Investigation of the Humm Wadsworth Temperament Scale: Revision, Development and Application

This thesis is presented in partial fulfillment of the requirements for the degree

of

Master of Arts

in

Psychology

at Massey University, New Zealand

Kimberley Severinsen

2006

ABSTRACT

The present investigation examines the psychometric properties of a measure of temperament, the Humm Wadsworth Temperament Scale (Humm). To this end, participants (n = 27,245) completed the Humm questionnaire as part of either a recruitment and selection process initiated by a prospective employer, a promotion and development assessment initiated by their current employer, or career guidance advice sought of their own volition. Quantitative theoretical analysis based on Thurstone's method of paired comparisons and conceptual analysis by Humm experts and users were utilised for both the single-loading items for each of the seven components of the Humm, as well as the remaining multi-loading items. Thurstone's method was used to rank order items conceptually from 'best' predictor to 'worst' predictor of a certain component, which in turn were used to identify which items should remain in the Humm and which items should be discarded. The conceptual judgments generated by Humm experts and users, followed by confirmatory factor analysis, were used to increase the validity of the Humm through revising the set of items in the version of the Humm currently in use. The study concludes with a discussion of issues surrounding psychometric test revision, applicability of the Humm to the wider community including culturally diverse populations, as well as suggestions and recommendations for future research in this area.

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor, Dr Richard Fletcher for his encouragement, support and assistance throughout the research and thesis process. My appreciation is also extended to the human capital solutions company involved in this project for providing support for a scholarship application, and in particular thanks to my manager and the Research and Development team who have provided me with access to the psychological measurement instrument and data, without which, this thesis would not have been possible. I would also like to express my gratitude to the Tertiary Education Commission for providing me with support in the form of an enterprise scholarship. Finally, I would like to thank my colleagues, friends and family who have given me endless support, encouragement and motivation throughout my Masters study.

It is noted that this project was judged to be low risk. Notification of this was provided to the Massey University Human Ethics Committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

INTRODUCTION

The purpose of the current study was to evaluate the psychometric properties of a commercial psychological measurement instrument that is utilised to measure temperament, and attempt to improve the measure's statistical validity and interpretation. To investigate this, Thurstone's method of paired comparisons, conceptual judgements by experts and users of the measure, and confirmatory factor analysis methods were utilised in order to place the measure on a sound scientific foundation for further investigation and revision.

The measure chosen for the present study was the Humm Wadsworth Temperament Scale (Humm), developed by Humm and Wadsworth in 1935. The Humm is a psychological measurement instrument that a human capital solutions company utilises to measure temperament and predict behaviour. The current investigation aims to make the Humm more meaningful with respect to the validity and interpretation of the measure, as well as provide conceptual clarity with regards to the questionnaire items that are currently in the Humm. The present research will allow for further development and adaptation of the measure, and increase the ability to generalise the results across the wider community. The information will also aid in the possible adaptation, addition or deletion of current questionnaire items within the Humm. Initially the research will involve reviewing the Humm as a whole, and identifying any biased or ineffective test items through sound statistical analysis.

The concept of temperament and the operational measure, the Humm, are discussed, followed by an exposition on psychological instrument revision and measurement. The

quantitative technique of Thurstone's method of paired comparisons is also discussed, concluding with a summary of the rationale and aims of the current investigation.

The Concept of Temperament and Personality

In the past, the expressions temperament, character, and personality have been used to refer to what is now considered as the term personality (Endler, 1989). However, Endler suggests that temperament refers to the material that personality evolves from, whereas personality is the manner in which a person interacts with themselves and their environment. Whilst there has been much debate about whether personality and temperament are actually one term referring to the same concept (Strelau, 1987; Goldsmith & Campos, 1982), many researchers have continued to use the terms personality and temperament interchangeably (Pervin, 2002; Borkenau, 2001; Gray, 1973; Sheldon & Stevens, 1942). Furthermore, measures that have been designed to assess either temperament or personality may well have commonalities due to the possibility that they are actually measuring the same variables (Endler, 1989). Thus for the current study, the terms personality and temperament are assumed to be referring to the same concept and are indeed used interchangeably.

Interest in personality and temperament as a predictor of job performance has significantly grown in recent times. So too has the interest in personality measurement. This is in part due a growing number of studies demonstrating that the variables of an individual's personality can predict their future performance across a diverse range of occupations. Additionally, there are an increasing number of measures being made

available to assess temperament (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). This interest is further supported by other measures of personality that have been successfully used to predict a wide range of occupational performance criteria (Barrick & Mount, 1991, 1993; Hurtz & Donovan, 2000). Hogan and Nicholson (1988) also suggest that an appropriate methodology in many areas of personality and industrial psychology is that of personality assessment. It has now been widely published, acknowledged and accepted that temperament is a predictive measure of performance. When temperament is assessed using well-constructed, valid and reliable measures, the results can be used in personnel selection as a valid predictor of job performance across a variety of occupations (Salgado, 1999).

The term personality includes all factors entering into the make-up of an individual. This can include gender, physical appearance, aptitudes, abilities, talents, disposition and any other factor that may contribute to the whole person and differentiate them from other individuals (Humm & Wadsworth, 1935). Pervin, Cervone and John (2005) define personality as "those characteristics of the person that account for consistent patterns of feeling, thinking and behaving" (p.6). Temperament on the other hand, is used to designate those factors of personality that contribute to disposition, social reactions, emotional tone and attitudes. Temperament determines how a person will behave in a particular situation and how an individual will use their personal resources. In this instance reference is made to behaviour that is based on habits, feelings, attitudes and emotions, rather than behaviour based on purely rational grounds. Temperament is the non-rational and impulsive aspect of a person; for example, a person may possess a temperament style that sees them automatically taking a logical and unemotional stance. In fact, a person might exhibit this

style to such a degree so as to maintain a very unemotional attitude even though many people would suggest that an emotional response would be the "appropriate" response in a particular situation. In other words, even a rational style can be exhibited to an irrational degree. Some people will be naturally more inclined to exhibit "uncontrolled" temperament behaviour, leading to another important point on temperament, the issue of appropriateness. There is neither a right nor wrong temperament style. Being "strong" on one component or characteristic may not necessarily be better than being "weak". What is inappropriate, when referring to the workplace, is the fit between the person's temperament style, the job task at hand and the workplace environment. Temperament components are neutral and do not naturally carry positive or negative associations until placed within a context. An individual's balance of strengths and weaknesses will also vary from environment to environment. People possess all of the temperament characteristics to a certain degree. However, it is the relative degrees and blend of characteristics that create each person's individuality. Temperament is also quite enduring. Whilst significant life experience will change and alter a person, as will time and maturity, much of the impulsive and non-rational part of a person (their temperament) will remain the same. The question remains, if all people possess all characteristics of temperament, but to differing degrees, how are these characteristics measured? As temperament characteristics are displayed by people to varying degrees, it is possible that the differences and similarities can be compared and therefore can be measured.

Most people will generally display an "average" amount of any given temperament characteristic in comparison to the rest of the population. However, on closer inspection any given person can also fall outside what is considered displaying a "normal" amount of

a certain characteristic and it is these differences that can thus be measured. It is the areas where people exhibit characteristics to a greater or lesser degree than the average person that become apparent to the observer. These characteristics define a person's temperament style, and most people possess two or three components of temperament that are predominant (Pervin, Cervone & John, 2005; Bartram, 2005; Pervin, 2002). These dominant components will have the strongest influence on an individual's needs, goals and behaviour. Being relatively weak on one component can be somewhat compensated for by the presence of another component. In saying this, there is a dynamic balance of these characteristics. Some characteristics may be very weak within a person, whilst other characteristics may be very strong (Mayer, 2005). The implication being that the more a particular behaviour is exhibited, the stronger a characteristic is being represented. However, behaviours can also be exhibited due to a lack of a particular characteristic. Bartram (2004) adds that a person's unique characteristics are regarded as more important than qualifications, training or experience. This is because a person can be trained in order to develop new knowledge and skills. However, a person's attitude, honesty or way of dealing with people are characteristics that are relatively fixed and unchangeable. Therefore it is important to expand the factor of temperament further and consider ways in which to assess or discover an individual's predominant temperament 'style' (Humm & Wadsworth, 1935; Pervin, Cervone & John, 2005).

A component or characteristic of temperament refers to a combination of traits frequently found together and leading to behaviour that is recognisable as characteristic of that particular combination of traits. Further to the definition of personality provided by Pervin, Cervone and John (2005), a trait therefore refers to the consistent patterns of

behaviour, feeling and thought that an individual displays. A trait is a term used to describe a unit of behaviour that cannot be further subdivided, although it can manifest itself differently in different circumstances and in differing combinations with other traits. Behaviour resulting from such combinations of traits can sometimes be referred to as 'typical', however, it is noted that individuals rarely possess all the traits associated with any one characteristic or component and will generally demonstrate (through their behaviour) the possession of traits from several different components. A sub-component therefore refers to a combination of traits found within a component. A trait provides a useful method to summarise how one individual differs from another. Each trait may vary from weak to strong in its manifestations, so that the differences among people are both qualitative as to traits possessed and quantitative as to the strength of each trait. Moreover, the influence of the traits on each other is such that the quantitative differences in the traits themselves are expressed in qualitative differences through observable behaviour (Pervin, Cervone & John, 2005; Humm & Wadsworth, 1935; Humm, 1938).

Barrick and Mount (1991) and Tett, Jackson, and Rothstein (1991) conducted reviews of the relationship between personality and job performance, with particular focus on the Big Five model of personality. Both reviews reported that personality measurement was indeed useful for the prediction of an individual's future on the job performance. Humm and Wadsworth (1943) state that it is critical to understand temperament, because temperament is the combination of emotional tendencies that determine how an individual will react to situations that present throughout life. For example: whether a person is controlled or more emotionally reactive; whether an individual shoulders their responsibilities or evades them; whether they are loyal and trustworthy or unreliable; or

whether they are persistent or easily discouraged. Bartram (2004) adds that assessment is often carried out by an organisation as a method for predicting future on the job performance for both future and current employees, either as a part of a selection and recruitment process or for development and performance management. Performance is often measured by observing the specific behaviours a person displays on an assigned task and then rating them against specific key performance indicators or competencies to identify how appropriate or effective a particular person is at a particular job. Organisations have been using methods such as performance appraisals, reference checking, peer reviews, and other less sophisticated and equally subjective methods to identify an individual's level of performance. However, a proper evaluation of an individual's personal attributes and temperament will also help predict behaviour in any given environment and the validity of personality attributes for predicting job performance is well supported (Bartram, 2004; Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991; Humm & Wadsworth, 1943). Thus, the current investigation utilises a temperament measure that has been used for this purpose.

The Humm Wadsworth Temperament Scale

The Humm Wadsworth Temperament Scale (Humm), developed by Humm and Wadsworth in 1935, was chosen in the present investigation because it has been used for a substantial period of time for the purpose of temperament assessment, particularly with relation to predicting the future on the job performance of an individual. Despite being utilised for over fifty years, there is little established statistical information on the measure, particularly in terms of its application in recent years. Furthermore, the Humm is used in a

commercial environment where factors such as the burden of time required to complete the questionnaire can be a factor. Added to this, the items of the measure have not had a major revision since the measures conception in 1935. Thus given the lack of recent revision information available, the length of the current measure and the time it takes to complete the questionnaire, the Humm was seen as a measure worthy of further investigation.

The Humm is distinct from other measures in that an individual's responses do not generate a specific score or set of scores that fit into a predetermined number of temperament descriptors. In contrast, the Humm does not have a predetermined number of descriptors and as such the Humm has many more possible combinations and various strengths of temperament characteristics when compared with other temperament measures. One of the measure's great assets and point of difference is the ability to measure such a wide range of unique temperament characteristics. The Humm purports to measure seven components and 31 sub-components of temperament. These components and sub-components are rated on a nine point scale (a score of one indicating little or no presence of that component or subcomponent, a score of nine indicating a very strong presence of that component or subcomponent). This gives immense subtlety to the data that the interpreting psychologist has to work with and the possible combinations or styles and the number of variations is too large to contemplate. In this sense, the uniqueness of an individual can be appreciated. The Humm sheds significant light on the strengths and weaknesses (or development needs) of an individual's temperament. This includes, but is not limited to: what motivates an individual; what their stressors and stress reactions may be; how they approach their work; how they are likely to interact with others; how they can best be managed; and how they are likely to manage others (Humm & Wadsworth, 1935; 1941).

A human capital solutions company (the sole proprietor of the Humm), currently utilises the Humm to assess an individual's temperament for the purpose of recruitment and selection, promotion and development, or career guidance, and in the past the Humm has been proved reliable and valid[1]. However, it has been identified that there is some potential for the measure to be revised, improved and updated. Literature on the Humm is quite dated and predominantly ranges from the 1930s to 1950s. The company's founders gained worldwide rights to the instrument in the late 1950s and have successfully used it since as a good predictor of an individual's behaviour.

It is important to have an appreciation of how the current version of the Humm was constructed. The temperament characteristics under consideration first became of interest when observing people with psychological disorders. Around the turn of the century, a European American psychiatrist, Aaron Rosanoff (1927) developed a particularly useful way of looking at human temperament. Rosanoff was interested in clinical or abnormal behaviour, his perspective being that abnormal behaviour is only behaviour exhibited to a degree that impedes proper functioning in a given situation. Rosanoff's belief was that abnormal behaviour is driven by the same core components all humans possess, but to excessive levels and without any control. Rosanoff identified core characteristics and once he could measure them, he had a means by which to understand abnormal behaviour. Many years later an American industrial psychologist, Doncaster Humm identified the theoretical framework as being highly useful for looking at functional people in the workplace with the

[1] Investigations into the reliability of the scale found a mean test-retest reliability of .86, an internal consistency of .83, and concurrent validity studies yielded coefficients of .94 and .98 (Kruger, 1938; Humm & Humm, 1944; Smith, Gudmand & Marke, 1958).

view to improve selection and career planning. He identified that there was value in temperament analysis, and in partnership with Guy Wadsworth, a statistician, the Humm Wadsworth Temperament Scale was developed (Humm & Wadsworth, 1935). Humm and Wadsworth applied Rosanoff's theory to a functional 'normal' population, which enabled greater capability in predicting behaviour.

The original standardisation of the Humm consisted of preparing a questionnaire and selecting subjects who displayed known temperament characteristics to whom the questionnaire was administered. Then their responses were analysed to determine the value of each of the items in the questionnaire and the future significance of scores when the measure is administered to unknown subjects at a later date. When the original items were selected for the questionnaire, approximately 2,000 questions were compiled which appeared to have relevance to the traits of the Humm's seven components of temperament: Normal, Hysteroid, Cycloid Manic, Cycloid Depressive, Schizoid Autistic, Schizoid Paranoid and Epileptoid (Humm & Wadsworth, 1935). These components are now known as Normal, Hustler, Mover, Double-Checker, Artist, Politician, and Engineer for obvious commercial and politically sensitive reasons. Once this large number of questions had been compiled, Humm and Wadsworth met with Rosanoff and selected 221 items that seemed most likely to measure the above seven components of temperament. These items were tested on experimental groups and were found to give reasonable results, although not as good as would be necessary if the measure was to be valuable in appraising prospective employees. Subsequently, the items were all reconsidered. Those items which had proved useless in the first trial standardisation were discarded and enough additional items were

included to make a total of 318 questions which constitute the present form of the Humm Wadsworth Temperament Scale[2].

Humm and Wadsworth (1935) established the original norms from a sample taken from the general community. In later developments, further samples were taken from individuals in employment to develop norms for the industrial community. General population norms were established in 1950 and industrial norms in 1955. The owner of the measure revised the norms for the Australian population in 1966, again in 1977 and in 1999.

Several basic assumptions were necessary when the Humm was created. Firstly, that an individual answering selected questions was, in itself, a sampling of behaviour by which temperamental tendencies could be observed. The reasoning for this being that people of similar tendencies would answer the questions in a similar fashion, while differences in temperamental tendencies would lead to differing responses. The second assumption was that those individuals possessing the temperament characteristics to be measured could be recognised by some other method independent of the questionnaire in order to provide criterion groups for testing the questions, and for the standardisation of the measure. In this case, an alternative method such as behavioural observation could be employed due to the fact that temperament traits can be exhibited quite overtly and predominant components can usually be identified through observation (Humm & Wadsworth, 1935).

[2] It is noted that experimental and control groups were used, item analysis was conducted and raw-score norms were established due to the components of temperament not mapping directly on to a normal distribution curve. However, the details of these are beyond the scope of the current study and are therefore not reported on further.

In judging the usefulness of a measure, it is important to know whether or not it is a stable and consistent measure of the variables it is designed to investigate. Previous studies using the Humm have determined the reliability and validity of the current measure (Kruger, 1938; Humm & Humm, 1944; Smith, Gudmand & Marke, 1958). Dysinger (1939) yielded test-retest reliability scores for the seven components significant at the 0.1 % level, with a mean r of .847, indicating that if a person was to answer the questionnaire a second time, their result would typically be the same as their first results. The effectiveness of the Humm as one of the procedures to be used in personnel appraisals has also been reported favourably by a number of its users. Humm and Humm (1944) confirmed this with research involving one of their clients that yielded a correlation of +0.72 between test results and the ability to predict future performance on the job. A private follow-up study in 1974 replicated these results[3]. In 1997 the owner of the Humm surveyed 56 of their clients regarding the accuracy of the information provided to the client about their employees (based on the employees' results derived from the Humm). There were 225 appraisals in total and the clients surveyed were in a position to observe their employees' behaviour over time. Clients rated the accuracy of the information provided on an employee using a Likert Scale from one to five, one being the information provided was inaccurate with the observed behaviour, and five being the information provided was consistently accurate with the observed behaviour. Of the 56 clients surveyed, 91 percent rated the information provided as consistently accurate with the observed behaviour of their employees. These studies indicate that the Humm was constructed with considerable care to ensure that the measure contains a representative sample of items relating to temperament characteristics

[3] For commercial in confidence and privacy reasons further details of these clients and studies cannot be published.

of all types, and that the measure has a definite theoretical basis that can be used effectively by trained psychologists who have a sound understanding of this theory.

Additionally, the Humm provides a gauge of response bias for the total measure and for each of the seven components within the Humm (Humm & Humm, 1944). These 'fake' measures (sometimes referred to as social desirability or impression management), indicate the degree of defensiveness or suggestibility with which the individual has responded to the questionnaire (Dicken, 1963; McCrae & Costa, 1983; Salgado, 1999; and McCrae, 1986). This form of response bias can contaminate the overall test scores. Therefore, among other things, the Humm provides a means for evaluating the extent to which an individual displays this and adjusts their scores accordingly. This phenomenon is supported by Dicken (1963), McCrae and Costa (1983), and McCrae (1986), who suggest that if people respond to the desirability of an item, rather than the content of the item itself, controlling for this response bias should enhance the validity of scores derived from the measure. In the original research by Rosanoff (1927) it was discovered that some of the institutional subjects under-reported their faults to the extent that their responses returned a profile similar to those of normal subjects. Similarly, some of the normal subjects over-reported their faults such that their profiles were similar to those of the institutional subjects. Investigations showed that the former invariably answered predominantly 'No' to the items in the questionnaire, while the latter answered predominantly 'Yes'. In addition, it was found that subjects whose profiles were in agreeance with their case histories tended to distribute their answers fairly evenly between 'Yes' and 'No'. Two measures of response-bias were developed to counter the effect caused by an imbalance between 'Yes' and 'No' responses. Firstly the 'No Count' or number of times an individual responds 'No', and

secondly the profile count or the amount by which the profile positions of all the components except Normal vary from the zero or 'typical' position.

Interestingly, as well as the total scores for each of the components and subcomponents of temperament that are used for the Humm's interpretation, the Humm also provides measures for the accuracy of the information obtained. Firstly, as discussed above, it provides an overall measure for an individual's responses, being the total number of times an individual responds 'No'. Values ranging from 120 to 220 out of a possible 318 responses are deemed an acceptable range for this 'No Count' measure. Secondly, the Humm provides seven individual measures, one for each of the components. These measures are referred to as corrective factors and scores ranging from .75 to 1.75 are deemed as being within an acceptable range. However, less reliable information on one component does not necessarily mean that the whole test for an individual is unusable and this decision is open to the interpretation of a qualified psychologist. When the Humm's measures of response bias are triggered and are deemed significant, meaning that the scores do not fall within the acceptable range for standard interpretation, the psychologist will often administer a second personality measure such as the NEO-PIR or the 16 Personality Factor Questionnaire (16PF) as a confirmatory measure.

The owner of the Humm has been using the measure since the company's beginning, first under licence and then as the proprietary holder when it purchased the rights to the measure on Doncaster Humm's retirement. The measure was brought to Australia from America and re-normed for the Australian general population. The Humm is purported to identify an individual's temperament characteristics and their respective levels

with accuracy, and can make population comparisons to provide useful information that can be used for predicting future performance in the workplace.

There are very few publicly available or published studies examining the Humm from an empirical standpoint, and to present knowledge this is the first study to attempt to reduce the number of items in the questionnaire without decreasing the statistical validity of the measure. Another problem identified by consumers of the Humm is the burden of time it takes an individual to complete the questionnaire due to the number of items being quite large. Therefore one of the goals of the current study was to reduce the number of questions without reducing the statistical validity of the measure. In essence, one of the aims was to make the questionnaire shorter whilst still having an appropriate and acceptable level of statistical validity.

The Humm is a psychological measurement instrument administered as a questionnaire to measure those characteristics which Humm and Wadsworth describe as making up an individual's temperament. Whilst all 318 items of the Humm address issues pertaining to work and life in general, and contribute to the overall response bias measure of the Humm, only 164 of these items load and group into the Humm's seven components and 31 subcomponents. Through an individual's responses to these questions, a profile can be generated which provides information about an individual's temperament across the components and sub-components measured. The responses to the questionnaire were once scored by hand but are now computer scored to produce an output of results that can be interpreted by a trained psychologist. The Humm has consistently maintained its interpreting integrity in that the owners of the measure do not allow other than fully

qualified and accredited psychologists to engage in its interpretation. The owners enforce a number of regulations to ensure this occurs. The owners: only permit the measure to be used for industrial purposes; confine the use of the Humm to psychologists whom the company can readily monitor, namely, psychologists trained by and employed through the owner; require their psychologists to undergo a rigorous six month training programme irrespective of their professional background; and require their psychologists to undergo monthly audits to ensure the ongoing quality of their interpretation skills.

As aforementioned, the possible number of combinations of subcomponents that can occur within the limits of the acceptable range is approximately one billion. One of the measure's great assets is the ability to account for such a wide range of unique temperaments. However, it is important to remember that the results of the Humm, as used by psychologists for the purposes of indicating how a person will perform in a variety of differing employment circumstances, are used in conjunction with other relevant information such as cognitive abilities and previous experience. Through careful interpretation of the seven components and 31 subcomponents by trained and accredited psychologists, a detailed picture of an individual's temperament is generated. The information gathered can shed light on such things as an individual's general potential, motivation, leadership style, business acumen, interpersonal style, work approach, team approach, stress tolerance, level of initiative, and self-confidence. This information is most commonly used for forecasting the future behaviours of an individual and can be utilised for, among other purposes, recruitment, career guidance, team building, career development and consideration for promotion. Currently the Humm is predominantly used as a measure

to identify an individual whose temperament is both suited to the work that is to be completed and the environment within which the individual is to operate.

The Humm's subcomponents were developed by breaking down the seven components of the Humm into a finer and more detailed analysis. This arose from the discovery that any given individual may manifest some of the tendencies associated with a given component, but not necessarily all of the component tendencies. Humm and Wadsworth's first attempt to subdivide the components resulted in 40 subcomponents being created. This was a more detailed breakdown than was justified, since some of the subcomponents were identifiable by too few items. Humm later reduced the number of subcomponents to 31, none of which had less than 12 questions attributed to it.

The current Humm distributes the 31 subcomponents of temperament across the seven components as follows (and illustrated in Figure 1 on p.18): The Normal component has four subcomponents; the Hustler component has six subcomponents; the Mover component has four subcomponents; the Double-Checker component has five subcomponents; the Artist component has five subcomponents; the Politician component has three subcomponents; and the Engineer component has four subcomponents. These subcomponents can be described in more detail, however, for the purposes of the current analysis we will focus on the seven major components only. Whilst the 31 subcomponents give immense subtlety and uniqueness to the data gathered on an individual, knowledge of the seven components of temperament alone can provide an appropriate level of information for management and human resource practices, and it is to these seven components that this report now turns.

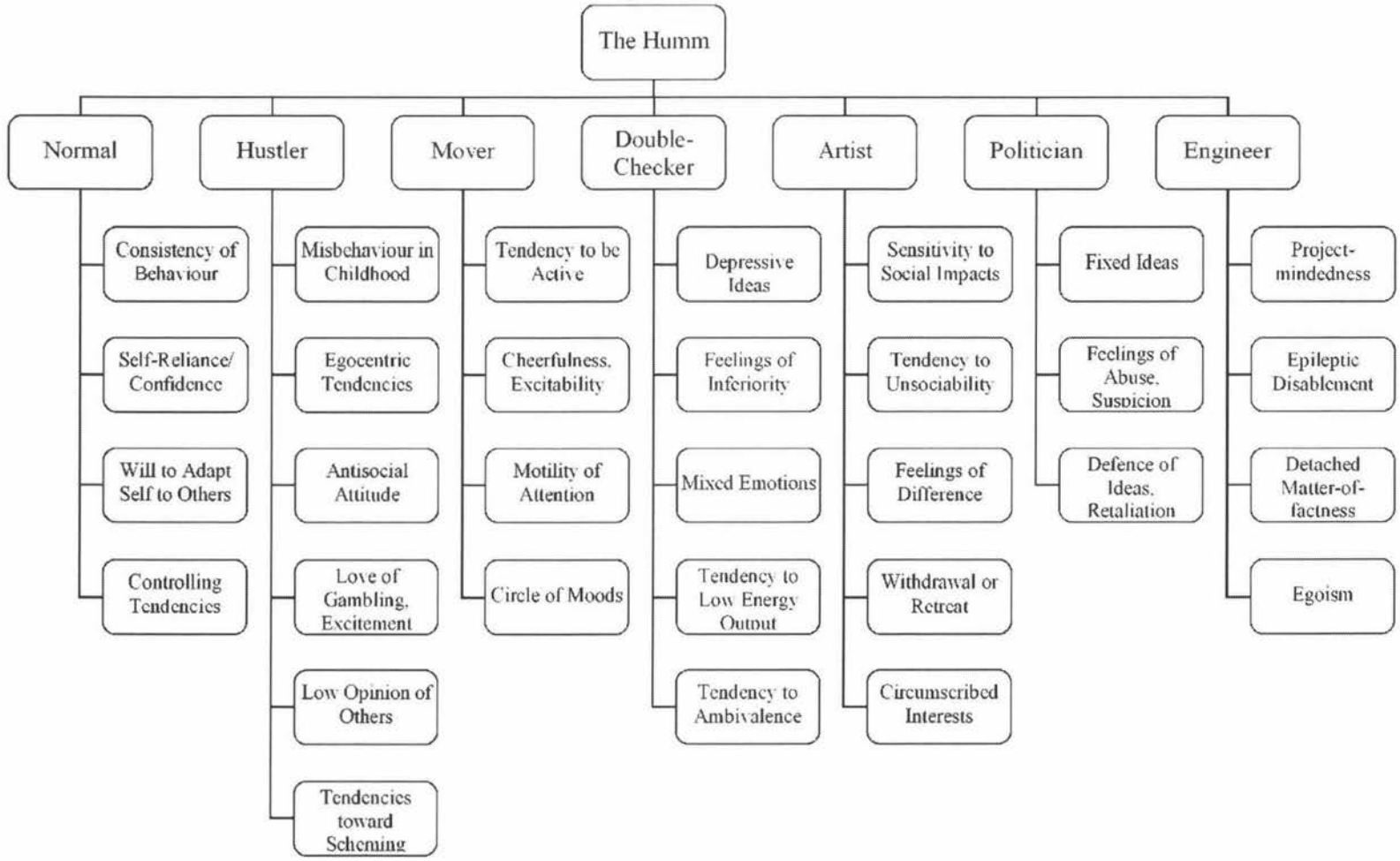


Figure 1: The Seven Components and 31 Sub-components of the Humm

The Humm's Seven Components of Temperament

The descriptions that follow are based on Rosanoff's (1927) work and the later work of Humm and Wadsworth (1935), and refer to the stereotypical behaviour associated with each of the seven (Normal, Hustler, Mover, Double-Checker, Artist, Politician and Engineer) components of temperament within the Humm.

Normal

The Normal component includes a group of traits or tendencies that provides a certain amount of control over the other six components. It is responsible for the power of self-direction and self-mastery, conservatism and the desire to conform. A person displaying a high level of the Normal component seeks self-improvement and applies a high degree of control over their emotional reactions even in stressful situations. They can also have a conservative attitude to rules and regulations, and can be highly adaptable to social or peer group expectations. This component is frequently described in terms of emphasising its inhibitory and repressive functions and importantly its directive and integrative functions. It not only prevents unfavourable manifestations of the other six components, but also enables the valuable and constructive manifestations of the other six components to present themselves. The Normal component also acts as a measure of self-mastery to evaluate the degree to which an individual's temperament characteristics integrate and is measured by the relationship between the overall score of the Normal component and the scores of the other six components. The Normal component is effective in discriminating between individuals who are masters of themselves and individuals who

give way to impulses and may behave erratically. This measure is so effective in fact, that with a considerable degree of certainty, an individual who has a high level of the Normal component is generally well integrated, while an individual who has a low level of the Normal component is likely to have some difficulty coping, particularly if faced with challenging or difficult situations.

Hustler

The Hustler component includes the group of traits that leads to a preoccupation with self-interest, the furthering of personal agendas and the satisfaction of personal desires, to the point of not considering the interests and desires of others. The Hustler component includes such attributes as the desire for financial gain, the need for excitement and short-term gratification, the possession of diplomacy, tact and persuasion skills, as well as business acumen and commercial astuteness.

Mover

The Mover component is responsible for an individual's activity, energy, motility and sociability. Many associated tendencies found in the Mover component include cheerfulness, enthusiasm, and jocularity, responsiveness to others, versatility, hopefulness, and the ability to multi-task.

Double-Checker

Closely related to the Mover component is the Double-Checker component. This component also centres on feelings, emotions and associated manifestations. However the Double-Checker component is responsible for negativity, caution, self-critical behaviour, pessimism, anxiety, empathy, and the manner in which a person makes decisions.

Artist

The Artist component is responsible for shy, sensitive, introspective behaviour. An individual displaying a great deal of the Artist component will be socially sensitive, frequently experiencing some difficulty in expressing their ideas and opinions in face-to-face situations. They are imaginative and creative people, who may be subject to reclusive reactions resulting from feelings of difference. The Artist component includes attributes such as insightfulness, self-consciousness, embarrassment and withdrawal.

Politician

Responsible for ego-defensive behaviour, assertiveness, competitiveness, stubbornness and defensiveness to criticism is the Politician component. Individuals with a high level of the Politician component may be argumentative and can display suspicious, vengeful or aggressive behaviour. They are generally ambitious individuals who are driven by the desire for status, power and prestige.

Engineer

The Engineer component is responsible for systematic, precise, matter-of-fact behaviour, as well as emphasising organisation, procedure, detail and method. Deliberate in approach, people displaying the Engineer component can be quite meticulous, task-orientated, objective, and detail-minded and gain satisfaction through accomplishments.

The Humm in relation to other Personality Measurement Instruments

How does the Humm framework compare with more widely used and published personality measurement instruments? One of the most commonly used sets of traits today is that of the Big Five (Costa & McCrae, 1992). The dimensions of the Big Five are as follows: Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness. Whilst these five factors do not directly align with the seven components of the Humm, some similarities are evident. Neuroticism includes such characteristics as worrying and nervousness and seems most closely aligned to the Double-Checker component of the Humm. The Extraversion trait includes characteristics such as being person-oriented and talkative which seems strongly related to the Mover component. Openness relates to creative and imaginative characteristics, which appear similar to the Artist component. Agreeableness refers to whether a person is cynical, suspicious, vengeful or manipulative and these characteristics seem to align with both the Hustler and the Politician components of the Humm. Finally, Conscientiousness includes characteristics such as being organised, self-disciplined, ambitious and hard working and these appear similar to the Normal and Engineer components of the Humm. The questionnaire that Costa and McCrae (1992)

developed through factor analyses of personality ratings that incorporated the Big Five factors is called the NEO-Personality Inventory Revised (NEO-PI-R). Each of the Big Five factors is further broken down into six facets (as illustrated in Figure 2 on p.24), and eight questionnaire items measure each of these six facets, equating to a total of 240 items in the questionnaire. This is a similar concept to the Humm questionnaire, although subjects indicate for each item the extent to which they agree or disagree, using a five-point rating scale which differs from the forced 'Yes' or 'No' choice for the Humm. Providing further support for the Big Five approach is a similar framework referred to as the Great Eight competency structure as discussed by Bartram (2005). Whilst the relationship between the Great Eight and the Big Five is not exact, the Great Eight does incorporate most of the aspects of the Big Five approach. It appears that frameworks such as the Big Five are similar in concept to that of the Humm. Whilst the components are referred to by different names and the characteristics of the components are grouped slightly differently, the majority of the Humm traits are represented in some way across the overall measure.

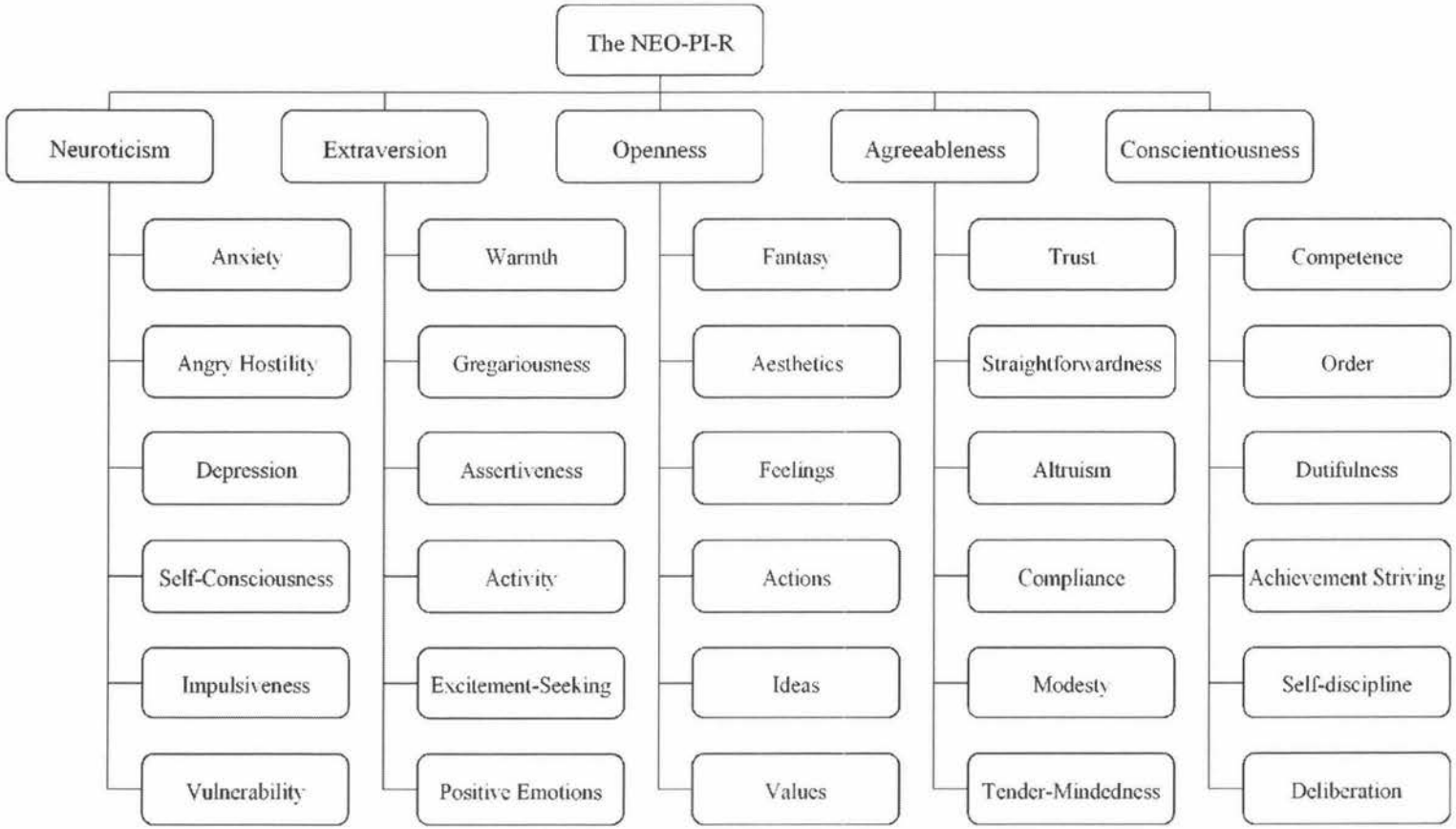


Figure 2: The Big Five Factors and 30 Facets of the NEO-PI-R

Evidence for Revision of Psychological Measurement Instruments

Butcher (2000) discussed guidelines for personality test revision, stating that many psychological measures require updating in order to ensure that their timeliness and effectiveness can be maintained. However, when revising a measure, certain aspects must be maintained to make sure that the revision exercise does not create a new measure altogether. The revised version of a measure must be similar, if not identical to that of the original measure with regards to its structure and configuration. Butcher goes on to say that the revised version of a measure must also be a distinct improvement from the original version, so that the assessment standards of the original version of the measure are raised. Furthermore, Butcher (2000) states that a revision exercise for any measure should be based on and supported by clear empirical justification and rationale, not merely pressure from market forces or other commercial interests. Commercial viability considerations such as the time taken to complete the questionnaire did, in part, drive the current investigation. However, the lack of statistical information available on a measure widely used to assess an individual's temperament, as well as the empirical vulnerability of the Humm, were the key drivers for the present research.

Butcher (2000) suggests that it is also important to gain input from a variety of sources during the revision process, thus qualified Humm experts and users were consulted. To this end, employees from the psychological services team within the owner of the Humm from were asked to contribute to the research, of which six employees responded. In addition, at the conclusion of a revision empirical evidence on the validity of the revised version of the measure is required. Hull, Lehn and Tedile (1991) state that most measures

can indeed have their goodness-of-fit statistics improved through modification post-development and this can sometimes be due to chance fluctuations in the sample data. Therefore, it is important that measures with post-development modifications are always replicated. Thus in the current study a second confirmatory factor analysis was conducted using one half of the data collected to account for this possible effect.

Measurement

It was proposed by Torgerson (1962) that in the social sciences there is a tendency to concentrate on psychological measurement instrument construction, where the means becomes the end of the measurement process. Furthermore, measurement in the discipline of psychology has always been controversial as psychology is not as tangible as other disciplines, such as physics, for example. Some common methods of measurement include: the ordinal assignment of numbers to primary physical qualities, for example, height; counting units which are of the same magnitude; or solving inequalities through ordering and cancellation. The last measurement method is worthy of consideration for applying to the field of psychology (Krantz, Luce, Suppes, & Tversky, 1971; Luce & Tukey, 1964; Torgerson, 1962). Luce (1963) states that fundamental measurement is based on additivity, going on to say that an additive psychological variable has not yet been discovered. Nevertheless, Thurstone (1959) conducted an experiment that showed that additivity of psychological values was indeed possible. However, to adopt classical measurement methods when measuring psychological attributes, additivity must be proved, not simply assumed. Assigning numbers according to rules does not automatically denote that the entity has been measured; rather it only indicates that the entity under investigation has

been classified (Grimm & Yarnold, 1995; Stevens, 1946; Torgerson, 1962). Consequently it is essential to establish the measurement properties of a temperament measure in order to ensure the legitimate and appropriate application of the analyses. Borkenau (2001) adds that there is a distinct difference between measuring a person's abilities, and measuring a person's personality. Measuring abilities involves sampling a person's relevant behaviour, whereas measuring personality is generally based on questionnaires and self-rating scales. Personality measurement most commonly relies on judgmental instruments, that is, the person being assessed makes judgments about their own personality and their typical behaviour (Borkenau, 2001).

To date, the common methods of investigating a psychological construct (including personality and temperament), have taken three forms: The first is that of the total score approach (the summing of subcomponents that have equal weighting); Secondly is the individual score approach (where results for each subcomponent are reported on); and thirdly is the regression approach (which uses multiple regression analysis simultaneously on each of the subcomponents) (Hull, Lehn & Tedile, 1991).

There are advantages and disadvantages for each of the above three methods. The total score approach may also include subcomponents that are weak or useless, which can lower the overall effectiveness of a construct. Whilst the individual score approach overcomes this problem, it also introduces ambiguity and complexity due to the analyses of multiple subcomponents. The regression approach also identifies the unique effects of each of the subcomponents. However, this approach can be limited by multicollinearity (where strong relationships exist between subcomponents) and this may cause estimated regression

coefficients to become unstable. The regression approach can also suffer from differential reliability problems whereby less important subcomponents may be measured as being more reliable than more important subcomponents that have been poorly measured (Hull, Lehn & Tedile, 1991).

Thurstone's Method of Paired Comparisons

A promising method for applying psychophysical measurement theory to psychology rests with Thurstone's method of paired comparisons (Coombs, Dawes & Tversky, 1970; Mosteller, 1963; Thurstone, 1927b, 1927c, 1931; Torgerson, 1962). Thurstone's original experiment, based on the seriousness of different crimes, provided a list of paired offences to university students and asked them to rate which was the most serious offence of each pair. The responses enabled the construction of a frequency matrix containing the relative frequency of the preferred choice in each pair.

The method of paired comparisons assumes that a stimulus arouses a discrimination process in an individual that creates a value that can then be placed along a psychological continuum (Edwards, 1957; Michell, 1990; Thurstone, 1927a). Different individuals vary in their opinion, and therefore their point of discrimination on a particular stimulus along a psychological continuum. However, the responses of a number of individuals will converge on a normal distribution as the number of participants increases, as predicted by the central limit theorem (Howell, 1997).

Thurstone (1927a) asserted that a single question alone cannot provide sufficient information for the construction of a psychological measurement instrument. In contrast, a succession of paired questions measuring the psychological distance between pairs of stimuli can provide data that can be mapped on a psychological continuum, thus enabling psychological values to be calculated for each stimulus. The proportion of respondents preferring one or the other paired stimuli from the frequency matrix is then used to create a proportion matrix. The proportion matrix is converted to a standardised matrix, providing a consensus set of psychological values for the stimuli under consideration.

While Thurstone's method is straightforward and succinct, unidimensionality and quantitativity of the psychological attribute is assumed, but not established (Michell, 1990). Luce (1963) argues that Thurstone's equal variance assumption has not yet been proved. However, Lord and Novick (1968) suggest that Thurstone's method can lead to an interval scale, which is enough for the current purposes. Although Thurstone's assumption of equal standard deviations of stimuli is sometimes challenged, Moesteller (1963) argues that Thurstone's assumption is reasonable and any arguments to the contrary are not enough to discourage further use of this method.

In terms of the present investigation, by conducting a thorough conceptual analysis of the items in the Humm, the aim of developing a better selection of items for each of the seven temperament components can then be addressed. The application of Thurstone's methods providing psychological values for items can then be applied to potentially achieve increased reliability of the revised version of the measure (Thurstone, 1927b). The application of Thurstone's psychometric values to the evaluation of items in the Humm can

provide a quantitative foundation for the revision of the measure. Establishing the quantitative attributes of the stimuli plausibly places the Humm construct on a more sound theoretical foundation than is presently the case. Thurstone's method of paired comparisons appears to be a promising tool for evaluating some of the psychological attributes underlying the Humm. The current study aims to utilise this method to provide psychological values that can be applied to ascertain a more valid set of items, and therefore create an empirically stronger psychological measurement instrument.

Aims and Rationale of the Current Study

In sum, the present study addresses three main issues, which to present knowledge no previously published study has attempted to investigate. The first is the evaluation of the psychometric properties of the Humm Wadsworth Temperament Scale (Humm), with the view of reducing the number of items in the measure without reducing the empirical validity of the measure. The second is the application of quantitative theory, in this case Thurstone's method of paired comparisons, to evaluate the items in the Humm conceptually and to possibly improve construct validity by choosing appropriate items as required by the psychological domain under investigation (Anastasi & Urbina, 1997). The final issue is to utilise confirmatory factor analysis to refine the items identified conceptually in order to create the most statistically appropriate set of items possible in the present circumstances. The above aspects of this study, based on proven psychometric methods, can then be integrated into future revisions of the Humm (and similar temperament measures), and provide a solid foundation for future enhancements.

METHOD

In the present investigation there were five tasks to be undertaken. The first task at hand was to conduct confirmatory factor analysis and structural equation modelling on the current Humm model for baseline comparisons; the second task was to conduct Thurstone's method of paired comparisons on the single loading items of the Humm; the third task was to gain conceptual judgments from Humm experts and users on where the current multi-loading Humm items fit best in order to achieve unidimensionality across items; the fourth task was to compile this information to create a common set of items for each of the seven components which could later be used to create one large seven factor model; finally confirmatory factor analysis and structural equation modelling was conducted to gain statistical and empirical information on the effectiveness of the second and third stages, and compare the revised models with the original version of the Humm.

When revising the Humm's current set of items two key methods were employed in order to make item selection decisions. First, the single loading items for each of the seven components were isolated and a panel of Humm experts and users employed Thurstone's method of paired comparisons to rank order each set of items against themselves and conceptually identify the best items for each component. Second, of the remaining items that loaded on more than one component, each expert reassigned these items to the component they believed the item to conceptually measure best. Items with an agreement between experts of 67 percent or more (as there were a total of six experts, agreement between four or more was required for a majority) were identified and retained for further consideration and analysis. Once these conceptual decisions had been made, the

components (using the newly assigned set of items) were subjected to confirmatory factor analyses (CFA) and structural equation modelling (SEM).

Maintaining the current model of the Humm, which has seven components, was attempted, however, it was necessary to create unidimensionality through eliminating the multi-loading items currently used by either reassigning them to one component only, or by removing these items completely. Data on goodness-of-fit indices as well as the magnitude of item loadings were examined for each of the seven individual models, and for the large seven- component model.

Participants

Twenty seven thousand, two hundred and forty five participants (17,372 male, 7310 female, 2563 unspecified) from a wide range of vocational disciplines, completed the Humm as part of either a recruitment and selection process initiated by a prospective employer, promotion and development assessment initiated by their current employer, or career guidance advice of their own volition. The participants' ages ranged from 15 to 71 years of age with a mean of 34.2 years (SD=9.62). The ethnicity of participants was not recorded, however, 26,787 completed the questionnaire in Australia and 508 completed the questionnaire in New Zealand. The participants' level of education as well as the length of time the participant had been speaking English (see Table 1 on p.33) was recorded.

Table 1: Participants level of English at time of completing questionnaire (n=27,245)

Time speaking English	# of participants	% of participants
Since birth	21,064	77.3
< 1 year	33	0.1
1 – 2 years	70	0.3
3 – 5 years	279	1.0
6 – 10 years	482	1.8
10 + years	2754	10.1
Unspecified	2563	9.4

As part of Thurstone's method of paired comparisons, and for the conceptual decisions made for multi-loading items, a panel of six experts or users of the Humm were consulted. This expert panel, consisting of employees from the measure's owner's included an associate psychologist who had less than 12 months exposure to the Humm, an experienced test-developer within the Research and Development team with less than 12 months exposure to the Humm, a senior consultant psychologist, the current principal psychologist and trainer of the Humm, a former principal psychologist who currently acts as a consulting panel psychologist for the owner of the measure (both with more than 35 years experience with the Humm), and a psychological services manager and former senior consultant psychologist.

Apparatus

The Humm questionnaire itself consists of 318 questions (some examples of the questionnaire items for each of the components are provided in Table 2 on p.34) to which a participant answers either 'Yes' or 'No'. The participants completed the questionnaire either on a personal computer, or by hand with pencil and paper (the results of which were

then entered on to the computer system for scoring). In most instances a questionnaire gathering biographical information was also administered, as well as a combination of cognitive ability assessments. There are no predetermined right or wrong responses for each of the Humm items, rather a participant's response of 'Yes' will either load onto the scoring system of a subcomponent or it will not[4].

Table 2: A sample of Humm questionnaire items for each component

Component	**Item Description**
Normal	Have you gone through life without nervous upsets? Has more than one person called you a hot-head?
Hustler	Do you think nearly everyone would tell a lie to keep out of trouble? When you are cornered, do you tell that portion of the truth which will do you no harm?
Mover	Are there times when you feel especially alert, and can make up your mind much more readily than at others? Do you sometimes get so excited that you find it hard to get to sleep?
Double-Checker	Have you several times had a change of heart about your life work? Do you find yourself at times very cheerful, and at others very "blue"?
Artist	Are you easily embarrassed? Do you ever have to fight against bashfulness?
Politician	Do we all demand more respect for our own rights than we five those of other people? Do you try to figure out the reason another person may have for doing something nice to you?
Engineer	Would you prefer a line of work requiring much attention to detail, to one which involves a number of different activities? When you have undertaken a task, do you find it hard to set it aside even for a short time?

Of the 318 questions in the Humm questionnaire, 164 items load into the scoring system, and 154 are superfluous items. Of the 164 loading items, 90 load on to one subcomponent only, and 74 cross load on to up to three different subcomponents. Of the

[4] Before data analysis could occur it was necessary to reverse-code 26 negatively geared items that loaded on to the scoring system when the answer was 'No' rather than 'Yes'.

single-loading items, eight items load on the Normal component; 29 items load on the Hustler component; six items load on the Mover component; eight items load on the Double-Checker component; 20 items load on the Artist component; eight items load on the Politician component; and 11 items load on the Engineer component.

There were six single-loading items from the Hustler component that were excluded from the analysis completely, reducing the total number of single loading items to 84, 23 of these loading on the Hustler component. These excluded items were items that were generally not reported on or used by psychologists today, and related to temperament characteristics that were conceptually in complete isolation and independent of any other characteristics, components or subcomponents.

Biographical Data Task

The biographical data questionnaire consisted of questions assessing a participant's age, gender, time spent speaking English, level of education, industry and level of occupation (where applicable). These questions were useful to capture salient group differences that may be of interest to construct validation of the Humm. However, in this instance, the results are beyond the scope of the current investigation, although they may be useful for future research and investigation.

Test Procedure

Groups of up to 10 participants at a time completed the questionnaire in a quiet computer laboratory. The questionnaire was administered untimed, however, it was planned to take participants approximately forty to fifty minutes to complete. Participants were given verbal instructions about the questionnaire according to administration guidelines in the Humm user manual. This gave participants insight in to the questionnaire that would follow.

Participants were advised that the questionnaire would look at how they see themselves as a person and provide an indication as to their behavioural characteristics, preferred work approach and general motivation, as well as their style in dealing with others. They were also advised that there were no right or wrong answers and that for a number of the questions they may find they can honestly answer both 'Yes' and 'No' in different situations. Participants were asked to respond the way they thought was most typical of their behaviour, allowing them to seemingly contradict themselves when questions are repeated with slight changes in wording or emphasis. Answering 'Yes' on one occasion and 'No' to a similar question later indicates that while they possess a certain attitude or display a particular behaviour on occasions, it will not always be the case. The participants were advised that the most appropriate approach to the questionnaire was to work quickly, giving their initial impression rather than thinking too deeply about any particular question or trying to work out how their responses might be interpreted.

A trained administrator remained nearby for the duration of the assessment to field any participant queries as they arose. An English dictionary was also made available so that participants were able to look up the meaning of any unfamiliar words contained within the questionnaire. At the completion of the questionnaire participants were given an opportunity to receive feedback on their results from a registered psychologist. All the participant responses were retrieved through a central networked personal computer. The data was then retained for statistical analysis.

Thurstone's Method of Paired Comparisons for Single-loading items

The quantitative method conducted in this experiment was Thurstone's method of paired comparisons, modelled on Thurstone's (1927b) method of paired comparisons for social values. The paired comparison test provided a method for evaluating the conceptual attributes and consistency of the items in the Humm. Each of the single loading items from each of the seven components was conceptually compared against each other by the six Humm experts and users. The rankings of each of the items by the six experts were then compared to create a combined rank order of items for each of the seven components. Once this had been completed, the items with the lowest rankings were removed from the group of items and structural equation models were created to investigate the statistical significance of the new set of items for each of the seven components.

Thurstone's technique was included with the express purpose of providing a means for investigating the properties of items within the seven components of the Humm. A limited number of items (only those that were single-loading on one of the seven

components) were evaluated using Thurstone's method of paired comparisons. This method was time consuming as it required n (n-1)/2 questions to be asked for each of the seven components. There were varying numbers of single-loading items in each of the seven components, ranging from six to 23 items, therefore the number of questions asked for each model ranged from 15 to 253 questions. In terms of inter-rater reliability, the conceptual judgments of the six Humm experts and users on the single loading items were relatively consistent, gaining percentage of agreement values that ranged from 51 to 90 percent, with an average of 71 percent.

Thurstone's method of paired comparisons provides a psychological scaling method for determining values of psychological phenomena (Edwards, 1957; Thurstone 1927b). Scale construction using Thurstone's methods provides an alternative technique that can potentially increase the validity of a measure. The Humm item weights were obtained from the conceptual judgments by utilising scoring weights to gain a consensus across the experts' judgments which in turn produced a common rank order of items for each of the seven components. This resulted in one list of items for each component that ranked from the item most likely to measure the component to the item least likely to measure the component. This could then be used to select the highest ranking items to be included in further analysis.

Expert judgments for Multi-Loading items

Conceptual decisions were made by each of the six experts as to whether the items that loaded on more than one subcomponent could be reassigned to load on only one of the

seven components. Once each of the experts had reassigned the items to the component they thought was best measured by the item, the percentage of agreement between experts was calculated for each of these items. Items that achieved 67 percent or more agreement between experts were retained for further analysis, whilst the remaining items were disregarded. In all, 75 multi-loading items were reviewed by the Humm experts. Sixty two items gained percentage of agreement between experts of 67 percent or greater. Thirteen items were discarded as they gained 50 percent or less agreement between experts. Inter-rater reliability was examined to ensure consensus between the conceptual judgments of the six Humm experts and users on the multi loading items (see Table 3 on p.39). The average percentage of agreement value for judgments made on multi loading items was 81 percent, indicating a strong degree of consensus between experts.

Table 3: Percentage of agreement between Humm experts for multi loading items

Percentage of agreement	**Number of items**
17	0
33	2
50	11
67	14
83	15
100	33
Total	**75**

For both the single and multi-loading items, some analytical freedom was necessary. Whilst the conceptual judgments of the experts were relied on to identify which items were of initial interest and in which components they belonged, ultimately the items included in the final models were identified through a combination of conceptual judgments and empirically based decisions.

Confirmatory Factor Analysis

A relatively recent approach that has been developed and addresses some of the concerns with relation to testing multifaceted personality constructs is that of structural equation modelling (SEM). SEM adopts a confirmatory approach whereby a hypothesised model can be assessed statistically in order to examine the extent to which the theoretical construct explains the observed data. When utilised to analyse theoretical constructs, two underlying assumptions are vital to the SEM approach: firstly, that the construct under investigation can be represented by a series of structural equations; and secondly that these equations can generate a model that conceptualises the construct under examination (Byrne, 2001).

In the SEM technique, subcomponents of a personality construct can be measured separately. However, SEM begins by estimating the extent to which the subcomponents correlate with each other, and assumes that they will not correlate perfectly due to measurement error and the unique traits of each subcomponent (Hull, Lehn & Tedile, 1991). SEM provides statistically significant tests for the size of each of the subcomponents' relationships in a model. SEM also provides statistics for: assessing the overall fit of a model; examining the extent to which the subcomponents indicate the latent variables; and examining the extent to which the latent variables are related to each other. Confirmatory factor analysis, represented as a measurement model within SEM, is a common statistical procedure that can be utilised (in appropriate settings) for investigating the relationships that exist between a construct's variables and subcomponents (Byrne,

2001). Confirmatory factor analysis was adopted in the current study because the theoretical structure of the Humm was, to some extent, already known.

The SEM approach retains many of the advantages of classical methods. SEM yields an estimate of the variable that is greater than simply adding its subcomponents together, it maintains the uniqueness of each of the subcomponents, and it takes into account any relationships that exist among subcomponents. However, SEM also tests the interrelation of the subcomponents, and the extent to which the subcomponents indicate a variable that is distinct from other variables. Relationships involving the latent variables are also free from possible measurement errors. SEM can simultaneously detect the specific effects of a subcomponent as being distinct from the general effects of a construct. SEM can also retain relative simplicity when compared to traditional approaches and provides several statistics that consider the overall fit of the model (Hull, Lehn & Tedile, 1991).

As mentioned above, a variety of statistics exist in SEM to assess the overall fit of a model with its variables. The chi-square goodness-of-fit test assesses the adequacy of the model in its ability to recreate the observed covariance matrix. If the predicted matrix significantly deviates from the observed matrix then the model is deemed inadequate; therefore statistically significant chi-square values result in the model being rejected (that is, when the chi-square goodness-of-fit statistic is not significant, it actually indicates that the model has good fit). However, in instances where the chi-square statistic is significant, it is widely accepted that theoretically sensible models are often rejected by the chi-square test and other fit indices should be considered (Roberts, Chernyshenko, Stark & Goldberg, 2005).

Bentler and Bonett (1980) suggested that the chi-square statistic is dependant on sample size and should not be used in isolation as it is likely to produce a significant result even in cases where there is a relatively good fit to the data. They proposed the Normal Fit Index (NFI) which is claimed to be less susceptible to sample size variation and also accounts for the variance in the observed covariance matrix that is accounted for by the theorised model. NFI values of below .9 suggest that the model could be improved (Bentler & Bonett, 1980). However, this too has been suggested to underestimate the adequacy of models; therefore Bentler (1990) proposed the Incremental Comparative Fit Index (CFI), which appeared to be less biased by sample size whilst still retaining the statistical interpretation properties of the NFI. Byrne (2001) also suggested that the Normed Fit Index (NFI) and Comparative-Fit Index (CFI) were adequate goodness-of-fit statistics. Values of .9 or above indicate sufficient fit, suggesting that the hypothesised model represents an adequate fit to the data. Furthermore, Byrne suggests the Tucker Lewis Index (TLI,) where values close to .95 are indicative of good fit.

Additionally, The Root Mean Square Error of Approximation (RMSEA) statistic takes into account error approximation in the data population. Values below .08 indicate reasonable fit, and values less than.05 indicate good fit between the hypothesised model and the observed data. However, if the RMSEA value is small, but the confidence interval (Lo90 and Hi90) is wide, it is not possible to determine accurately the degree of fit. On the other hand, small RMSEA values with very narrow confidence intervals can reflect good model fit (Byrne, 2001). It is also important to acknowledge whether a subcomponents' association with the variables, and the variables' association with each other are significant (Hull, Lehn & Tedile, 1991). The SEM approach allows for more accurate parameter

estimation; however models should not only be judged on their statistical merits, but also on their theoretical adequacy.

For the present purposes, models that have improved goodness-of-fit and maximum likelihood estimate statistics from that of the original model are sought, as well as increased maximum likelihood estimate regression weights with positive values for each of the items, which for the current purposes is considered the minimum threshold.

Analysis

The data collected was divided roughly into two: one half was used for the first confirmatory factor analysis (n = 13,622); and the other half of the data was used for a second confirmatory factor analysis (n = 13,623).

The current version of the Humm is based on a seven component, 31 sub-component model. However, the current study is limited to investigating how items fit into the seven components (as depicted in Figure 3 on p.45). Instead of conceptually allocating items under each of the 31 sub-components, efforts concentrated on allocating items under the seven major components of the measure only. Due to the intricate nature of the original version of the Humm (with many items cross loading on several different components or subcomponents), only fit indices for those models were recorded.

For each of the seven components several confirmatory factor analysis models were created for comparison. Firstly, the original component with all items including those that

loaded on to other components was created. Then a model with only the original single-loading items was prepared, followed by a model that contained only the single loading items identified by the Humm experts using Thurstone's method of paired comparisons. These three models were then compared using goodness-of-fit indices and the model that had the strongest statistical information was retained for further analysis.

The next step was to create models for each of seven components using the information gathered from the expert judgments on the multi loading items. Initially, models that only contained items with 100 percent agreement between experts were created. If the number of items was too small, then items that had 83 percent agreement were added to the model, followed by items that had 67 percent agreement. In the event that there were too few items identified in the 100 percent agreement set, 83 percent agreement items, followed by 67 percent agreement items were added until there were enough items for a robust model. These models were also compared using goodness-of-fit indices and the model that had the strongest statistical information was then combined with the strongest model from the single loading items analysis. This created a revised model for each of the seven components that could then be used to create a revised seven factor model. Following this, the maximum likelihood estimates for each of the items within the models were taken into consideration and any estimate with a negative value was removed. What remained for each of the seven components was a core set of items, all of which only loaded on one of the seven components, with sound goodness-of-fit statistics and maximum likelihood estimate regression weight values for each of the items that were significantly higher than those of the original set of items for each of the seven components.

Finally, once the development of each of the new seven individual models was complete, a new seven-factor model was created (using each of the empirically superior single factor models). This new large model was then compared with that of the original Humm (which was recreated for confirmatory factor analysis as per the current version of the measure).

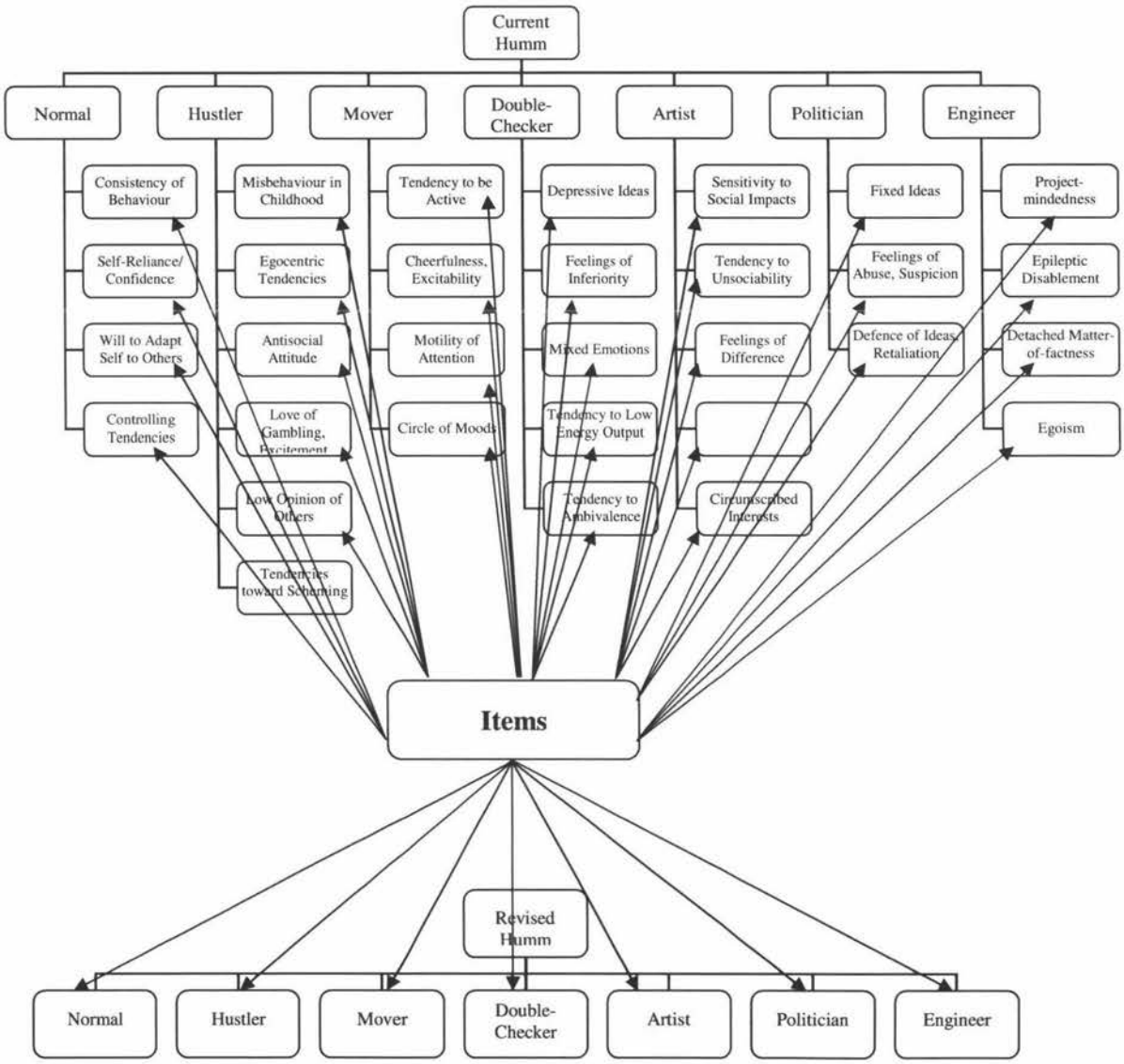


Figure 3: Allocation of Items across the Current and Revised versions of the Humm

RESULTS

The seven individual models for components Normal, Hustler, Mover, Double-Checker, Artist, Politician, and Engineer are referred to as "single factor models", whereas the large model, which incorporates all of the single factor models, is referred to as the "seven-factor model".

Analysis of the Current Humm

Confirmatory factor analysis was conducted and goodness-of-fit statistics were calculated (n = 13,622) for each of the single factor models (as reported in Table 4 on p.49) and the original seven-factor model to determine the statistical properties of the current version of the Humm. Accordingly, chi-square values with associated degrees of freedom and p values were calculated. However, as previously discussed, the chi-square statistic can be problematic (Bentler & Bonett, 1980; Byrne, 2001; Roberts, Chernyshenko, Stark & Goldberg, 2005), particularly as large sample sizes are more likely to yield significant chi-square values thus rejecting potentially worthy models (as was the case in the present investigation where all models gained a significant chi-square value). Additional goodness-of-fit statistics were calculated in order to address the chi-square limitations. As such, Tucker Lewis Index (TLI), Normed Fit Index (NFI), Comparative-Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA) were also considered and retained for comparison with later analyses.

Current Normal Single Factor Model

The results of the CFA analyses on the single factor model for the Normal component indicated goodness-of-fit indices of χ^2 = 12,910.678, df = 779, p < .0001, NFI = .718, TLI = .716, CFI = .730, and RMSEA of .034. A closer inspection of the maximum likelihood estimates showed that regression weights for each of the items in the Normal model ranged from -.484 to 0.513 and were all found to be significantly different from zero at the .001 level (two-tailed).

Current Hustler Single Factor Model

The results of the CFA analyses for the Hustler component indicated goodness-of-fit indices of χ^2 = 31,276.479, df = 945, p < .0001, NFI = .508, TLI = .492, CFI = .515, RMSEA of .049. Maximum likelihood estimates revealed that regression weights for each of the items in the Hustler model ranged from -.193 to 0.510. Regression weights were found to be significantly different from zero at the .001 level (two-tailed) with the exception of one item which was significant at the .141 level.

Current Mover Single Factor Model

The single factor model for the Mover component indicated goodness-of-fit indices of χ^2 = 7191.930, df = 324, p < .0001, NFI = .815, TLI = .807, CFI = .822, and RMSEA of .039. Regression weights ranging from -.015 to .553 were produced for each of the items in

the Mover model. Regression weights were found to be significantly different from zero at the .001 level apart from one item which gained a probability value at the .107 level.

Current Double-Checker Single Factor Model

The goodness-of-fit indices for the single factor model for the Double-Checker component were $\chi^2 = 14{,}199.095$, df = 702, $p < .0001$, NFI = .814, TLI = .811, CFI = .821, and RMSEA of .037. Regression weights for each of the items in the Double-Checker model ranged from -.454 to .522 significant at the .001 level (two-tailed).

Current Artist Single Factor Model

The current Artist component gained goodness-of-fit statistics of $\chi^2 = 14{,}682.242$, df = 702, $p < .0001$, NFI = .707, TLI = .701, CFI = .717, and RMSEA of .038. The Artist model produced significant regression weights at the .001 level for each item in ranging from -.156 to 0.563.

Current Politician Single Factor Model

The CFA analyses for the Politician component produced goodness-of-fit indices of $\chi^2 = 9037.660$, df = 377, $p < .0001$, NFI = .736, TLI = .724, CFI = .744, and RMSEA of .041. Maximum likelihood estimates showed regression weights ranging from -.214 to .572 for each of the items in the Politician model significant at the .001 level.

Current Engineer Single Factor Model

Goodness-of-fit statistics for the Engineer component single factor model were produced (χ^2 = 11,257.524, df = 560, p < .0001, NFI = .620, TLI = .608, CFI = .631, and RMSEA of .037). Regression weights for each of the items in the Engineer model ranged from -.467 to .452. Most regression weights were found to be significant at the .001 level (two-tailed), however, three items gained significance values of .005, .030 and .414.

Table 4: Goodness of Fit indices for current Humm single factor models (n = 13,622)

Model	NFI	TLI	CFI	RMSEA	Lo90	Hi90	# of Items
N (Original all items)	.718	.716	.730	.034	.033	.034	41
N (Original single items)	.855	.804	.860	.040	.037	.043	8
H (Original all items)	.508	.492	.515	.049	.048	.049	45
H (Original single items)	.607	.574	.613	.053	.052	.054	23
M (Original all items)	.815	.807	.822	.039	.039	.040	27
M (Original single items)	.925	.885	.931	.028	.023	.033	6
D (Original all items)	.814	.811	.821	.038	.037	.038	39
D (Original single items)	.881	.836	.883	.057	.053	.060	8
A (Original all items)	.707	.701	.717	.038	.038	.039	39
A (Original single items)	.813	.797	.819	.041	.040	.043	20
P (Original all items)	.736	.724	.744	.041	.040	.042	29
P (Original single items)	.644	.506	.647	.067	.064	.070	8
E (Original all items)	.620	.608	.631	.037	.037	.038	35
E (Original single items)	.711	.648	.718	.043	.041	.045	11

Current Seven-Factor Model

The results of the CFA analyses for the seven factor model indicated goodness-of-fit indices (as reported in Table 8 on p.61) of χ^2 = 142,389.365, df = 13,090, p < .0001, NFI = .573, TLI = .587, CFI = .596, and RMSEA of .027. Maximum likelihood estimates showed that regression weights for each of the items in the seven factor model were at a concerning

level, ranging from -1.372 to 2.394. The majority of regression weights were found to be significantly different from zero at the .001 level (two-tailed), however, 78 of the 258 items (remembering that this figure also includes multi-loading items) gained significance values between .002 and .940. The standardised correlations between components ranged between -.99 and .86 (as reported in Table 5 below and illustrated in Appendix I).

Table 5: Correlations between components for current Humm seven factor model (n = 13,622)

Component	N	H	M	D	A	P	E
N	-						
H	-0.85	-					
M	-0.50	0.57	-				
D	-0.72	0.60	0.62	-			
A	-0.50	0.36	0.30	0.70	-		
P	-0.99	0.82	0.47	0.66	0.45	-	
E	-0.81	0.77	0.80	0.86	0.53	0.77	-

Development of a Revised Humm

Confirmatory factor analysis (n = 13,622) was utilised to create revised models of the Humm. Applying Thurstone's method of paired comparisons and Humm experts' conceptual judgements to each of the seven components of the Humm yielded empirically superior single factor models as outlined below (and presented in Table 6 on p.59). As was the case with the models for the current single factor and seven-factor models, all chi-square statistics were found to be significant and subsequent goodness-of-fit indices were calculated. In contrast to the current Humm models, however, was the finding that all

regression weights across the all single factor and seven-factor models were significantly different from zero at the .001 level (two-tailed).

Revised Normal Single Factor Model

The goodness-of-fit statistics for the Normal component increased from those of the current Normal model (NFI = .852, TLI = .811 CFI = .858, RMSEA = .039, and χ^2 = 594.272, df = 27, $p < .0001$). Maximum likelihood estimates showed that regression weights for each of the items in the revised Normal model also increased and were at an acceptable level, ranging from .166 to .452 (as illustrated in Figure 4 below).

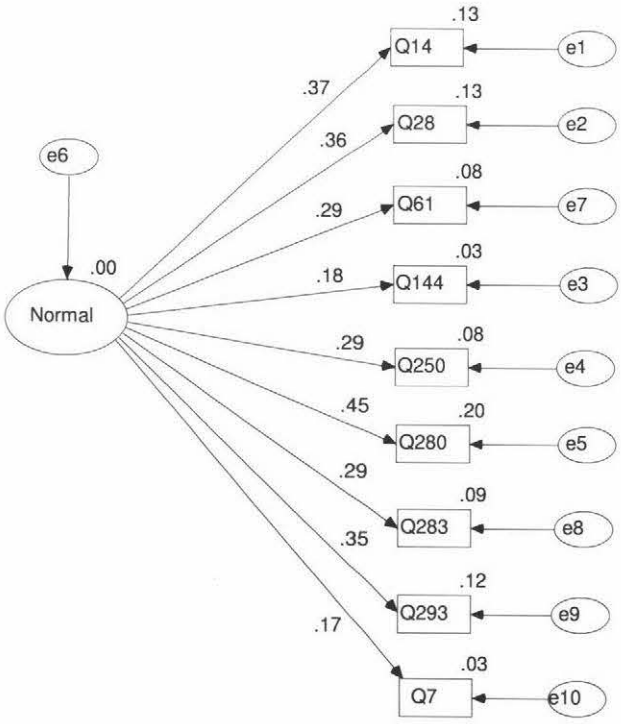


Figure 4: Standardised Regression Weights for the Revised Normal Model

Revised Hustler Single Factor Model

The Hustler component's goodness-of-fit statistics increased (NFI = .869, TLI = .852 CFI = .873, RMSEA = .042, and $\chi^2 = 2220.340$, df = 90, $p < .0001$). Regression weights for each of the items in the revised Hustler model also increased and were at an acceptable level, ranging from .259 to .557 (as illustrated in Figure 5 below).

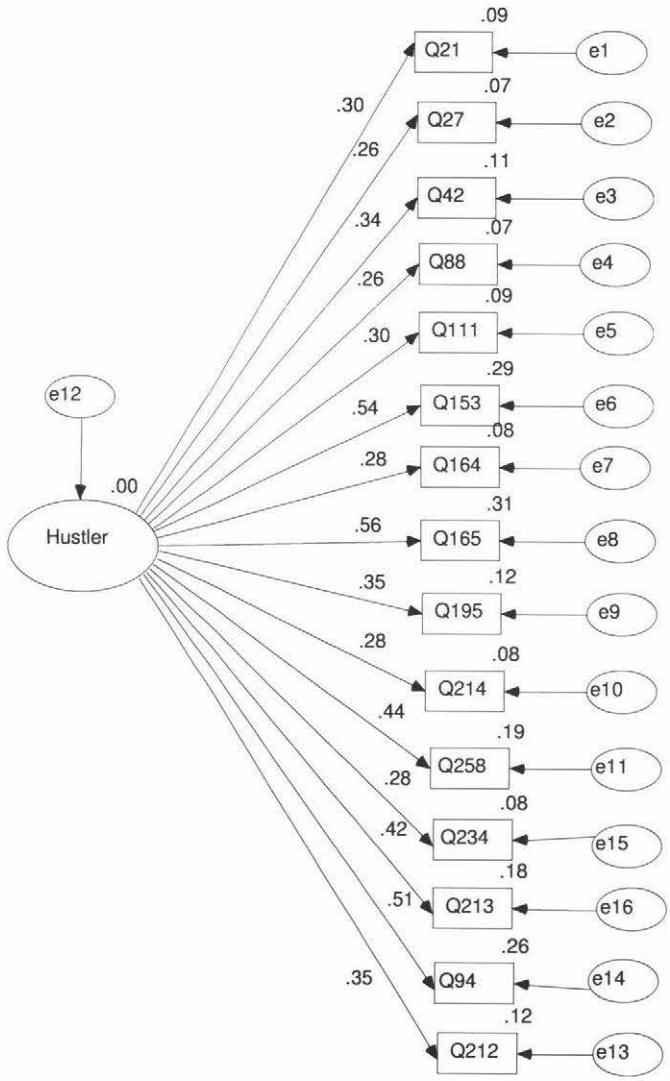


Figure 5: Standardised Regression Weights for the Revised Hustler Model

Revised Mover Single Factor Model

The goodness-of-fit statistics for the Mover component improved and provided a reasonably good fit (NFI = .922, TLI = .906 CFI = .925, RMSEA = .041, and χ^2 = 1067.416, df = 44, p < .0001). Each of the items in the revised Mover model had improved regression weights at an acceptable level, ranging from .230 to .537 (as illustrated in Figure 6 below).

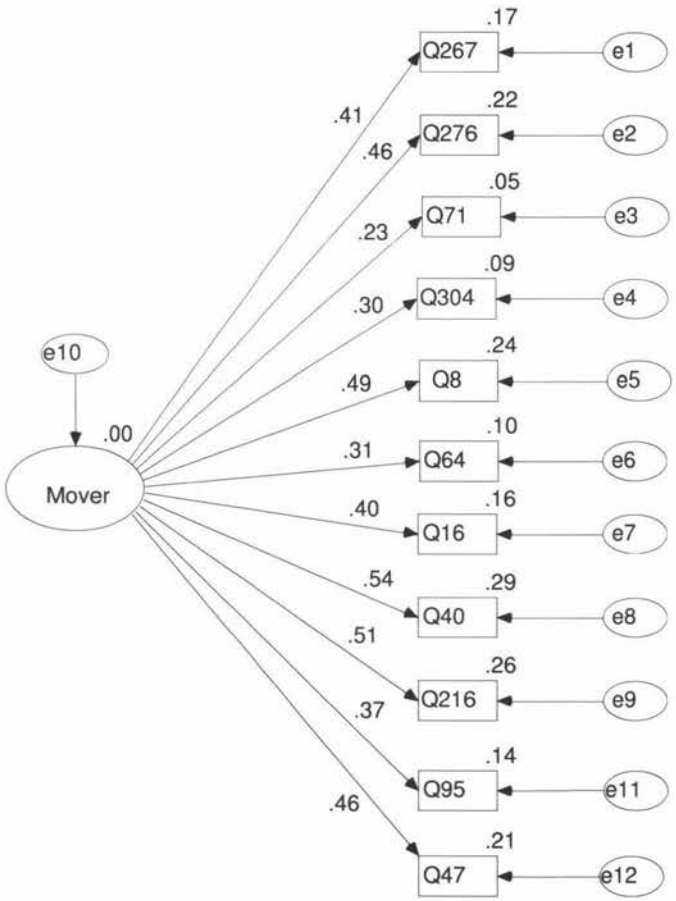


Figure 6: Standardised Regression Weights for the Revised Mover Model

Revised Double-Checker Single Factor Model

The goodness-of-fit statistics increased for the revised Double-Checker component (NFI = .945, TLI = .936 CFI = .948, RMSEA = .034, and $\chi^2 = 729.502$, df = 44, $p < .0001$). The regression weights for each of the items in the revised model also improved, ranging from .272 to .530 (as illustrated in Figure 7 below).

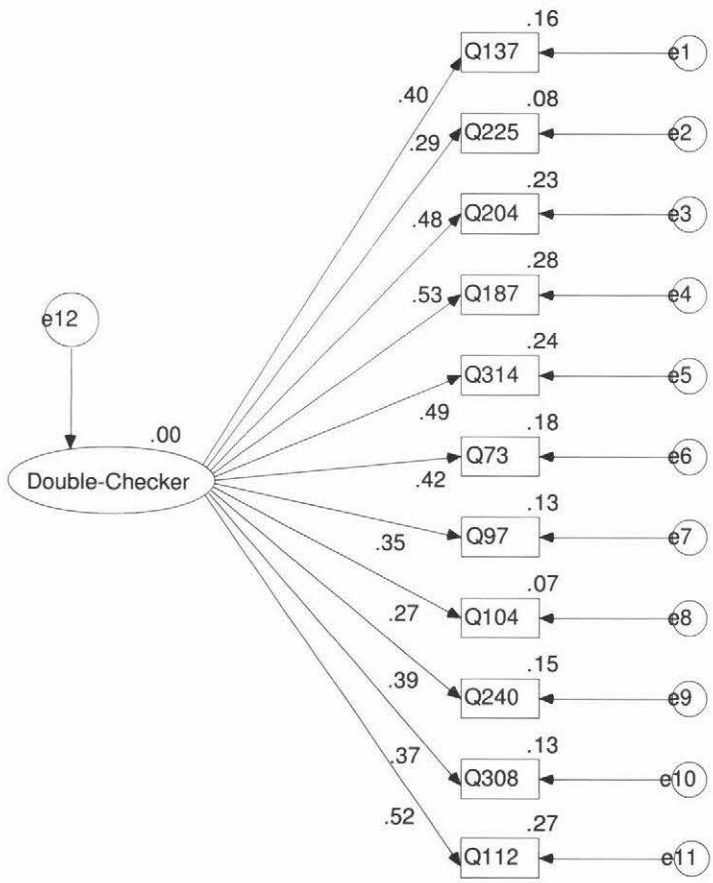


Figure 7: Standardised Regression Weights for the Revised Double-Checker Model

Revised Artist Single Factor Model

The revised Artist component's goodness-of-fit statistics increased from those of the current model (NFI = .913, TLI = .859 CFI = .914, RMSEA = .044, and χ^2 =241.880, df = 9, $p < .0001$). The revised model also had increased regression weights ranging from .161 to .459 at an acceptable level (as illustrated in Figure 8 below).

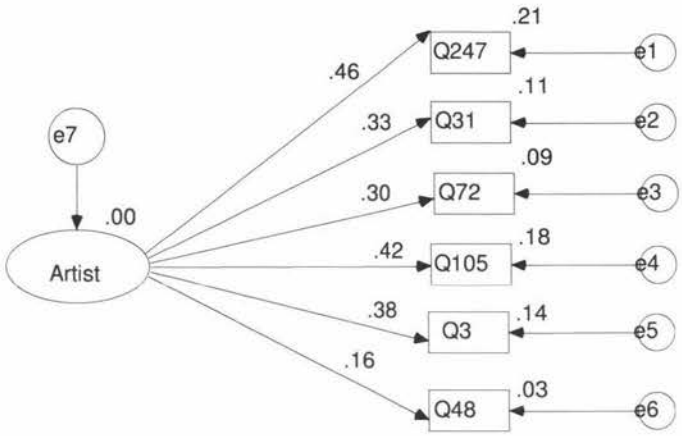


Figure 8: Standardised Regression Weights for the Revised Artist Model

Revised Politician Single Factor Model

The goodness-of-fit statistics were NFI = .889, TLI = .869 CFI = .893, RMSEA = .043, and χ^2 = 1437.884, df = 54, $p < .0001$for the Politician component which was an improvement from those of the current Politician model. The items in the revised Politician model had increased regression weights at an acceptable level that ranged from .179 to .645 (as illustrated in Figure 9 on p.56).

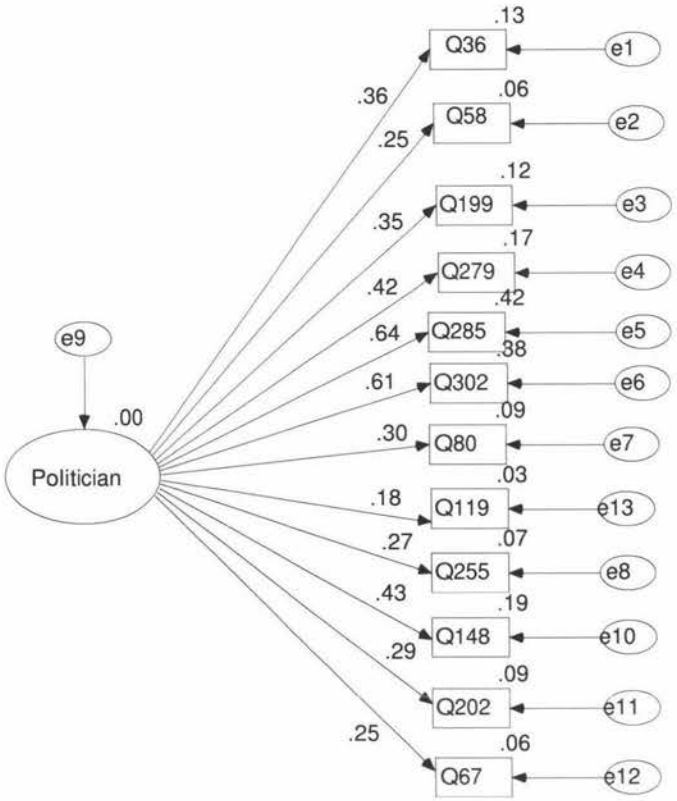


Figure 9: Standardised Regression Weights for the Revised Politician Model

Revised Engineer Single Factor Model

The Engineer component also increased in fit (NFI = .801, TLI = .730 CFI = .807, RMSEA = .041, and $\chi^2 = 484.684$, df = 20, $p < .0001$). Maximum likelihood estimates showed that regression weights for each of the items in the revised Engineer model also slightly increased. Whilst they were significant at the .001 level, they were not yet in an acceptable range, gaining scores between -.401 to .485 (as illustrated in Figure 10 on p.57). This was an unexpected finding; therefore, an alternate method was employed whereby items from the original Engineer component were selected by examining the maximum

likelihood regression weights, rather than through conceptually choosing items. Items that yielded positive scores were used to create an alternate model for the Engineer component. The goodness-of-fit statistics for the revised Engineer model where items were selected statistically increased and provided a reasonable fit (NFI = .934, TLI = .924 CFI = .941, RMSEA = .023, and χ^2 = 277.828, df = 35, p < .0001). Maximum likelihood estimates showed that regression weights for each of the items in the newly revised Engineer model also increased and were now at an acceptable level, ranging from .147 to .519 (as illustrated in Figure 11 on p.58).

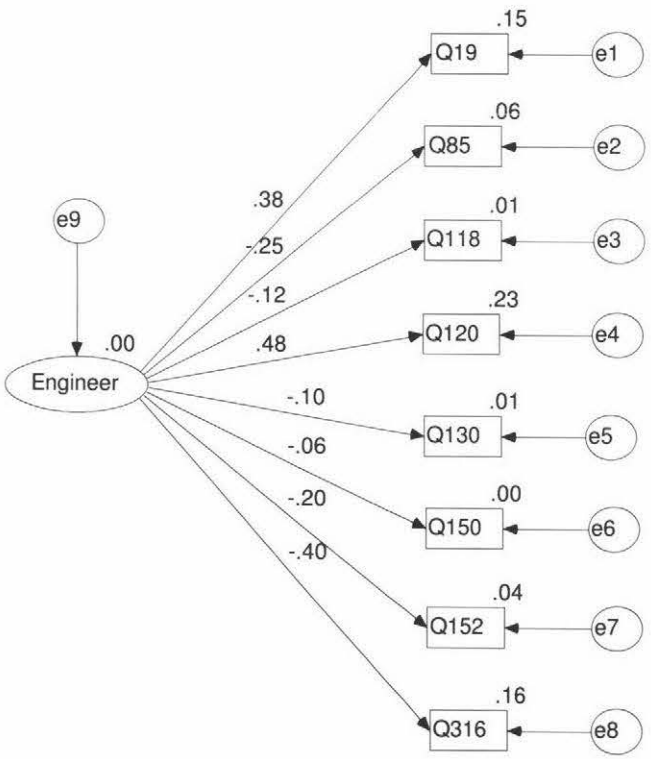


Figure 10: Standardised Regression Weights for the Revised Engineer Model where items were selected conceptually

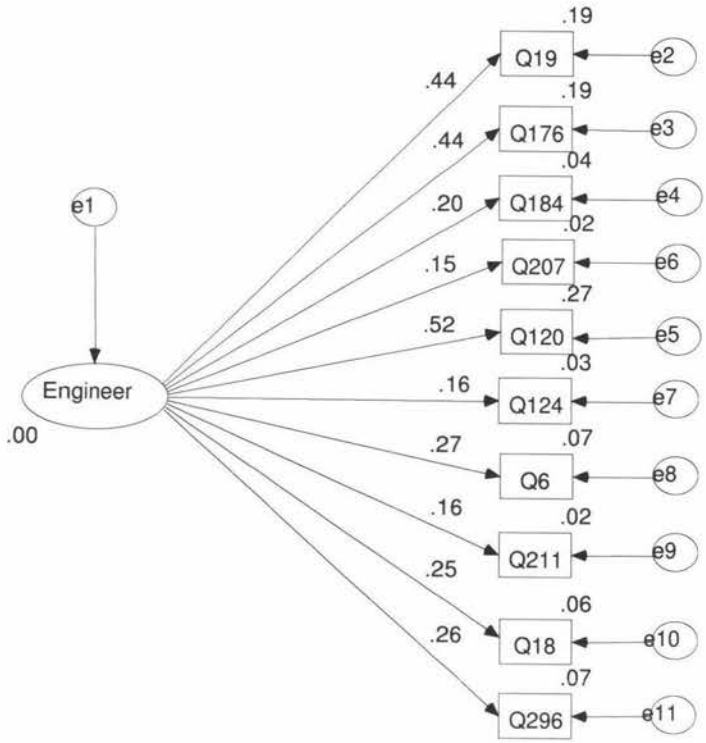


Figure11: Standardised Regression Weights for the Revised Engineer Model where items were selected statistically

Table 6: Goodness of Fit indices for revised Humm single factor models (n = 13,622)

Model	NFI	TLI	CFI	RMSEA	Lo90	Hi90	# of Items
N (Thurstone's single items)	.812	.725	.817	.047	.043	.050	7
N (Expert Judgements 100, 83 & 67 %)	.876	.769	.884	.027	.021	.034	5
N (Best of Single and Expert)	.852	.811	.858	.039	.037	.042	9
H (Thurstone's single items)	.622	.579	.626	.061	.060	.062	19
H (Expert Judgements 100, 83 & 67 %)	.813	.797	.819	.041	.040	.043	20
H (Best of Single and Expert)	.869	.852	.873	.042	.040	.043	15
M (Thurstone's single items)	.992	.992	.996	.008	.000	.016	5
M (Expert Judgements 100, 83 & 67 %)	.909	.890	.912	.046	.043	.048	11
M (Best of Single and Expert)	.922	.906	.925	.041	.039	.043	11
D (Thurstone's single items)	.883	.807	.884	.076	.072	.082	6
D (Expert Judgements 100 %)	.932	.913	.935	.040	.038	.043	9
D (Expert Judgements 100, 83 & 67 %)	.904	.896	.908	.038	.037	.040	19
D (Best of Single and Expert)	.945	.936	.948	.034	.032	.036	11
A (Thurstone's single items)	.833	.815	.838	.045	.044	.047	17
A (Expert Judgements 100 & 83 %)	.826	.782	.830	.045	.042	.047	10
A (Best of Single and Expert)	.913	.859	.914	.044	.039	.048	6
P (Thurstone's single items)	.822	.707	.824	.062	.057	.067	6
P (Expert Judgements 100%)	.844	.803	.847	.057	.055	.060	10
P (Expert Judgements 100, 83 & 67 %)	.827	.796	.830	.053	.052	.055	13
P (Best of Single and Expert)	.889	.869	.893	.043	.041	.045	12
E (Thurstone's single items)	.801	.730	.807	.041	.038	.045	8
E (Expert Judgements 100, 83 & 67 %)	.939	.885	.943	.031	.025	.038	5
E (Best of Single and Expert)	.801	.730	.807	.041	.038	.045	8
E (With statistically selected items)	.934	.924	.941	.023	.020	.025	10

Revised Seven-Factor Model

The results of the CFA analyses indicated that the overall goodness-of-fit indices for the current version of the Humm were lower than the revised version (as reported in Table 8 on p.61), so too were the goodness-of-fit indices for all of the single factor models. However, when the revised single factor models were combined to create a revised seven-factor model, the Engineer component gained some negative maximum likelihood estimates. The goodness-of-fit statistics for the revised seven-factor model, which included

the Engineer model that was revised conceptually, did increase from those of the current version of the Humm (χ^2 = 37,972.957, df = 2464, p < .0001, NFI = .708, TLI = .711, CFI = .722, and RMSEA of .033). However, on closer inspection, regression weights for each of the items in the revised seven-factor model were not at an acceptable level, ranging from -.452 to .584. As negative factor loadings are not satisfactory and it appeared that the component responsible for this was Engineer, CFA analyses were conducted on the revised seven-factor model a second time. In this second analysis, the conceptually revised Engineer model was replaced with the statistically revised model of Engineer. The goodness-of-fit indices did not alter greatly (χ^2 = 38,958.200, df = 2607, p < .0001, NFI = .710, TLI = .713, CFI = .723, and RMSEA of .032). However, maximum likelihood estimates showed that regression weights for each of the items in the revised model did increase and were now within an acceptable range, between .148 and .583. The standardised correlations between components for the seven-factor model (including statistically revised Engineer) ranged between -.92 and .96 (as reported in Table 7 below and illustrated in Appendix II). This was a slight improvement on the correlation values for the current seven-factor model.

Table 7: Correlations between components for revised Humm seven factor model (with statistically selected E) (n = 13,622)

Component	N	H	M	D	A	P	E
N	-						
H	-0.79	-					
M	-0.74	0.57	-				
D	-0.92	0.61	0.94	-			
A	-0.67	0.59	0.70	0.79	-		
P	-0.90	0.86	0.58	0.61	0.59	-	
E	-0.92	0.70	0.86	0.96	0.71	0.78	-

As a matter of interest, the Engineer component was removed from a subsequent analysis and CFA was conducted on a revised six-factor model. The goodness-of-fit indices (as presented in Table 8 below) did not vary greatly (χ^2 = 32,703.266, df = 2001, p < .0001, NFI = .720, TLI = .721, CFI = .732, and RMSEA of .034). Maximum likelihood estimates showed that regression weights for each of the items in the revised six factor model did not change greatly either, ranging from .089 to .586.

Table 8: Goodness of Fit indices for current and revised Humm seven factor models (n = 13, 622)

Model	NFI	TLI	CFI*	RMSEA	Lo90	Hi90	# of Items
7 Factor Model (Original all items)	.573	.587	.596	.027	.027	.027	164
7 Factor Model (Revised Best of)	.708	.711	.722	.033	.032	.033	72
7 Factor Model (Revised with statistically selected E)	.710	.713	.723	.032	.032	.032	74
6 Factor Model (Without Engineer)	.720	.721	.732	.034	.033	.034	64

* Whilst all the CFI values are poor, the values of the revised models have all increased from the original by a minimum of .126.

Validation of Results

A second maximum-likelihood confirmatory factor analysis was conducted (as reported in Table 9 on p.62) using the remaining half of the data (n = 13,623) in order to confirm that the current results were not due to chance fluctuations in the sample data or a statistical anomaly. The data obtained in the resulting analysis indicated that results for both

CFAs did not differ greatly statistically. Thus it is reasonable to conclude that the results obtained in the current study were due to the item manipulations and not due to chance.

Table 9: Second CFA Goodness of Fit indices for current and revised Humm seven factor models (n = 13, 623)

Model	NFI	TLI	CFI	RMSEA	Lo90	Hi90	# of Items
7 Factor Model (Original all items)	.571	.585	.594	.027	.027	.027	164
7 Factor Model (Revised Best of)	.706	.709	.720	.033	.032	.033	72
7 Factor Model (Revised with statistically selected E)	.707	.710	.721	.032	.032	.033	74
6 Factor Model (Without Engineer)	.718	.719	.730	.034	.033	.034	64

DISCUSSION

The main aim of this study was to examine the psychometric properties of the Humm Wadsworth Temperament Scale (Humm) and attempt to generate a more statistically sound and valid model. In order to achieve this, Thurstone's method of paired comparisons was conducted on the current items of the Humm. This was followed by Humm experts and users making conceptual judgements regarding the appropriateness and suitability of the current item allocation across the Humm's seven components. Further to this, investigation into whether it was possible to obtain superior results through quantitative methods and confirmatory factor analysis was conducted. In the passages that follow, the exposition centres on the implications of the current findings in light of these aims.

In terms of some of the statistical challenges experienced in the present analysis, none of the models created (current or revised) were spared from producing a significant chi-square value, suggesting that all models did not provide an adequate fit to the data. However, additional goodness-of-fit statistics that were calculated in order to address the chi-square limitations proved significant and were consistent with the assertions of researchers such as Bentler and Bonett (1980), Byrne (2001), and Roberts, Chernyshenko, Stark and Goldberg (2005). Regression weights did vary in their level of significance for some models, although on the whole, the statistics calculated typically reached an appropriate level for reasonable assumptions to be made from the findings.

For the most part, the current investigation improved the fit of the seven components and the overall model. Applying Thurstone's method of paired comparisons to gain a 'tighter' set of items resulted in a significant increase in goodness-of-fit statistics for each of the seven components when analysed in one large seven-factor model. However, the negative correlations produced for estimates of the Engineer model were indicative of this component potentially not measuring what it is intended to measure, and this suggests that there is a significant need for a full revision of the items currently used in the Engineer component. This is further supported by the finding that when the Engineer component was removed from analysis completely (creating a six-factor model), the goodness-of-fit indices for the over all model did not alter significantly. This is a disturbing discovery given that the Humm (including the Engineer component) is currently being utilised commercially and brings to light significant ethical and legal implications for the Humms continued use in its current state.

Empirical evidence gathered during the current study indicates that a full revision of the Humm is not only justified, but is of critical importance in order to maintain the integrity of the measure. The goodness-of-fit indices for the current seven-factor model were all considerably below the values suggested by Bentler & Bonett (1980), Bentler (1990) and Byrne (2001) as demonstrating that the hypothesised model represents an adequate fit to the data. The only exception was the existence of small RMSEA values, combined with narrow confidence intervals, which can be reflective of a good model fit (Byrne, 2001). Regression weights for each of the items in the current seven-factor model were also outside the acceptable range as they included negative values. The existence of negative regression weights indicated that there were a number of items in the current

version of the Humm that were non-significant or did not adequately contribute to a component. This effect was most apparent in the Engineer component.

In contrast, the goodness-of-fit indices for the revised seven-factor model (with the Engineer component reviewed statistically rather than conceptually), whilst not quite reaching the suggested thresholds indicative of sufficient fit, were a significant improvement from that of the current seven-factor model Regression weights for each of the items also increased and were within an acceptable range. In essence, the statistical qualities of the revised version were a vast improvement from that of the current version. It is noted, however, that correlations between the Humm's components in both the current and revised seven-factor models, were relatively consistent in their range (although, for the revised model, correlations were typically stronger).

Additionally, the number of items in the questionnaire, and therefore the time it takes an individual to complete it, has been significantly reduced. Data suggests that this shorter version of the Humm questionnaire is feasible without reducing the statistical significance of the measure. A positive consequence of the current research is that of commercial viability in terms of the burden of time it takes an individual to complete the questionnaire. In the present study the number of items in the questionnaire has been reduced from 318 to 74 and the statistical significance of the measure has actually increased in the process. However, a possible repercussion of this reduction in items could be that some of the components of the Humm may have too few questions in the revised version. In saying this, empirical evidence suggests that the revised version of the Humm is still statistically significant, and in fact data suggests that the level of significance has increased

from that of the original version. Furthermore, unidimensionality was achieve in the revised model, as all items that loaded on to more than one component were removed or reassigned to a single component. It is also possible that the wording of the remaining items is too similar, and therefore may overlap in which components they measure. Whilst steps were taken to ensure that the final items were significantly different from each other, an investigation into the quality of item wording should be conducted to ensure that items in the final questionnaire are suitably distinct.

Humm and Wadsworth (1935) ensured that no less than twelve items were attributed to a subcomponent. However, this is not the case in the revised version where there are nine items in the Normal component, fifthteen in the Hustler component, eleven in the Mover component, eleven in the Double-Checker component, six in the Artist component, twelve in the Politician component, and ten items in the Engineer component. A suggested next step in investigating the Humm further could therefore include field testing the revised measure in order to gather new data on the validity and factor structure. Basic psychometric statistics on both the revised measure and the original measure can then be provided to the current users of the Humm for comparison, and to gain current user feedback and input into the revision process.

The use of Thurstone's method of paired comparisons is further supported by the findings from the current study. This quantitative revision technique was included in the present research for the purpose of testing for item quality conceptually and for extracting psychological scale values for use in forming an alternative set of items for each Humm model. Thurstone (1959) and other researchers also assert that this method is robust enough

to stand on its own merits (Luce, 1963; Mosteller, 1963). The overriding benefit of using Thurstone's method is that it has a long history of use and a superior theoretical basis, which in the current research has facilitated the statistical improvement of the revised measure and therefore enabled the achievement of the aims in the current research.

Related to this, in the present investigation the quality of useable information yielded from Thurstone's method of parried comparisons was extremely high. This offers additional support for the use of this method. A unique feature of this technique is the ability to review information from a conceptual standpoint whilst still being able to derive tangible results that can be used as a basis for decision making. The broader application of Thurstone's method of paired comparisons beyond psychometric test revision is therefore strongly encouraged.

Limitations

Whilst the results from the present study were validated with a second confirmatory factor analysis, and robust statistical techniques and procedures were employed to ensure that valid and reliable findings have been reported, it is an inevitable truth that the research and findings of the current study possesses limitations. This is particularly the case with regards to the applicability of the results and the ability to generalise the findings to domains outside the scope of the current project.

With regards to the negative correlations in the Engineer model, the question does arise, are the results still valid considering that one of the seven components, the Engineer

component, appears to be flawed in some way? If so, is the Humm therefore fatally flawed and should its use be discontinued? In the measure's favour, there is empirical evidence to suggest that the other six components have some redeeming qualities. When the model was reduced to six components (that is when the Engineer component was removed altogether), the overall goodness of fit statistics did not alter greatly from those of the seven-factor model. When the Engineer component was evaluated by employing a different item identification method, by statistical comparisons to identify and remove items that yielded negative estimate values, a more statistically valid model was created. When this model was included in the seven-factor model, negative estimate values were no longer present. However, it should be noted that the items selected in this instance were chosen purely for their statistical validity rather than for their conceptual validity. Thus it is worth considering that utilising Thurstone's method of paired comparisons on the items contained within the Engineer component may not be appropriate. Alternatively, it is possible that the Engineer component (and associated items) is perhaps a component that experts and users had more difficulty conceptually evaluating in the first place. Another explanation could be that the Engineer component may not conceptually belong in the Humm model. It may not properly relate to or align with the definition and framework of temperament that currently underpins the Humm. The other six components being investigated individually still maintained a level of statistical strength when combined together in one large model. Thus one could surmise that results borne from a model that includes a more statistically challenging component such as Engineer may still in fact be valid and worthy of further consideration.

Practicality considerations offer support for and against the revision of the Humm to the extent that the structure is significantly altered. The current training required to become accredited in the interpretation of the Humm is typically six months in duration. Reducing the number of items, subcomponents and possible components in the measure may also have benefits in terms of reducing the time it takes to train new users. However, current experts and users of the Humm would have to re-learn many aspects of the measure and this may be difficult depending on how entrenched their previous understanding of the measure is.

The very large sample size utilised in the current study, as well as the confirmation of findings through conducting a second confirmatory factor analysis strongly suggests that the results obtained in the present investigation are not an anomaly and in all likelihood can be replicated using other data samples. However, the limited number of Humm experts and users (of which there were only six) who contributed to the conceptual analysis of the items in the Humm may have impacted upon the obtained results, particularly as this part of the analysis relied heavily on an individual's personal subjectivity and understanding. Therefore, it may be worthwhile to replicate this study with additional or different Humm experts and users to corroborate the reported findings.

Another consideration is the fact that the only official way of scoring[5] the Humm is by using the scoring system in the sole possession of the measures owners. All the participants in the current study and indeed all individuals subjected to the measure in a

[5] The Humm Wadsworth Temperament Scale scoring keys are not published, nor are they available for purchase.

commercial setting were assessed against Australian population norms. This normative sample has been standardised to the general Australian population, and is unlikely to be as culture-fair as scoring techniques that are calibrated for the local ethnic mix of regions in both Australia and New Zealand.

Although the number of years a participant had been speaking English was recorded, and often cognitive abilities were evaluated at the same time as completing the Humm, for the present study cognitive ability was not taken into consideration. This included an individual's verbal reasoning skills and the terminology and level of sophistication used in the wording of Humm items. Further investigation into the impact that an individual's cognitive abilities may have on their responses to temperament and personality measures is therefore recommended. This position is supported by Salgado (1999), who stated that personality measures should be used together with other predictors of performance such as ability tests and previous experience. Whilst in summarising a person's temperament, factors such as their cognitive abilities, qualifications, previous experience and their ability to do the job at hand are taken into consideration, research into the impact that these factors may have on how a person actually responds to a temperament questionnaire in the first place have not been taken into consideration. In some cases this is despite psychologists commenting on whether an individual has fully understood the questions being asked in the temperament questionnaire (particularly in the event of an individual gaining a very low verbal reasoning score).

Further to this, whilst the original Humm has a built in 'faking' measure, the current investigation did not examine or make allowances for the possibility of participants faking

their responses. Indeed within the current sample it is safe to assume some participants did not answer the questionnaire as honestly as they could. The impact of such responses is yet to be investigated. Also worthy of consideration is the effect a person's temperament has on how they actually approach and respond to the questionnaire; for example, whether a person is objective, or has a tendency to over analyse each question. In the past, such characteristics in the extreme have been known to cause distorted temperament profiles.

Moreover, although the exact number is unknown, a large proportion of participants in the data sample for the present study were aware, prior to completing the Humm questionnaire, that their responses would be used as part of a decision-making process that could influence their professional career. This adds value in terms of the applicability of the reported results in a recruitment or selection setting. However, it would be interesting in future to assess the differences between various groups by using responses that were provided for the sole purpose of research, rather than as part of some kind of potentially life-changing decision-making process.

Future Directions

Additional to the abovementioned, the present study has also raised many issues that could be examined in future research. For example, it may be beneficial to investigate and compare the similarities and differences of Aboriginal and Maori descendants as well as other minority populations, with individuals of European descent. When developing a normative sample, Butcher (2000) suggests that it is important to sample diverse ethnic group membership. This is to ensure that the sample is balanced for ethnic group

membership and represents the national census of the country in which the measure will be used. For these reasons, comparisons between female and male participants should also be investigated to see whether norms based on gender may also be necessary. Whilst gender was recorded in the current research, possible gender differences when understanding or responding to an item, was not considered and warrants further investigation, particularly as in the current case, 63 percent of participants were male.

Analysing responses by gender, age or other factors may also be worthwhile to gain a greater appreciation of the way in which different individuals approach a self-rating questionnaire. Future research may also seek to examine differences by sector, and across a variety of jobs and occupations to determine whether unique differences exist between employees working in a variety of roles or hierarchies.

Related to the above is the possibility that responses from participants with a Western background may more closely align with a fair representation of self, compared with participants from other ethnic backgrounds. Bartram (2004) suggests that international organisations are more commonly seeking to adopt consistent selection and recruitment practices across a variety of different countries. This raises the issue of the practicality of using a measure that is not "culture-fair" in a multicultural environment such as Australia or New Zealand. Some of the terms used within an item have a culturally inappropriate aspect, for example, questions of a gambling nature may discriminate against cultures where gambling is strictly prohibited. Redevelopment, rewording or the creation of new items in order to counter these concerns, is thus recommended. Butcher (2000) supports this stating that when revising a measure it is important to consider improving the wording

and reducing the culture-bound aspects of a measure, making as much use of culturally neutral language as possible. As there is no guidance on score adjustments for ethnic groups provided in the Humm manual, in order to increase the ability to generalise the results gained from the Humm, a review or survey of the prominent cultures and ethnicities that may be exposed to the Humm is suggested. A more detailed investigation of the measure to identify any between-group differences and investigate whether there are items within the Humm that are discriminative in any way is also recommended.

The current analysis was conducted on the seven major temperament components of the Humm, and not at the 31 subcomponent level. Further analysis into the possibility of breaking down the new components is now possible. This could include investigating whether certain items within a component cluster together in order to identify the potential sets of items that could create new subcomponents. Breaking down the seven components into a finer analysis would address the concern that any given individual may manifest some of the tendencies associated with a given component, but may not possess all of them. Therefore, a microanalysis of the measure is recommended.

Related to the above, some of the correlations between the seven Humm components are exceptionally high. This may indicate that the current seven factor structure should be reviewed as the question does arise, are the seven components really distinct and separate? It is therefore feasible that the temperament model underpinning the Humm may benefit from being broken down further and the logical starting point for this is at the current 31 sub-component level.

As an aside, at the commencement of this research it was hoped that Item Response Theory (IRT), a relatively recent psychometric approach, might be employed to enhance more traditional psychometric procedures (Embreston, 1996). However, this was not possible within the scope of the current project. Despite this, it is suggested that future research on the Humm incorporates this approach. To present knowledge, the current research is the first to deviate from using more classical analysis methods. Therefore, utilising a more modern technique such as IRT would increase the brevity of methods used to analyse the psychometric properties of the Humm.

CONCLUSION

With the hindsight gained from the current research, the chosen path appears to have been an appropriate first step for evaluating temperament and the Humm Wadsworth Temperament Scale (Humm). The foundations for further research into temperament first required a detailed investigation into the validity of any conceptually questionable items in the Humm. Quantitative theory demonstrated the viability of an alternative revision protocol that could be applied to all items across the questionnaire. The data produced by Thurstone's method of paired comparisons increased the statistical validity of six of the seven single factor models and these single factor models went on to improve the large seven-factor measure. The current research supports the assertion that a multi-stage, mixed methods approach to psychometric test development is not only appropriate, but also yields strong findings that can be utilised to create a more robust measure.

The present study also reduced the number of items in the questionnaire from 318 to 74, as poor loading items were identified as reducing the validity of the measure and were subsequently removed. Concern over the validity of the measure is of commercial, ethical and legal significance to all stakeholders in the continued use of the Humm. Fortunately it appears that the psychometric flaws identified may be alleviated by the removal of ambiguous or poor items, and a comprehensive review and restructure of the current components.

The burden of time required for individuals to complete the questionnaire was reduced; this was also an important result for increasing the continued commercial viability

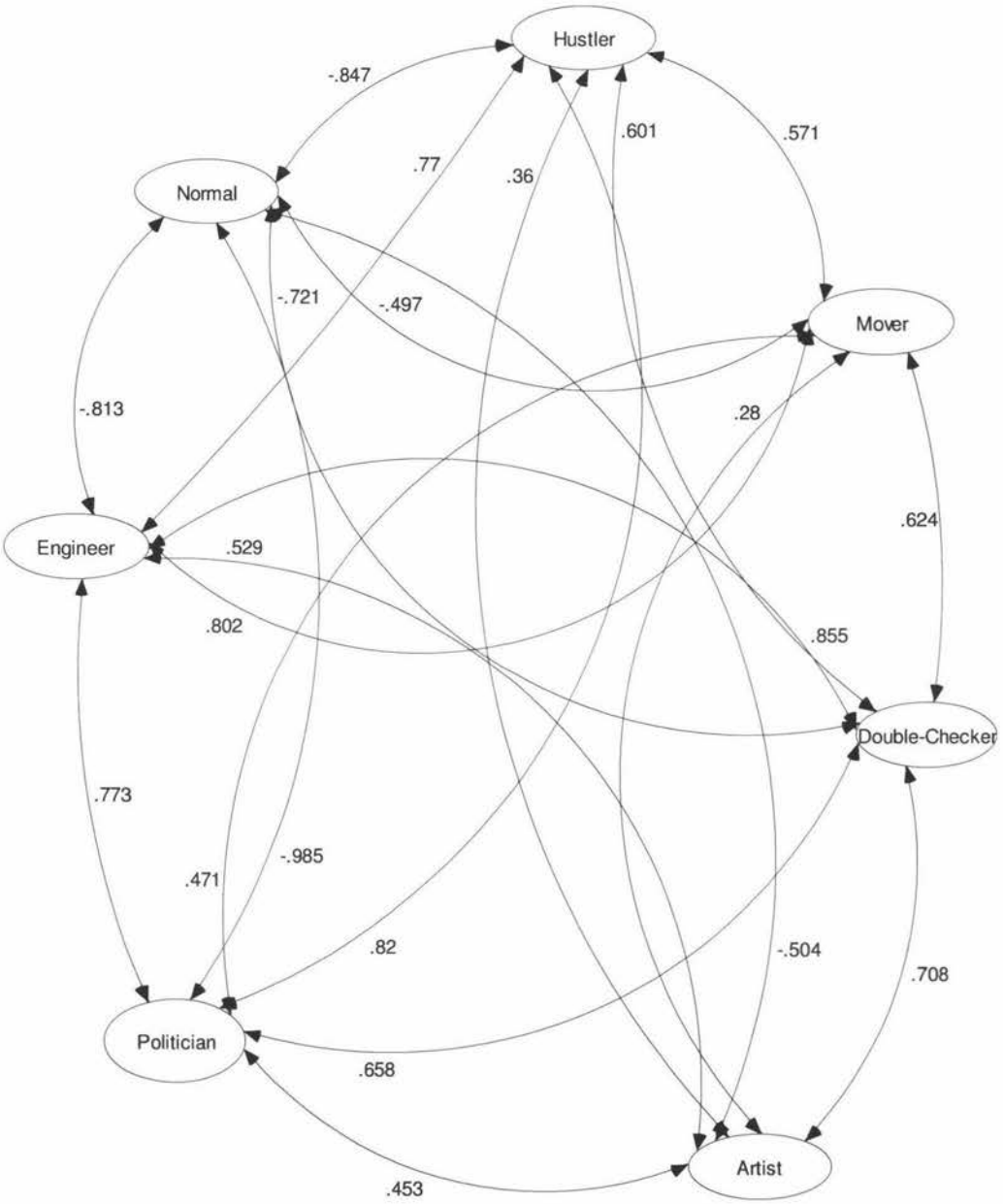
of the measure. The current study clearly demonstrated the benefits of utilising Thurstone's method of paired comparisons, and shows great promise for the future application of this technique to the construction and review of temperament and personality measures.

The application of temperament and personality measures such as the Humm to human resource assessment is continuing to experience major growth worldwide. It is noted, however, that without promptly addressing the psychometric properties of the Humm and the norms used to score the Humm, further encouragement for organisational psychologists, human resource managers and other professionals to make use of the information gained from the Humm, could result in unfair treatment of employment-seekers and those currently employed. Thus investigation and revision activities are strongly recommended as a matter of urgency in order to maintain the reliability of the measure.

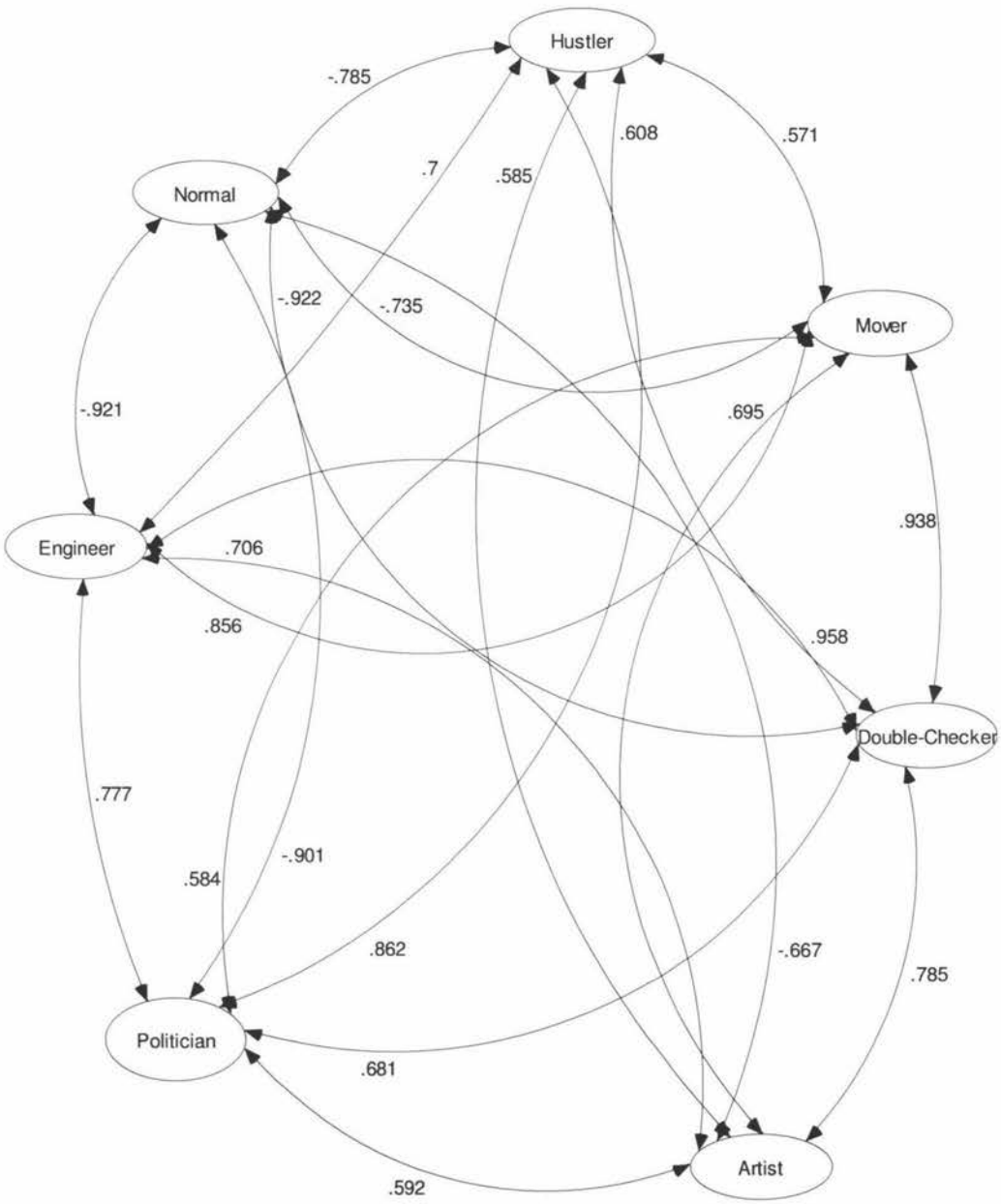
The results gained from the present investigation certainly suggest that the Humm requires a major revision. The current version of the Humm does display some reliability and validity characteristics, however, the present analysis has offered an alternate model that is more statistically valid, adds conceptual clarity and has fewer items. Whilst the major criteria required for robust psychometric test revision have not all been satisfied, the revised model of the Humm is certainly an improvement on the original. In sum, whilst some statistical challenges have been identified, the Humm is not flawed to the extent that the measure should be disregarded. However, proactive steps are required to ensure the ongoing integrity of the measure.

The psychometric problems identified in the Humm are, however, by no means isolated. There is an urgent need to apply improved measurement techniques and revised methodologies to many of the temperament and personality measures that are currently being used in an industrial or organisational setting. The primary concern: to minimise the potential unfair treatment of individuals and prevent a loss of confidence in psychometric assessment.

Appendix I: Standardised Correlations for Current Seven-Component Model



Appendix II: Standardised Correlations for Revised Seven-Component Model



BIBLIOGRAPHY

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, New Jersey: Simon & Schuster.

Barrick, M.R., & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M.R., & Mount, M.K. (1993). Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. *Journal of Applied Psychology, 78,* 111-118.

Bartram, D. (2004). Assessment in organisations. *Applied Psychology: An International Review, 53*(2), 237-239.

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*(6), 1185-1203

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Borkenau, P. (2001). Issues in the measurement of temperament and character. In: Collis, J.M., & Messick, S. (Eds). *Intelligence and Personality Bridging the Gap in Theory and Measurement.* pp. 99-112, Mahwah, New Jersey: Lawrence Erlbaum Associates.

Butcher, J.N. (2000). Revising psychological tests: Lessons learned from the revision of the MMPI. *Psychological Assessment, 12*(3), 263-271.

Byrne, B.M. (2001). *Structural equation modeling with AMOS Basic concepts, application and programming.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Coombs, C.H., Dawes, R.M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Englewood Cliffs, New Jersey: Prentice-Hall.

Costa, P.T., & McCrae, R.R. (1992). *The NEO-PI-R professional manual.* Odessa, FL: Psychological Assessment Resources.

Dicken, C. (1963). Good impression, social desirability, and acquiescence as suppressor variables. *Educational and Psychological Measurement, 23*, 699-720.

Dysinger, D.W. (1939). A critique of the Humm Wadsworth Temperament Scale. *Journal of Abnormal and Social Psychology, 34*, 73-83.

Edwards, A.L. (1957). *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts.

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349.

Endler, N.S. (1989). The temperamental nature of personality. *European Journal of Personality, 3*, 151-165.

Goldsmith, H.H., & Campos, J.J. (1982). Toward a theory of infant temperament. In: Emde, R.N. & Harmon, R.J. (Eds). *The Development of Attachment and Affiliative Systems*. pp. 161-193, New York: Plenum Press.

Gray, J.A. (1973). The psychophysical nature of introversion-extroversion: A modification of Eysenck's theory. In: Nebylitsyn, V.D., & Gray, J.A. (Eds). *Biological Bases of Individual Behaviour*. pp. 182-205, New York: Academic Press.

Grimm, L.G., & Yarnold, P.R. (1995). Introduction to multivariate statistics. In: Grimm, L.G., & Yarnold, P.R. (Eds). *Reading and Understanding Multivariate Statistics*. pp. 1-18, Washington: American Psychological Association.

Hogan, R., & Nicholson, R.A. (1988). The meaning of personality test scores. American *Psychologist, 43*(8), 621-626.

Howell, D.C. (1997). *Statistical methods for psychology*. Belmont, CA, USA: Duxbury Press.

Hull, J.G., Lehn, D.A., & Tedile, J.C. (1991). A general approach to testing multifaceted personality constructs. *Journal of Personality and Social Psychology, 61*(6), 932-945.

Humm, D.G., & Humm, K.A. (1944). Validity of the Humm Wadsworth Temperament Scale: With consideration of the effects of subject's response bias. *Journal of Psychology, 18*, 55-64.

Humm, D.G., & Wadsworth, G.W. (1935). The Humm Wadsworth Temperament Scale. *American Journal of Psychiatry, 92*, 163-200.

Humm, D.G., & Wadsworth, G.W. (1941). Using the Humm Wadsworth Temperament Scale. *Journal of Applied Psychology, 25*, 654-659.

Humm, D.G., & Wadsworth, G.W. (1943). Temperament in industry. *Personnel Journal, 21*, 314-322.

Hurtz, G.M., & Donovan, J.J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869-879.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement.* (Vol. 1). New York: Academic Press.

Kruger, B.L. (1938). A statistical analysis of the Humm Wadsworth Temperament Scale. *Journal of Applied Psychology, 22*, 641-52.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

Luce, R.D. (1963). On the possible psychophysical laws. In: Luce, R.D., Bush, R.R., & Galanter, E. (Eds.). *Readings in mathematical psychology.* (Vol. 1). New York: John Wiley and Sons.

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Mayer, J.D. (2005). A tale of two visions. Can a new view of personality help integrate psychology? *American Psychologist, 60*(4), 294-307.

McCrae, R.R. (1986). Well-being scales do not measure social desirability. *Journal of Gerontology, 41*, 390-392.

McCrae, R.R., & Costa, P.T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882-888.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mosteller, F. (1963). Remarks on the method of paired comparisons. In: Luce, R.D., Bush, R.R., & Galanter, E. (Eds.). *Readings in mathematical psychology.* (Vol. 1). New York: John Wiley and Sons.

Pervin, L.A. (2002). *Current controversies and issues in personality, (3rd Ed.)*. New York: John Wiley & Sons.

Pervin, L.A., Cervone, D., & John, O.P. (2005). *Personality Theory and Research, (9th Ed.)*. Hoboken, New Jersey: John Wiley & Sons.

Roberts, B.W., Chernyshenko, O., Stark, S., & Goldberg, L.R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58(1)*, 103-139.

Rosanoff, A.J. (1927). *Manual of Psychiatry, (6th Ed.)*. New York: John Wiley & Sons.

Salgado, J.F. (1999). Personnel selection methods. In: Cooper, C.L., & Robertson, I.T. (Eds.). *International review of industrial and organizational psychology*. Vol. 14. Chichester, United Kingdom: Wiley

Sheldon, W.H., & Stevens, S.S (1942). *The Varieties of Temperament*. New York: Harper & Row.

Smith, T., Gudmand, R., & Marke, S. (1958). The internal Consistency of the Humm Wadsworth Temperament Scale. *Journal of Applied Psychology, 42*, 234-240.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Strelau, J. (1987). The concept of temperament in personality research. *European Journal of personality*, 1, 107-117.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Thurstone, L.L. (1927a). A law of comparative judgement. *Psychological Review, 34*, 273-286.

Thurstone, L.L. (1927b). The method of paired comparisons for social values. The *Journal of Abnormal and Social Psychology, 21*, 384-400.

Thurstone, L.L. (1927c). Psychophysical Analysis. *American Journal of Psychology, 38*, 368-389.

Thurstone, L.L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology, 26*, 249-269.

Thurstone, L.L. (1959). *The measurement of values*. Chicago: The University of Chicago Press.

Torgerson, W.S. (1962). *Theory and methods of scaling*. New York: John Wiley & Sons.