression: prevalence, diagnosis, treatment, and disease progression. *Biological Psychiatry*, *54*(3), 376-387. doi: 10.1016/S0006-3223(03)00423-2

Robinson, R. G., & Coyle, J. T. (1980). The differential effect of right versus left hemispheric cerebral infarction on catecholamines and behavior in the rat. *Brain Research*, *188*(1), 63-78. doi: 10.1016/0006-8993(80)90557-0

Robinson, R. G., Kubos, K. L., Starr, L. B., Rao, K., & Price, T. R. (1984a). Mood disorders in stroke patients. Importance of location of lesion. *Brain, 107 (Pt 1)*, 81-93. doi: 10.1093/brain/107.1.81

Robinson, R. G., & Spalletta, G. (2010). Poststroke depression: a review. *Canadian Journal of Psychiatry, 55*(6), 341-349. doi: 10.1176/appi.ajp.2015.15030363

Robinson, R. G., Starr, L. B., Kubos, K. L., & Price, T. R. (1983). A two-year longitudinal study of post-stroke mood disorders: findings during the initial evaluation. *Stroke, 14*(5), 736-741. doi: 10.1161/01.STR.14.5.736

Robinson, R. G., Starr, L. B., Lipsey, J. R., Rao, K., & Price, T. R. (1984b). A two-year longitudinal study of post-stroke mood disorders: dynamic changes in associated variables over the first six months of follow-up. *Stroke, 15*(3), 510-517. doi: 10.1161/01.STR.15.3.510

Robinson, R. G., & Szetela, B. (1981). Mood change following left hemispheric brain injury. *Annals of Neurology, 9*(5), 447-453. doi: 10.1002/ana.410090506

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton: Princeton University Press.

Roth, A. J., Kornblith, A. B., Batel-Copel, L., Peabody, E., Scher, H. I., & Holland, J. C. (1998). Rapid screening for psychologic distress in men with prostate carcinoma. *Cancer, 82*(10), 1904-1908. doi: 10.1002/(SICI)1097-0142(19980515)82:10<1904::AID-CNCR13>3.0.CO;2-X

Royal College of Physicians. (2012). National Clinical Guidelines for Stroke, 4th edition. Prepared by the Intercollegiate Stroke Working Party. London: RCP.

Ruckmick, C. A. (1921). A preliminary study of the emotions. *Psychological Monographs, 30*(3), 30-35. doi: 10.1037/h0093142

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews, 32*(4), 863-881. doi: 10.1016/j.neubiorev.2008.01.001

Russell, J. A. (1978). Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology, 36*(10), 1152-1168.

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology, 37*(3), 345-356. doi: 10.1037/0022-3514.37.3.345

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178. doi: 10.1037/h0077714

Russell, J. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502. doi: 10.1037/0022-3514.57.3.493

Russell, J. A. (1995). Facial expressions of emotion: what lies beyond minimal universality? *Psychological Bulletin, 118*(3), 379-391. doi: 10.1037/0033-2909.118.3.379

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*(5), 805-819. doi: 10.1037//0022-3514.76.5.805

Russell, J. A., & Bullock, M. (1985). Multidimensional scaling of emotional facial expressions: Similarity from preschoolers to adults. *Journal of Personality and Social Psychology, 48*(5), 1290-1298. doi: 10.1037/0022-3514.48.5.1290

Russell, J. A., & Carroll, J. M. (1999a). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*(1), 3-30. doi: 10.1037/0033-2909.125.1.3

Russell, J. A., & Carroll, J. M. (1999b). The phoenix of bipolarity: Reply to Watson and Tellegen (1999). *Psychological Bulletin, 125*(5), 611-617. doi: 10.1037/0033-2909.125.5.611

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*, 273-294.

Sagen, U., Vik, T. G., Moum, T., Mørland, T., Finset, A., & Toril, D. (2009). Screening for anxiety and depression after stroke: Comparison of the Hospital Anxiety and Depression Scale and the Montgomery and Åsberg Depression Rating Scale. *Journal of Psychosomatic Research, 67*(4), 325-332. doi: 10.1016/j.jpsychores.2009.03.007

Salter, K., Bhogal, S. K., Foley, N., Jutai, J., & Teasell, R. (2007). The assessment of poststroke depression. *Topics in Stroke Rehabilitation, 14*(3), 1-24. doi: 10.1310/tsr1403-1

Salter, K., Teasell, R., Bitensky, J., Foley, N., & Bhogal, S. (2009). Chapter 18: Post-Stroke Depression *The Evidence Based Review of Stroke Rehabilitation*: www.ebrsr.com.

Saxena, S. K., Ng, T. P., Yong, D., Fong, N. P., & Koh, G. (2008). Subthreshold depression and cognitive impairment but not demented in stroke patients during their rehabilitation. *Acta Neurologica Scandinavica, 117*(2), 133-140. doi: 10.1111/j.1600-0404.2007.00922.x

Scarborough, P., Peto, V., Bhatnagar, P., Kaur, A., Leal, J., Luengo-Fernandez, R., Gray, A., Rayner, M., & Allender, S. (2009). Stroke Statistics: 2009 Edition. *British Heart Foundation*.

Schlosberg, H. (1941). A scale for the judgment of facial expressions. *Journal of Experimental Psychology, 29*, 497-510. doi: 10.1037/h0061489

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44*(4), 229-237. doi: 10.1037/h0055778

Schlosberg, H. (1954). Three dimensions of emotion. *The Psychological Review, 61*(2), 81-88. doi: 10.1037/h0054570

Schulz, R., Beach, S. R., Ives, D. G., Martire, L. M., Ariyo, A. A., & Kop, W. J. (2000). Association between depression and mortality in older adults: the Cardiovascular Health Study. *Archives of Internal Medicine, 160*(12), 1761-1768. doi: 10.1001/archinte.160.12.1761

Scott, J., & Huskisson, E. C. (1976). Graphic representation of pain. *Pain, 2*(2), 175-184. doi: 10.1016/0304-3959(76)90113-5

Scott, J., & Huskisson, E. C. (1979). Vertical or horizontal visual analogue scales. *Annals of the Rheumatic Diseases, 38*(6), 560. doi: 10.1136/ard.38.6.560

Shewan, C. M., & Kertesz, A. (1980). Reliability and validity characteristics of the Western Aphasia Battery (WAB). *Journal of Speech and Hearing Disorders, 45*, 308-324. doi: 10.1044/jshd.4503.308

Sieger, T., Serranova, T., Ruzicka, F., Vostatek, P., Wild, J., Stastna, D., Bonnet, C., Novak, D., Ruzicka, E., Urgosik, D., & Jech, R. (2015). Distinct populations of neurons respond to emotional valence and arousal in the human subthalamic nucleus. *Proceedings of the National Academy of Sciences U S A, 112*(10), 3116-3121. doi: 10.1073/pnas.1410709112

Singh, A., Black, S. E., Herrmann, N., Leibovitch, F. S., Ebert, P. L., Lawrence, J., & Szalai, J. P. (2000). Functional and neuroanatomic correlations in poststroke depression: the Sunnybrook Stroke Study. *Stroke, 31*(3), 637-644. doi: 10.1161/01.STR.31.3.637

Singh, A., Herrmann, N., & Black, S. E. (1998). The importance of lesion location in poststroke depression: a critical review. *Canadian Journal of Psychiatry, 43*(9), 921-927. doi: 10.1177/070674379804300907

Sinyor, D., Amato, P., Kaloupek, D. G., Becker, R., Goldenberg, M., & Coopersmith, H. (1986). Post-stroke depression: relationships to functional impairment, coping strategies, and rehabilitation outcome. *Stroke, 17*(6), 1102-1107.

Sriwatanakul, K., Kelvie, W., Lasagna, L., Calimlim, J. F., Weis, O. F., & Mehta, G. (1983). Studies with different types of visual analog scales for measurement of pain. *Clinical Pharmacology & Therapeutics, 34*(2), 234-239. doi: 10.1038/clpt.1983.159

Stanley, D. J., & Meyer, J. P. (2009). Two-dimensional affective space: a new approach to orienting the axes. *Emotion, 9*(2), 214-237. doi: 10.1037/a0014612

Stern, R. A. (1999). Assessment of mood states in aphasia. *Seminars in Speech and Language, 20*(1), 33-49; quiz 49-50. doi: 10.1055/s-2008-1064007

Stern, R. A., Arruda, J. E., Hooper, C. R., Wolfner, G. D., & Morey, C. E. (1997). Visual analogue mood scales to measure internal mood state in neurologically impaired patients : description and initial validity evidence. *Aphasiology, 11*(1), 59-71. doi: 10.1080/02687039708248455

Stern, R. A., & Bachman, D. L. (1991). Depressive symptoms following stroke. *American Journal of Psychiatry, 148*(3), 351-356. doi: 10.1176/ajp.148.3.351

Stern, R. A., Hooper, C. R., & Morey, C. E. (1990). Development of visual analogue scales to measure mood in aphasia. *The Clinical Neuropsychologist, 4*, 300.

Stewart-Knight, K., Parry, R., Abey, A., & Seymours, J. (2012). Does the distress thermometer work? A systematic review of the evidence for its use and

validity. *BMJ Supportive & Palliative Care, 2*(A30). doi: 10.1136/bmjspcare-2012-000196.87

Sutcliffe, L. M., & Lincoln, N. B. (1998). The assessment of depression in aphasic stroke patients: the development of the Stroke Aphasic Depression Questionnaire. *Clinical Rehabilitation, 12*(6), 506-513. doi: 10.1191/026921598672167702

Temple, R. O., Stern, R. A., Latham, J., Ruffolo, J. S., Arruda, J. E., & Tremont, G. (2004). Assessment of mood state in dementia by use of the visual analog mood scales (VAMS). *American Journal of Geriatric Psychiatry, 12*(5), 527-530. doi: 10.1176/appi.ajgp.12.5.527

Thayer, R. E. (1967). Measurement of activation through self-report. *Psychological Reports, 20*(2), 663-678. doi: 10.2466/pr0.1967.20.2.663

Thayer, R. E. (1978). Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion, 2*(1), 1-34. doi: 10.1007/BF00992729

Thayer, R. E. (1986). Activation-deactivation adjective check list: current overview and structural analysis. *Psychological Reports, 58*(2), 607-614. doi: 10.2466/pr0.1986.58.2.607

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. New York: Oxford University Press.

The Stroke Association. (2006). R11 Stroke Statistics Resource Sheet. Retrieved from http://www.stroke.org.uk/document.rm?id=877

The Stroke Association. (2015). State of the nation stroke statistics - January 2015. Retrieved from https://www.stroke.org.uk/sites/default/files/stroke_statistics_2015.pdf

Thomas, S. A., & Lincoln, N. B. (2006). Factors relating to depression after stroke. *British Journal of Clinical Psychology, 45*(Pt 1), 49-61. doi: 10.1348/014466505X34183

Thomas, S. A., & Lincoln, N. B. (2008). Predictors of emotional distress after stroke. *Stroke, 39*(4), 1240-1245. doi: 10.1161/STROKEAHA.107.498279

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Townend, B. S., Whyte, S., Desborough, T., Crimmins, D., Markus, R., Levi, C., & Sturm, J. W. (2007a). Longitudinal prevalence and determinants of early mood disorder post-stroke. *Journal of Clinical Neuroscience, 14*(5), 429-434. doi: 10.1016/j.jocn.2006.01.025

Townend, E., Brady, M., & McLaughlan, K. (2007b). A systematic evaluation of the adaptation of depression diagnostic methods for stroke survivors who have aphasia. *Stroke, 38*(11), 3076-3083. doi: 10.1161/STROKEAHA.107.484238

Townsend, N., Wickramasinghe, K., Bhatnagar, P., Smolina, K., Nichols, M., Leal, J., Luengo-Fernandez, R., & Rayner, M. (2012). *Coronary heart disease statistics 2012 edition*. London: British Heart Foundation.

Triandis, H. C., & Lambert, W. W. (1958). A restatement and test of Schlosberg's theory of emotion with two kinds of subjects from Greece. *Journal of Abnormal and Social Psychology, 56*(3), 321-328.

Turner-Stokes, L., Kalmus, M., Hirani, D., & Clegg, F. (2005). The Depression Intensity Scale Circles (DISCs): a first evaluation of a simple assessment tool for depression in the context of brain injury. *Journal of Neurology, Neurosurgery and Psychiatry, 76*(9), 1273-1278. doi: 10.1136/jnnp.2004.050096

van de Port, I. G., Kwakkel, G., van Wijk, I., & Lindeman, E. (2006). Susceptibility to deterioration of mobility long-term after stroke: a prospective cohort study. *Stroke, 37*(1), 167-171. doi: 10.1161/01.STR.0000195180.69904.f2

van Dijk, M. J., de Man-van Ginkel, J. M., B., H. T., & Schuurmans, M. J. (2016). Identifying depression post-stroke in patients with aphasia: A systematic review of the reliability, validity and feasibility of available instruments. *Clinical Rehabilitation, 30*(8), 795-810. doi: 10.1177/0269215515599665

Vastfjall, D., Friman, M., Garling, T., & Kleiner, M. (2002). The measurement of core affect: a Swedish self-report measure derived from the affect circumplex. *Scandinavian Journal of Psychology, 43*(1), 19-31. doi: 10.1111/1467-9450.00265

Velicer, W. F., Eation, C. A., & Fava, J. L. (2000). Chapter 3: Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determinng the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy* (1st ed., pp. 41-71). New York: Springer-Verlag.

Wade, D. T., Legh-Smith, J., & Hewer, R. A. (1987). Depressed mood after stroke. A community study of its frequency. *British Journal of Psychiatry, 151*, 200-205. doi: 10.1192/bjp.151.2.200

Watkins, C., Leathley, M., Daniels, L., Dickinson, H., Lightbody, C. E., van den Broek, M., & Jack, C. I. A. (2001). The signs of depression scale in stroke: How useful are nurses' observations? *Clinical Rehabilitation, 15*, 447.

Watson, D., & Clark, L. A. (1994). *Manual for the Positive and Negative Affect Schedule (expanded form)*. Unpublished Manuscript. University of Iowa. Iowa City.

Watson, D., & Clark, L. A. (1997). Measurement and mismeasurement of mood: recurrent and emergent issues. *Journal of Personality Assessment, 68*(2), 267-296. doi: 10.1207/s15327752jpa6802_4

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS. *Journal of Personality and Social Psychology, 54*, 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*(2), 219-235. doi: 10.1037/0033-2909.98.2.219

Watson, D., & Tellegen, A. (1999). Issues in the dimensional structure of affect - effects of descriptors, measurement error, and response formats: Comment on Russell and Carroll (1999). *Psychological Bulletin, 125*(5), 601. doi: 10.1037/0033-2909.125.5.601

Watson, D., Wiese, D., Vaidya, J., Tellegen, A., & Tellegen, A. (1999). The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology, 76*, 820-838. doi: 10.1037/0022-3514.76.5.820

West, R., Hill, K., Hewison, J., Knapp, P., & House, A. (2010). Psychological disorders after stroke are an important influence on functional outcomes: a prospective cohort study. *Stroke, 41*(8), 1723-1727. doi: 10.1161/STROKEAHA.110.583351

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health, 13*(4), 227-236. doi: 10.1002/nur.4770130405

Whyte, E. M., & Mulsant, B. H. (2002). Post stroke depression: epidemiology, pathophysiology, and biological treatment. *Biological Psychiatry, 52*(3), 253-264. doi: 10.1016/S0006-3223(02)01424-5

Whyte, E. M., Mulsant, B. H., Vanderbilt, J., Dodge, H. H., & Ganguli, M. (2004). Depression after stroke: a prospective epidemiological study. *Journal of the American Geriatrics Society, 52*(5), 774-778. doi: 10.1111/j.1532-5415.2004.52217.x

Willey, J. Z., Disla, N., Moon, Y. P., Paik, M. C., Sacco, R. L., Boden-Albala, B., Elkind, M. S., & Wright, C. B. (2010). Early depressed mood after stroke predicts long-term disability: the Northern Manhattan Stroke Study (NOMASS). *Stroke, 41*(9), 1896-1900. doi: 10.1161/STROKEAHA.110.583997

Wong, D. L., & Baker, C. M. (1988). Pain in children: comparison of assessment scales. *Pediatric Nursing, 14*(1), 9-17. doi: http://www.wongbakerfaces.org/wp-content/uploads/2010/08/pain-in-children.pdf

Woodworth, R. S. (1938). *Experimental psychology* (Vol. Henry Holt & Co.). New York.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1983). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of Psychiatric Research, 17*(1), 37-49. doi: 10.1016/0022-3956(82)90033-4

Yik, M. S. M., & Russell, J. A. (2003). Chinese affect circumplex: I. Structure of recalled momentary affect. *Asian Journal of Social Psychology, 6*(3), 185-200. doi: 10.1046/j.1467-839X.2003.00120.x

Yik, M. S. M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of Core Affect. *Emotion, 11*(4), 705-731. doi: 10.1037/a0023980

Zealley, A. K., & Aitken, R. C. (1969). Measurement of mood. *Proceedings of the Royal Society for Medicine, 62*(10), 993-996.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67*(6), 361-370. doi: 10.1111/j.1600-0447.1983.tb09716.x

Zung, W. W. (1965). A self-rating depression scale. *Arch Gen Psychiatry, 12*, 63-70. doi: 10.1001/archpsyc.1965.01720310065008