Towards a hierarchical and multifaceted model for the measurement of academic self-concept in science

Graham Hardy

Towards a Hierarchical and Multifaceted Model for the Measurement of Academic Self-Concept in Science

Graham Hardy

Submitted in Partial Fulfilment of
the Requirements for the
Degree of Doctor of Education
School of Education, University of Durham

2007

ACKNOWLEDGEMENTS

Undertaking the EdD course has been a wonderful and inspiring journey. I have learnt so much and worked with such astonishing people. I was blessed throughout the course to be taught by a group of excellent teachers and scholars. I would like to give a huge vote of thanks to all the tutors for their preparation, insight and enthusiasm. I think it is fitting to mention you by name; Keith Morrison, Mike Byram, Peter Tymms and David Gallaway. Thanks also to Joe Elliott for his guidance through the hardest phase. This thesis is the culmination of many years' work, and it has been worth every last minute.

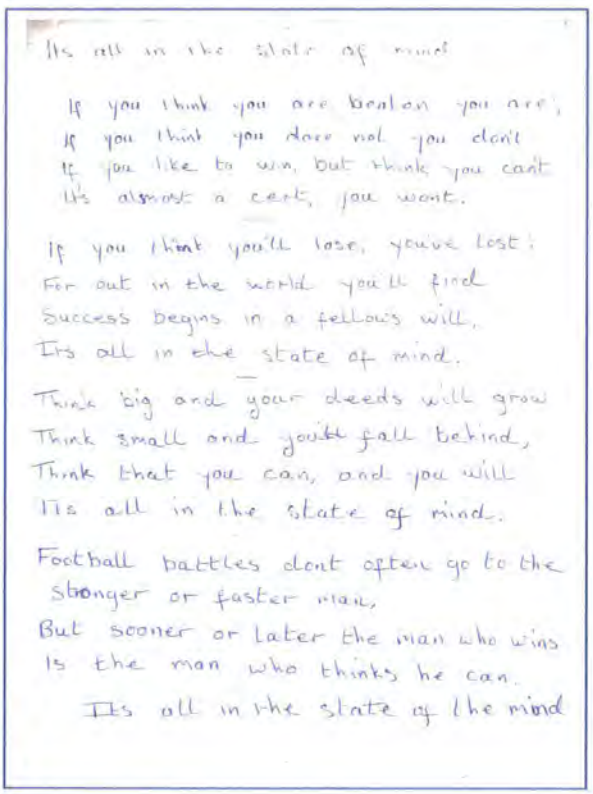
The most special thanks is reserved for my family, near and far, young and old, for without their constant and present willingness for me to succeed, the journey would have been longer, harder and not nearly as much fun. To all of you ... thank you.

Graham Hardy

DEDICATION

During the long hours and days of writing this thesis, my mind happened to wonder to my childhood and then to my school days. It rested upon a memory of a hand written poem my mother had put onto the wall of my bedroom to inspire me to achieve better in all that I did. Some months later, having completed the writing, my mind again fell onto this poem. My attic housed three small boxes each containing papers collected from my parents' home after they had passed away. In the third of these boxes, lodged against my old school reports and old class photographs was a slightly creased and aged tinted sheet of A4 paper on which my mother had written the poem.

Little did my parents know that such sentiments would form the corner stone of my doctorate study. If they had, they would have been very proud. The words now have a new resonance. Twee they might be, but they are printed below because in their unsophisticated way they communicate more wisdom than the 200 pages that follow.



It's all in the state of mind

If you think you are beaten you are;
If you think you dare not you don't
If you like to win, but think you can't
It's almost a cert, you won't.

If you think you'll lose, you've lost;
For out in the world you'll find
Success begins in a fellow's will,
It's all in the state of mind.

Think big and your deeds will grow
Think small and you'll fall behind,
Think that you can, and you will
It's all in the state of mind.

Football battles don't often go to the
stronger or faster man,
But sooner or later the man who wins
Is the man who thinks he can.

It's all in the state of the mind

This thesis is dedicated to the memory of my mother and father, who lived for too few years and died on 19 November 1997 and 25 November1995 respectively.

It was their belief in self-evaluation and the constant striving for improvement which so influenced my life.

ABSTRACT

Recent research into academic self-concept has included investigations into domain-specific self-concepts. Examples include Lau et al (1999) within English, Marsh et al. (1997) within physical education and Vispoel (2003) within Music. They have all indicated these subjects to be multidimensional in nature, consisting of distinct sub-domains. This is an important finding and helps teachers and researchers to understand how pupils feel about themselves as learners. In contrast there have been few, if any, studies about this within the subject area of science. Up to now self-concept in science has been conceived as a uni-dimensional construct.

Using structural equation modelling this study explored the multidimensional and hierarchical nature of self-concept in science. The outcomes show that science self-concept of secondary aged pupils is heterogeneous in nature and presents a consistent, stable and valid set of measures for the ways in which school pupils feel about themselves when learning science. It argues that learners have a multidimensional self-concept 'profile' which represents their psychological response to being a learner of science.

An instrument has been developed and validated for the measurement of science self-concept for secondary aged pupils (11 – 16 years). Carrying out model fit analysis using LISREL 8, the instrument has been shown to be extremely robust in measures of fit and construct validity, and has also shown itself to be invariant across sex and age subgroups

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DIAGRAMS

LIST OF CHARTS

Towards a Hierarchical and Multifaceted Model for the Measurement of Academic Self-Concept in Science

Graham Hardy

Submitted in Partial Fulfilment of
the Requirements for the
Degree of Doctor of Education
School of Education, University of Durham

2007

INTRODUCTION

Rationale

The purpose of this study was to explore the multidimensional and hierarchical nature of academic self-concept in science; to provide a model for the way in which academic self-concept in science may be conceived within the context of an extension of the Marsh/Shavelson Model (Marsh et al., 1985); to use the model as a basis for forming a comprehensive instrument for the measurement of academic self-concept in science; to explore gender and age-group issues in relation to academic self-concept in science; and to explore the potential utility of the self-concept measurement instrument for the teachers and learners of science.

The study was motivated by two perceived gaps within the self-concept research literature. First, research into academic self concept in science has generally been sparse, and research specifically into the discrete and distinct components of science self-concept has been particularly under investigated. Where such research has occurred, (for example, Ireson, Hallam and Plewis, 2001; Krogh and Thomsen, 2005; Pajares, Britna, and Valiante, 2000) existing scales for measuring academic self-concepts in other subjects have been adapted to investigate science. Notably, this has been achieved by simply substituting the word 'science' (or 'physics' in the case of Krogh et al.) for the subject label for which the instrument had previously been used, e.g. 'mathematics'. Little research, if any, has been carried out on a deconstruction of the term 'science' to explore how the impact of its individual facets affects the self-concept of learners.

Second, the hierarchical nature of self-concept had generally been under researched. Yeung et al., (1999) reported that 'despite the voluminous literature on self-concept, there is little evidence for a hierarchical academic self-concept' (p.378). They point to their own study of commercial studies students and Marsh et al.'s (1997) study of physical self-concept among elite athletes as being some of the very few studies to demonstrate conclusively both the multidimensionality and hierarchical nature of academic self-concept. Also, with a few notable exceptions, (see Yeung, 1999; Lau, Yeung, Jin and Low, 1999; Vispoel, 1995) research into the hierarchy of self-concept has tended to stop at the general subject level. Explorations into possible extensions of the structure beyond the general subject level have been

neither widespread nor systematic, despite Marsh, Shavelson and Byrne (1988) calling for such research activity nearly twenty years ago.

The Case for Deconstructing the Term 'Science'

Science Education in England and Wales has undergone considerable change in the last 40 years. These changes have been driven partly by modifications to school structures, such as the move from selective schools to comprehensive education; partly through curriculum innovation and reform, such as Nuffield Science, Warwick Process Science and latterly Twenty First Century Science; partly through major funding streams, such as TVEI; and partly through HMI reports and papers, such as Science 5-16: A Statement of Policy. The largest wholesale changes, however, have been brought about, not unsurprisingly, by two government statutory requirements, (at least for state maintained schools). The first of these was the National Curriculum for Science which first reached the classrooms in 1989, (DES 1988), and in its latest incarnation went into operation in maintained schools in September 2006 (DfES, 2006). The second was the Key Stage Three Strategy (now National Strategy) for Science, (DfES 2000) which has been around for half a dozen years although is concerned more with science pedagogy rather than the science curriculum[1]. One of the touch stones of good science education, post National Curriculum, has been the concept of a balanced science curriculum. The introduction of the National Curriculum brought with it, for the first time, an entitlement to a science education for all pupils attending state maintained schools up to the age of 16 years. Further to this the curriculum was required to be broad and balanced in nature. This essentially meant incorporating physics, chemistry, biology with a new earth science curriculum into a whole science scheme of work. Examination boards responded by designing GCSE courses which complied with the government requirements and which were variously structured as 'coordinated science', 'integrated science', 'modular science' or three separate sciences.

Some schools took the opportunity to blur the distinction between the traditionally strong notions and often distinct separation of physics, chemistry and biology teaching. Some of this blurring was pedagogically driven in the desire to unify the

[1] Curriculum is being conceived here in a narrow sense as the matters to be taught and performance indicators to be assessed.

study of science and to reduce what some people perceived as artificial boundaries between the subject (see Smithers and Robinson, 2005). Others perceived this to be driven by the need to 'manage' the real and urgent lack of physics teachers in secondary school[2]. In whichever form the new curriculum was 'delivered' to pupils, and despite the many and varied contexts[3] used to exemplify the knowledge, skills and understandings, the fingerprints of 'physics', 'chemistry' and 'biology' have remained ever present, whether consciously maintained or not.

The physics, chemistry and biology distinction still remains strong today. The experience of most secondary aged pupils in most secondary aged schools consists largely of being taught topics which can more or less be fitted into the physics, chemistry and biology categories. They are taught by science teachers who see themselves trained and working within their physics, chemistry and biology specialisms, within teaching laboratories which may be similarly named and equipped, and with their Key Stage 3 and 4 learners moving towards A-level courses comprising these very same subjects. Many pupils in their mid-secondary years already know the science subjects at which they think they are good and which they enjoy. Physics, for too many pupils for instance, is already conceived as overly hard (Angell, Guttersrud and Henriksen, 2004), and there are a disproportionately high number of girls for whom physics holds little association or interest (EMBO Reports 2005; van Langena, Rekers-Mombargb and Dekkersa, 2006; Zohar and Bronshtein, 2005). There is also well documented evidence of gender differences relating to pupils' post 16 science subject choices and this can be seen by the extraordinarily and disappointingly low number of girls opting into physics classes at schools, colleges and universities (see Murphy and Whitelegg, 2006) and then consequently into physical science based careers (EOC, 2002). For the vast majority of school-girls the subject of physics holds no long term future for them. Pupil performance within physics, chemistry and biology is just as varied as the pupils' motivation to pursue the different subjects at a higher level of study (van Langena, Rekers-Mombargb, and Dekkersa, 2006). Students' A-level choice of science subjects is a

[2] See the Institute of Physics document to the House of Lords (IoP, 2006) and Smithers and Robinson, (2005), for an analysis of the scale and urgency of the problem.
[3] Much curriculum initiative has been forthcoming in science, particularly recently, to teach (or deliver) the science content through a variety of 'contexts' be these industrial, social, economic, sports or health related etc. The motivation for this can be captured under three broad groupings, (i) curriculum accessibility, (ii) linking learning to real life situations, and (iii) greater inclusivity and motivational impact, (see for example Bennett and Lubden, 2006; Gilbert, 2006; Schwartz, 2006).

key example (see for instance Smithers and Robinson, 2006). At post 16 level couplets of subject choices often include physics and chemistry, or chemistry and biology but rarely physics and biology except with chemistry as well (UCAS, 2006). Students' attitude to the different sciences together with their motivation for studying the different science subjects is evidently distinct and well established (Osborne and Collins, 2000; Miller, Blessing and Schwartz, 2006).

With this in mind there seems to be something incomplete with measures of a self-concept which aims to find a single measure for 'science' without recognising the diversity of the subject and the differential impact it has on young learners. This is not to say that *the sciences* do not belong together as members of a cognate group. There are some strong characteristics which typify scientific study; particularly in methodology, or in the kinds of knowledge being pursued and valued, or in the personal qualities science teachers aim to foster, e.g. respect for evidence, tolerating uncertainty and showing perseverance (NCC, 1989). These may provide some form of loose overarching dimension, but it is at best partial and less than watertight.

From the science teachers' perspective, it does seem reasonable to assert that if academic self-concept is to be a useful measure, with context validity, then it ought to recognise the multifaceted nature of science. Self-concept in science as a unitary construct is consistent with neither the curriculum structures nor pupils' attitudinal responses to it (Osborne, et al., 2000). From the psychological perspective, there are structural and methodological reasons, as argued here and by others (Marsh, Shavelson and Byrne, 1988), that an extension of the subject specificity hierarchy which has been reported in other subjects, like English and creative subjects (Lau, Yeung, Jin and Low, 1999; Vispoel, 1995) should be equally applicable and valid to the subject of science.

Theoretical Framework

In any research study a clearly defined theoretical and methodological framework should occupy the base upon which everything else is built. Indeed, past research on self-concept has been heavily criticized, by Shavelson et al. (1976) and others (Burns, 1979; Bracken 1976) and with some justification, for weaknesses in both the conceptual base and the instrumentation used in the research activity. In order that similar criticisms are not justified here, particular care has been taken to address these issues.

The definition of self-concept used in this research is consistent with the contemporary and dominant view of self-concept and in line with the multidimensional and hierarchical structure as first posited by Shavelson et al. (1976) and revised by Marsh et al. (1985). In this study, multidimensionality will imply that the self-concepts facets, although intercorrelated, can be measured as separate constructs. This means that for example, although within the overall structure, particular facets of self-concept such as 'science self-concept' and 'mathematics self-concept' may well correlate strongly, these 'different dimensions operate as separate interpretable quantities', (Byrne and Shavelson, 1996, p.600). The hierarchical nature of the structure by contrast, will be suggestive of a pattern of correlational strengths between the self-concept facets at various levels, where for instance, 'general self-concept correlates highest with academic self-concept, next highest with subject-specific self-concepts, and lowest with academic achievement (i.e. actual behavior)', (ibid., p 600).

An extension of the Marsh/Shavelson model would require an explicit disentangling of the term 'science' with the insertion of a mini-structure which would more realistically reflects young people's self-concept responses to the subject. To achieve this, specific facets which constitute the science framework were designed and placed into the structure. Young learners' knowledge and attitudes to science is primarily constructed as a response to their school experiences. Consequently, it seemed reasonable to base this disentangled mini-structure of science self-concept on the curricula and pedagogical frameworks to which the pupils were exposed on a daily basis. In this way the self-concept instrument would be embedded within the learners' actual experiences, and hence achieve greatest curriculum relevance. The

items relating to the indicator (or manifest[4]) variables were matched carefully to the school science curriculum[5]. It was equally important that the self-concept measurement had a full and comprehensive coverage of the science curriculum and not merely a non-representative subgroup of it. This was achieved by choosing latent variables and manifest variables which matched the science content across the entire Programmes of Study for either Key stage 3 or key Stage 4 Science (DfEE 1999).

[4] Manifest variables will be discussed later, but they are the measured responses to the instrumentation items.
[5] For the sample of this study it meant an explicit connection to the Science National Curriculum for England (see DfEE 1999).

LITERATURE REVIEW

Importance and Impact of Self-Concept

> In very broad terms, self-concept is a person's perception of himself (sic) . . . . We do not claim an entity within a person called 'self-concept'. Rather, we claim that the construct is potentially important and useful in explaining and predicting how one acts. One's perceptions of himself are thought to influence the ways in which he acts, and his acts in turn influence the ways in which he perceives himself (Shavelson, Hubner and Stanton, 1976, p. 411).

Understanding self-concept is becoming ever more important as there is a growing recognition that a positive self-concept can facilitate a whole range of desirable educational outcomes. Such outcomes have been shown to include the potential to, (i) maximise children's personal and social adjustment (Branden, 1994; Harter, 1990; Tracy, 2002), (ii) influence motivation, effort and anxiety (Skaalvik and Rinkin, 1995) and (iii) raise academic performance (Chapman, 1988; Marsh, Byrne and Yeung, 1999). Indeed, Lawrence (1996) has stated, 'one of the most exciting discoveries in educational psychology in recent times has been the finding that people's levels of achievement are influenced by how they feel about themselves' (p.xi).

Some would go further and state that an individual's feelings toward self, such as self-concept, have more than mere correlational relationships with behaviour outcomes. Marsh, Craven and McInerney (2005), for instance, report that self-concept researchers have made such advances in the last two decades that they have established appropriate paradigms and have begun to disentangle the *causal influences* between self-concept and educational outcomes such as attendance, participation, achievement, coursework selection and school enjoyment.

The importance of self-concept and the enhancement of positive feelings about self have not only been recognised as a means of facilitating these desirable educational outcomes but have also become an educational goal in their own right[1]. Self-enhancement has found itself being promoted as an explicit aim in many curricula, government policy documents and funded educational initiatives.

[1] Much to the regret of some, see for instance, Baumeister, Campbell, Krueger and Vohs (2003)

The OECD (2003) identified 'student engagement' as a key developmental objective. Engagement, which they defined as consisting of a range of psychological components, including self-concept, was found to be closely associated with both economic success and long term health and well-being and as such, they concluded, ought to be considered alongside academic achievement as an important educational outcome. Another example can be found in Australia, the Adelaide Declaration on National Goals for Schooling in the Twenty first Century have stated as Goal 1.2:

> When students leave school they should have qualities of self-confidence, optimism, high self-esteem, and a commitment to personal excellence as a basis for their potential life roles as family, community and workforce members (p.13).

The California Department of Education (2002) recently published that schools should be producing:

> Programs and strategies that develop a student's sense of family and school connectedness, self-esteem, personal and social responsibility, character, and ability to resolve conflict in a positive, constructive way (online).

The heavily government funded National Strategy for England (DfES, 2004) promotes the teaching of Key Skills which 'improve pupils' capacity for independent learning and thus their self-esteem and motivation,' (p. 14).

There may well be good reason why government, local education authorities and schools should make such initiatives a priority, for there is substantial evidence to suggest that a balanced and successful adult life may well be contingent on positive perceptions of self. Parares (2002) contends that there is strong empirical data to show that beliefs about self competence touch virtually all aspects of people's lives; it controls whether their thinking is productive, debilitative, pessimistic or optimistic; how well they are motivated and their degree of perseverance; their vulnerability to stress and depression; and the life choices they make. This conception of scale and reach is echoed by Bandura (1986), who insists that:

> People who regard themselves as highly efficacious act, think, and feel differently from those who perceive themselves as inefficacious. They produce their own future, rather than simply foretell it (p. 395).

Indeed, it is sometimes only when things go wrong that it becomes most clear about the importance of things going right. This is summed up with Branden's (1994) much repeated reflection when he reported:

> I cannot think of a single psychological problem – from anxiety to depression, to underachievement at school or at work, to fear of intimacy, happiness or success, to alcohol or drug abuse, to spouse battering or child molestation, to co-dependency and sexual disorders, to passivity and chronic aimlessness, to suicide to crimes of violence – that is not traceable, at least in part, to the problem of deficient self-esteem (p. xv).

Research into self and particularly self-concept is, as highlighted by some of the most eminent thinkers in educational psychology, are of significant importance to many fundamental issues relating to the individual and or the individual's place in society. Its universal importance and its multidisciplinary nature makes self-concept a particularly appropriate choice for a focal point of a doctoral thesis.

Patchy Background

Research on self-concept has had an uneven history. Like many ideas undergoing rapid development, the early past of self-concept is chequered with research outcomes yielding a good proportion of inconsistent and contradictory outcomes. The conceptual definitions were vague, imprecise and there was an absence of empirically testable theoretical models available in the literature. Early research, and it is now thought wrongly, posited self-concept as a unitary construct, (Ellis et al., 2002). Development of instruments and investigations into self-concept construct validity suffered as researchers were unable to draw upon robust theoretical models or clear conceptual frameworks. Early studies were also hindered by a lack of empirically based conceptual models and measuring instruments that were psychometrically secure, (Shavelson et al., 1976; Wells and Marwell, 1976; Wylie, 1974). This, unsurprisingly, led to a confused message and prevented independent groups of researchers from establishing consistent, replicated outcomes and the generation of a robust theoretical base. The situation

underwent a step change in the mid 1970s when Shavelson called for a moratorium on research until a unified definition, based upon a sound theoretical framework, was established. Following a meta-analysis of influential studies, including five existing self-concept scales, Shavelson Hubner and Stanton (1976) finally established a seminal theoretical model.

These difficulties with early studies were compounded by the lack of appreciation of the importance of within network and between network construct validations. There was almost a total neglect for within network research (Byrne 1996a) and it was not until Shavelson et al.'s seminal work in 1976 that self-concept studies were categorised in these two different ways with due importance being given to the within network studies. This reconceptualisation laid the foundations for much of the structurally and theoretically robust studies that have emerged since then.

Defining Self-Concept: Conceptual Exploration

Differentiating Terms

In defining a concept, construct or measurement, it is often a useful process to undertake an exploration of what it isn't as well as what it is. This helps to set the boundaries and limits as well as testing the conceptual framework and refining the language. Defining self-concept is a good example of this. Self-concept, by its very name, is a construct that seems to give a sense of self-definition or self-explanation. It is not unreasonable for individuals, including those from the research community, to form an intuitive feel of the construct's elements, importance and association with other concepts, constructs or measurements. In these instances it can be an extremely useful process to place a construct, like self-concept, directly against related and sometimes overlapping constructs, comparing them, and teasing out the commonalities and distinctions. To this end the following section sets out to explore and define the nature of self-concept by making comparisons against such constructs and others which are sometimes mistakenly substituted for in name.

Terms such as self-esteem, self-perception, self-worth, self efficacy and self-concept have been used interchangeably in the literature for some time (Harter,

1999; Shavelson et al., 1976), and in fact still continue to be used inappropriately today. These self-terms are now known to represent quite different constructs.

Self-concept is recognised as being a global term which reflects an individual's belief about self. It is a hypothetical construct and helps explain and predict an individual's behaviour. This is so because it is thought that an individual's actions influence their self-perceptions, and perceptions of self, in turn, influence their actions (Shavelson et al., 1976). Self-concept is thought to incorporate both a cognitive and affective response to self, being heavily influenced by social comparison and contain both descriptive and evaluative elements (Byrne, 1996a). One definition puts it as the, 'self-perceptions formed through experience with the environment and, in particular, through environmental reinforcement and the reflected appraisals of others' (Shavelson, Hubner and Stanton, 1976).

Self-esteem, by contrast, is thought by some, (Byrne, 1996a; DuBois, Felner, Brand, Phillips and Lease, 1996; Watkins and Dhawan, 1989), to encompass evaluations of the descriptive aspects of self-concept, although Marsh and Hattie (1996) have found no empirical evidence to support this distinction and Marsh and Craven (2005) instead use the term self-esteem to refer to the global component of self-concept, i.e. measurement at a more general level of specificity. Others have defined it in yet a different way, as the 'global judgements of self-worth' (Crocker and Wolfe, 2001, p.594).

Self-efficacy relates to cognitive judgements of an individual's capability based on mastery criteria (Bong and Clark, 1999). Self-efficacy is context dependent, operates generally at a higher level of specificity when compared with self-esteem or self-concept and, unlike self-concept, does not include beliefs of self worth connected to perceptions of competence, (Byrne, 1996a). Self-efficacy has been defined by Bandura (1986) as representing a person's 'judgement of their capabilities to organise and execute courses of action to attain designated types of performances' (p.391).

Self-worth is thought to be less related to an individual's ability to be successful at specific tasks (c.f. self-efficacy) but more related to an individual's value judgement incorporating feelings of themselves. Crocker and Wolfe (2001) however, use the

terms overall self-worth and global self-esteem synonymously, Harter (1998) does likewise.

Distinguishing, at the theoretical level, between self-concept, self-esteem and self-efficacy has historically been problematic and in some ways continues to be problematic. The dust does seem to be settling today and there is a wider acceptance of definitions and meaning, and a growing consensus on distinctions and differences. Some researchers are content to deal with broad definitions which blur the edges and where the precision of definition is less important than causal influence, or where effect and consequence is more important than the nomenclature. For example, Watkins, Fleming and Alfon (1989) measure self-concept by using an instrument which they describe as comprising of self-esteem items. On the other hand there are those whose core research foundation is based on the distinctiveness and exactness of characterization, and for them, the research finding would be worthless without it (e.g. Marsh, Walker and Debus, 1991).

If distinguishing at the theoretical level has been difficult then making distinctions at the empirical and methodological level is a much greater challenge, and indeed has met with more limited success (for a review see Byrne ,1996). Part of the difficulty has arisen in differentiating between the outcomes of the different constructs. Correlations between measures of self-concept, self-esteem and self-efficacy constructs have been unhelpfully high (Marsh and Ayotte, 2003; Pajares, 1996).

Methodological difficulties also arise from the cross-over of one construct to another which may be caused by the particular focus of different research investigations. For instance, it is not uncommon for self-concept research to focus on feelings and evaluations of self, which have a strong connection to self-esteem measures, or alternatively, to focus on perceptions of capability which has a strong connection to self efficacy research (Byrne, 1996a; Pajares 1996). Additional methodological difficulties arise from the preferred data collecting instruments. Byrne (1996a) hypothesises that collecting data via self-reporting questionnaires necessarily taps both descriptive and evaluative components further adding to the difficulties of delineation.

In the case of self-concept and self-esteem, distinctions are usually, but not exclusively along the lines that self-concept refers to the descriptive component of self whilst self-esteem refers to the evaluative component of self (Byrne, 1996a). Hattie (1992) delineates the two by distinguishing between perceived *importance* and perceived *capability.* He asserts that the *importance* attributed to a domain or the amount of self-worth the individual invests in that area is linked with self-esteem, whereas the way in which an individual evaluates his or her *capability* and appraises their performance is linked with self-concept.

Self esteem measures usually arise from unidimensional scales which collect information on self confidence or self competence, for example the Rosenberg Self-Esteem Scale (SES: Rosenberg 1989). These measures are not typically related to any specific content area and are superordinate to specific content. Self-esteem measures are nearly always conceived and measured at this level of generality. Two issues emerge from this. Firstly, there seems to be scarce support for the distinction between the descriptive and the evaluative elements and the division between the two constructs with no theoretical reason why both constructs should not contain evaluative components (Marsh and Hattie, 1996). This view is supported by Bracken (1996) who asserts that both descriptive and evaluative aspects are incorporated into self-concept and that that self-esteem is part of self concept. Secondly, that the distinction is more likely to be one of unidimensionality verses multidimensionality with Marsh and Craven (2005) proposing that self-esteem measures and global self-concept measures should be treated identically. As evidence of this Marsh cites research which shows that self-esteem correlates approximately 0.95 with the highest order factor of multidimensional self-concept responses (Marsh and Hattie, 1996).

Distinctions between self-concept and self-efficacy are probably even more contested than those which try to distinguish between self-concept and self-esteem.

The literature is not always consistent on the matter, but there does seem to be commonality of distinction between those researchers who see the need to distinguish between the two The fact that separating these two constructs remains a current and continuing area of study, at the conceptual level, is illustrated by Bong and Skaalvik's (2003) recent and comprehensive analysis which explored

differences, commonalities and relationships between the two. In their article they explain that both self-concept and self-efficacy explain and predict an individual's thoughts, emotion and action, but that their axis of operation is different. In self-concept measurement, for instance, individuals are routinely asked to evaluate their skills and abilities which relates to a general perception of self within the domain of function. For example, *I am good at physics* or *physics calculations are easy for me.* In contrast self-efficacy judgements, rather than being an appraisal of skill or ability, relate to the level of competence and conviction that a specific outcome can be achieved, Bandura (1986). For example, *I can achieve at least a Grade B in my next physics test.* It is a personal judgement of the likelihood of success in specific circumstances rather than a judgement of general competence in the area.

This distinction was made in a different manner by Pajares and Schunk (2002) who gave us a conceptual framework to compare the two constructs. They posit that self-concept beliefs and self-efficacy beliefs form and emerge through asking different questions. Self concept beliefs arise through asking questions about 'feeling' and 'being'. For instance, significant questions might be, *Am I good at science? How do I feel about myself as a learner of science?* By contrast self-efficacy beliefs develop by asking 'can' questions. For example, *Can I complete this science task successfully?*

At the conceptual level at least five differences have at some time been used to distinguish between the self-concept and self-efficacy. The first distinction is at the level of specificity. Self-efficacy is considered to be measured at a higher level of specificity than self-concept. (Finney and Schraw, 2003; Pajares and Miller 1994). This is generally true, although Zimmerman and Bandura (1992) conducted self-efficacy measures at the subject specific level, including science, when investigating students' general performance in academic subjects. More recently self-concept has also been measured at the same level of specificity, (e.g. Haussler, and Hoffmann, 2002). In fact many recent studies have matched the two at exactly the same level in a direct comparison, (e.g. Choi, 2005; Pietsch, Walker and Chapman, 2003). The second distinction relates to the contribution of cognitive and affective responses. Both constructs incorporate a cognitive appraisal of self, but only self-concept incorporates an affective response to self (Bong and Clark, 1999; Pajares and Schunk, 2002).

The third distinction relates to the role that social comparison plays in forming the self beliefs (Bong and Clark, 1999). Self-concept formation is conceptualised as being heavily influenced by the process of social comparison. This entails the normative referenced process whereby an individual will make an evaluation of self partly by comparison with others, and also through the reflected appraisals of others, (Marsh, 1986). By contrast, self-efficacy formation is influenced by evaluation of self in relation to past performance; hence being a criterion referenced process rather than a norm referenced process (Bong and Skaalvik, 2003). Bandura (1997) however, has identified the impact that vicarious experiences can have on self-efficacy development, and freely acknowledges the impact that social comparison can make when forming performance capabilities, indeed Bandura (1986) states:

> Students judge how well they might do in a chemistry course from knowing how peers, who performed comparably to them in physics, fared in chemistry (p. 404).

He does however, contest elsewhere that prior mastery experience is by far the greatest sources of efficacy information, (Bandura 1997). The differences over social comparison are probably best observed at the level of measurement. Self-concept has for some time contained specific items which relate to the influence of social comparison. Within the ASDQI instrument for instance, Marsh (1999), items like 'compared with others of my age I am good at science' and 'In science I am one of the best students in my class' appear, where there are no such comparable items in self efficacy measurement with efficacy being gauged mainly against concrete performance standards (Bong and Clark, 1999; Zimmerman, 1996).

The fourth distinction is in the differential ability of self-concept and self-efficacy to predict future academic achievement. This superiority of each of the two constructs to predict subsequent academic performance relative to the other has been the matter of not small amounts of professional rivalry between the different camps. What has seemed to emerge recently, however, is that it is not which construct per se, is used, but at what level of specificity. Research data have consistently indicated that the greater the level of specificity relative to the performance outcome then the greater is the predictive utility of the instrument (Parjares and Miller, 1994). Expressed in an alternative way, when measured at the same level of generality there is no significant difference between the

constructs of self-concept and self-efficacy for predicting academic performance (Marsh et al., 2004).

There is also a difference between how the two constructs relate to past and future events and the consequences of this. Self-concept is past oriented and tends to be stable. Self-efficacy is future oriented and tends to be relatively more malleable (Bong and Skaalvik, 2003). Most self-concept items within self-completion measuring instruments tend to be reflective in nature, asking the individual to assess past performance or evaluate self against past experiences. For example, *I have always done well in ...,* or *I often need help with ...* (Pietsch et al., 2003). This is contrasted against self-efficacy items which, as well as being task oriented, are also concerned with performance in those tasks in the future. For example, *how well can you ...?*, or *I am confident that I will be able to ...*, (Pajares 1996).

Despite the fact that self-concept and self-efficacy point in different temporal directions they are both in fact contingent of the individual making judgements and/or evaluations based on past events. Even self-efficacy judgements, according to Markus and Nurius (1986), which are inherently future-oriented are based upon mastery criteria from the past in order that the individual can make judgements about their confidence to successfully complete imminent tasks.

In many cases, if self-concept and self-efficacy measurement were to occur near simultaneously, individuals would be drawing upon the same pool of experiences in order to make their judgements. However, even though they draw upon the same data the result may well be different. For example, a high achieving individual with regular success (in science, say) and with test scores in the top quarter of the class might reflect on these experiences in different ways when making their evaluation for self-concept and self-efficacy. Likewise an individual with a poor test record (in science) and a history of being poorly achieving in class would draw upon those experiences. Remembering that self-concept is related to an individual's evaluation of general competence, whilst self-efficacy is related to competence for successfully accomplishing future tasks, an interesting outcome could well emerge. The highly achieving individual may not judge themselves (or may not feel) that they are a good science student and therefore present a low science self-concept. This may well occur even though their presentation of self-efficacy is high owing to a strong self belief that future performance could be just as successful as past

performance. A student might well report about themselves 'I'm not good at science but I can usually get through the exams.'

The last words of the complexities of two competing or complementary constructs belong to Bong and Skaalvik (2003):

> Researchers express little disagreement as regards the purported differences between task-specific academic self-efficacy and subject-specific academic self-concept (e.g., Marsh et al., 1991; Pajares, 1996). However, when the two constructs are put side by side at the same level of measurement specificity, the opposing arguments collide. Academic self-efficacy researchers express pessimistic views that self-concept can ever be assessed at task-specific or problem-specific levels (Bong and Clark, 1999; Pajares, 1996). Academic self-concept researchers, on the other hand, question the practical utility of self-efficacy judgments beyond what they view as microlevel analyses of performance. The problem worsens because both self-concept and self-efficacy theories contend that their construct can be assessed at varying levels of measurement specificity (Bandura, 1997; Shavelson et al., 1976) (p17).

Two Views of Self-Concept

Over the course of the last several decades researchers have addressed the conceptual and methodological problems inherent in the early self-concept research much more explicitly. Recent self-concept research has proceeded within two distinctly defined research traditions. One tradition was set within a cognitive or social cognitive information processing perspective (e.g. Markus 1977; Markus and Wurf, 1987), and the other within the instrument and/or construct validation framework (Marsh, Relich and Smith, 1983; Marsh and Shavelson, 1985). A difference in the way self-concept was conceptualised has created two contrasting empirical perspectives forming quite different methodological strategies at the level of research design, instrumentation and data analysis. This has resulted in the emergence of two distinct bodies of literature with little cross referencing or opportunities for synergy.

The research described within this thesis is motivated by the prospect of developing an understanding of the structural organisation of self-concept and as such is positioned within and influenced by the theoretical and empirical framework of construct validation research. All further discussion therefore, will take place

within this perspective and self-concept will be thought of as being a theoretical model that explicates the relationships among the constructs embodying the particular theory and centres the investigations on measures based on that model.

Defining Self-Concept: Structural Exploration

Unidimensional

The earliest ways in which self-concept was conceptualised was as a unidimensional, general self-concept construct. Coopersmith (1959; 1967) contended that a general factor of self-concept so overwhelming dominated any other contributing specific factor that it was only possible to conclude that self-concept was a unitary construct. Coopersmith made a significant contribution to the research literature of the time with the 'Coopersmith Self-Esteem Inventory' being one of the most widely used self-concept tests (Hattie, 1992). This unidimensional model persisted for nearly three decades despite a growing weight of evidence that Coppersmith's research was fatally flawed in both a theoretical and empirical sense. Marsh and Hattie (1996) could find absolutely no support for the idea that a general factor of self-concept had this overwhelming strength and believed that Coppersmith's work was fraught with measurement and statistical error, (Marsh and Hattie 1996). Byrne (1996a) has commented at her surprise that the Coopersmith model lasted so long and Stein (1993) tried to rationalise it by suggesting that psychologists were attracted by its ease of use due to the fact that interventions could be at a single level rather than at the more complex multiple levels needing an assortment of methods.

Marsh and Hattie (1996) wrote unambiguously:

> In conclusion, there appears to be no support at all for a unidimensional perspective of self-concept or, apparently, even a unidimensional perspective of academic self-concept. Critical evaluations of previous research claiming support for the unidimensionality of self-concept suggest that these claims were apparently unwarranted (p. 44).

A distinction must be made about categories of self-concept structure validation research. A primary dichotomous division relates to differences between 'within network' and 'between network' research. The valid use of a construct, according to measurement theory, requires that both within network and between network characteristics are investigated. A process referred to by Cronbach et al. (1955) as validating a construct's nomological network. This investigative and delineation process involves locating a construct within a theoretical framework and examining the between network and across network characteristics. The within network validation relates to identifying and exploring the internal components or dimensions and identifying their characteristics and relationship between each other. Such investigations might, for instance, examine the relationships between multidimensional facets of self concept structure, explore correlational measures between global self-concept and general academic self-concept, or explore the correlation between general academic self-concept and self-concept in science. By contrast, between network validation centres on relationships between one aspect of self-concept measure and a potentially associated external construct. That is, developing its predictive validity. Such theoretically related constructs to self-concept might include school achievement, truancy levels or IQ scores.

In conducting self-concept research, meaningful outcomes in cross-network research is contingent on successful and appropriate outcomes in the between network research. It is therefore important that research relevant to both the theory and the construct measurement of the within network research is robustly and validly undertaken as a prerequisite to the between network research.

Moving towards Multidimensionalism

Very early ventures into self-concept thinking were undertaken by James (1890/1963). His seminal work had a rich theoretical and philosophical basis (Marsh and Craven, 2005) which remained largely ignored for seventy years. James was the first psychologist to propose that self-concept structure would be multidimensional in nature. His four divisions of self or 'classes' consisted *material-self*, *social-self*, *spiritual-self* and the *pure ego*. This conception of self, which was ahead of its time, can now be seen reflected in many of the self-inventory scales.

Today, very few psychologists would contend that self-concept is anything but multidimensional[2]; however the pathway taken in reaching agreement and acceptance of common principles and features has been long and winding with the role, nature and importance of a unidimensional element still the subject of some debate today. James (1963) laid the groundwork to this idea with his search for the core of self. This he called the 'sum total' referring to every influence that helped to make up the person. This directed many researchers (including Coopersmith, 1967, as previously discussed) to search for the elements that made up the 'sum total'.

Rosenberg (1979) later picked up on the idea by acknowledging that self-concept comprised both the *parts* and the *whole*. He reflected on the idea that self-concept consisted of a *global* form and separate *facets* of self-concept which, he asserted, should constitute legitimate focus points for future research activity. He never sought to explain the relationship between the 'components' to the 'global' but he did propose three broad domains under which they could be collected. These were (i) the extant self (actual self), (ii) the desired self (ideal self) and (iii) the presenting self (displayed self). Rosenberg (1979) proposed that self-concept was influenced by five principles: i) reflected appraisal, the influence of others' opinions on self, ii) social comparison processes, others as a standard by which to be judged, iii) self-attribution, decisions made after reflecting on internal states, iv) psychological centrality, self-concept as a 'whole' not as the sum of its parts, v) domain importance, comparison of perceived importance of different school subjects.

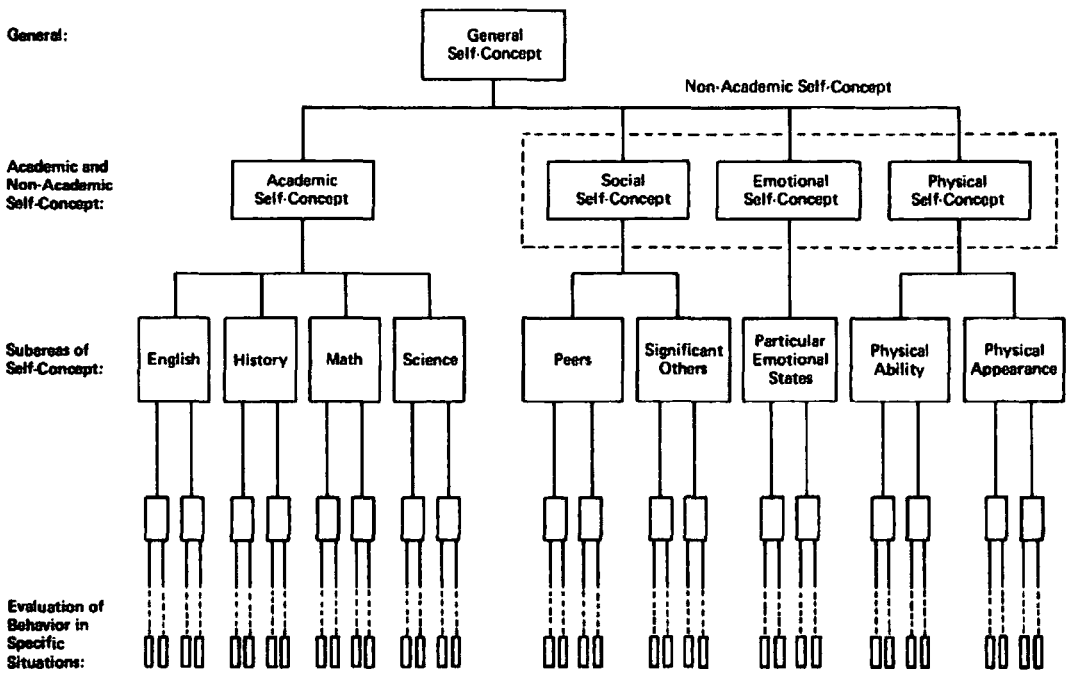
This work led to the Rosenberg Self-Esteem Scale (SES: Rosenberg 1989) which has been widely adopted, is still used today, and has been freely available since his death in 1992. The scale tends to focus on the global aspect of self-concept rather than specific facets or components of self-concept, and as such has fallen foul of criticism notably because of its neglect of differentiating the multidimensional aspects of its construct (see Marsh, 1990a).

[2] The concerns of Baumeister, Campbell, Krueger and Vohs (2003) will be discussed later in the chapter.

The Shavelson Model

Discontent with the confused state of self-concept research in the mid 1970s, Shavelson called for a moratorium on research until a unified definition based upon a sound theoretical framework was established. Shavelson, Huber and Stanton (1976) undertook a meta-analysis of influential studies including the five self-concept measures of the Michigan State Self-Concept of Ability Scale, the How I see Myself Scale, Piers-Harris Children's Self-Concept Scale, Self-Concept Inventory and Self-Esteem Inventory which finally established a seminal theoretical model of self-concept. The proposed new model characterised self-concept structure as being organised, multifaceted, hierarchical, stable, developmental and differentiable.

Diagram 2.1 Multidimensional and Hierarchical Structure of Self-Concept



Shavelson, Hubner and Stanton (1976). Multidimensional and Hierarchical Structure of Self-Concept

The multidimensional and hierarchical structural nature set the tone and frames of reference for much of the work which superseded it. In fact the theoretical structural model has withstood substantial subsequent investigation, has

dominated self-concept theory since its introduction and has come to be known as the 'Shavelson Model' (Byrne and Shavelson, 1996).

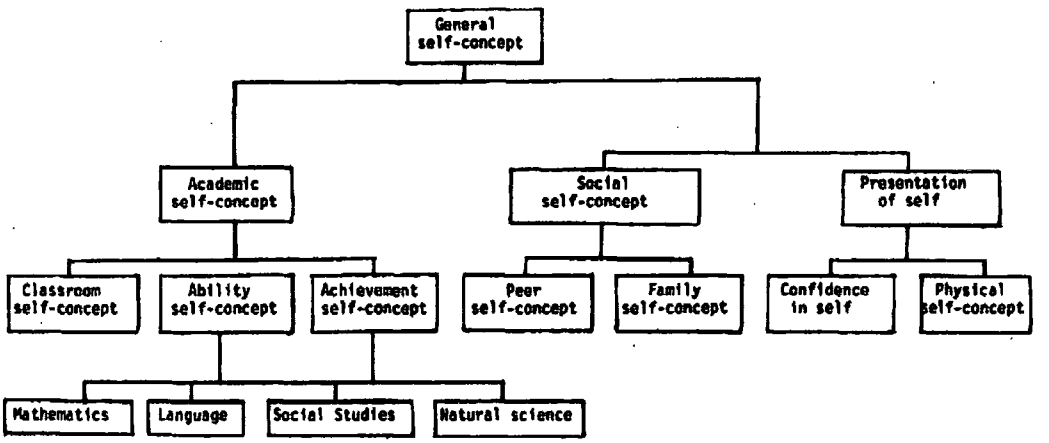
The hierarchical nature places the most general perceptions of self as a person, i.e. global self-concept, at the apex of a pyramid. A descent through the pyramidal structure sees different elements becoming ever more specific. Beneath the apex at the second strata of the hierarchy sees global self-concept being separated into the two facets of 'academic' and 'non-academic' self-concepts. The non-academic side subsequently is further separated into 'social, 'emotional' and 'physical' self-concepts (Byrne 1996a; Byrne and Shavelson, 1986; Shavelson and Bolus, 1982). The academic dimension of self-concept is itself sub-divided into more specific and distinct academic components and includes English, mathematics and science. Further levels of sub-division were hypothesised for each of these specific self-concepts so that at the lowest level of the pyramid are actual observed behaviours. Between, and across the various levels of the hierarchical structure can be found horizontal and vertical correlational links. The nature and strength of these correlational links were subsequently to give clues which guided developments and refinements of the model, notably, from Marsh and colleagues (Marsh and Hocevar, 1985; Marsh and Shavelson,, 1985; Shavelson and Marsh, 1986). This ultimately, brought research into self-concept to the position we enjoy today. Thirty years has passed since this model was first proposed and since then it has seen a number of important modifications and revisions.

Refinements to and Adaptations of the Shavelson Model

Shavelson et al.'s model of self-concept has undergone numerous adaptations over the past three decades, but still serves to provide the core basic principle on which self-concept structural research is based. It was revisited by Song and Hattie (1984) who tested four slightly different models. Their best fit model was both multidimensional and hierarchical in nature although their higher order facets differed from those proposed by Shavelson which were widely accepted by researchers. They made a number of changes to the non-academic side of the structure notable combining Shavelson's 'Emotional' and 'Physical' into a new facet called 'Self regard / presentation'. The other half of the structure, which is of more interest to us here, 'Academic' remained as a discrete first order facet although

instead of dividing into individual school subjects as Shavelson had proposed, Song and Hattie found favour with 'Achievement self-concept ' (confidence in achievement at a particular point in time), 'Ability self-concept' (confidence in capability to achieve) and 'Classroom self-concept' (confidence in classroom abilities). This was eventually formed into a new self-concept measure called the Song and Hattie Test (Hattie, 1992) and was determined by means of a 35 item self-completion questionnaire, using a Likert type scale with each of the second-order facets, including Ability, Achievement and Classroom being measured by five items each.

Diagram 2.2 Song and Hattie Test of Self-Concept



Song and Hattie Test of Self-Concept (Hattie, 1992)

Hattie (1992) undertook additional revisions when further research seemed to indicate the existence of stronger relations between Classroom self-concept and Social self-concept than there were between Classroom self-concept and academic self-concept. He also introduced the fourth order facets of Maths, Language, Science and Social which related to the first order academic facets of Ability and Achievement. This brought it back closer to the original Shavelson model although a distinction remained as Hattie continued to differentiate between Ability self-concept and Achievement self-concept. This model has found support from some researchers, for example Waugh (2001) but generally has received less empirical support than some of the other models considered below.

Harter (1982) conceived her own self-concept scale based upon the original Shavelson design. Critical of the Coppersmith Self-Esteem Inventory (1967) and the Piers Harris Self-Concept Scale (1964) for summing heterogeneous items and calculating a total score which was interpreted as an index of Global Self-Regard, she developed her own multidimensional scale which unlike Coopersmith assumed that, in forming their self-perception of competence, children make distinctions between different domains of experience in their lives. Harter found a very clear and stable factor structure which revealed that children as young as 8 years of age can make meaningful distinctions between her four proposed domains of (i) cognitive competence in school, i.e. being smart and feeling good about own performance (ii) social competence with peers, i.e. possessing many friends and having good status among friends, (iii) physical competence in sports, i.e. being good at sport and enjoying sport participation, and (iv) general self worth, i.e. being sure, being happy feeling good and feeling you're a good person. The measurement data showed:

> dramatic subscale differences, (which) suggest that those instruments (e.g. the Coopersmith Self-Esteem Inventory) which yield a single score are masking important distinctions which children can make about their competence in different domains (p.95).

She fashioned this research outcome into a new scale called Perceived Competence Scale for Children (PCSC: Harter 1982), targeted at children in the age range 8 – 12 years of age.

Based on the PCSC as parent instrument, Harter (1985) later introduced an alternative scale, still consisting of the four sub-scales, called Self-Perception Profile for Children (SPPC). The SPPC contains additional scales measuring physical appearance and behavioural conduct self-concept, with the PCSC physical scale becoming a physical ability scale.

The factorial validity of the scale has been supported by a number of independent researchers. Exploratory factor analysis of the PCSC scale has identified social, physical, academic and general self-concept facets with children in US Grades 3 – 9 (Harter, 1992) and with Australian children in Grades 7 – 9 (Marsh and Gouvernet, 1989). Further exploratory and confirmatory factor analysis of the PCSC scale of Australian children covering Grades 5 -9 (Marsh and McDonald,

1990; Marsh, 1990b), with Canadian children within Grades 5 – 8 (Byrne and Schneider, 1988) and with Dutch children (Van Dongen-Melman et al., 1993), using a translated instrument, identified four facets.

The correlation coefficients measured among the different self-concept facets were not so high as to show that the scales are indistinguishable and therefore non-differentiable (Harter, 1985). For example correlation coefficients ranged from 0.08 to 0.62 with Grade 5 and 6 children from the US, which therefore supports the PCSC and SPPC instruments' differentiability of dimensions of self-concept.

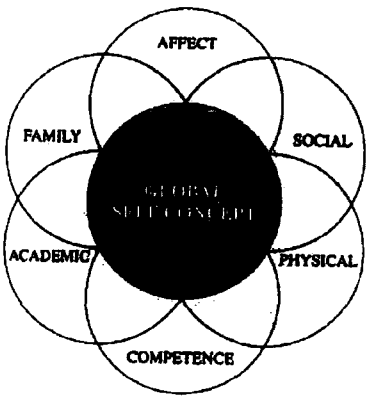
These and other studies have shown both the SPPC and the PCSC scales to perform well in multiple samples with much cross-cultural validity testing (e.g. Eapen and Abbas, 2000; Smith and Mao 1985; Alva and de Los Reyes, 1999). This SPPC scale is still a popular instrument with researchers and clinicians alike and is still regularly used today (e.g. see Fetsch and Yang, 2002).

Harter (1990) asserts that individual conceptions of self-concept change as chronological age increases with self-concept changing from concrete conceptions to abstract conceptions. In keeping with this, her age related scales also increase in complexity beginning with a pictorial version for preschool children (Harter and Pike 1984) administered orally indicating a two factor structure, through the SPPC scale to a specific instrument designed for adolescents (Harter, 1988).

Bracken (1992) designed an alternative Multidimensional Self-Concept Scale (MSCS) which demonstrated a rather unique circular symmetry. Bracken reported six sub-components that overlap and all of which contribute equally to a seventh component of Global Self-concept. This characterisation of self-concept consisted of the same seven features hypothesised by the Shavelson self-concept scale although in this model they are more explicitly delineated and the formation and development of self-concept is more strongly linked with behavioural measures (Crane and Bracken, 1994; Keith and Bracken, 1996). The self-concept instrument is intended for use with young people aged 9 years to 19 years with the six areas of self-concept being (i) social, (ii) a more specific family, (iii) academic, (iv) physical, (v) affective, (vi) competence in relation to attainment goals (Bracken et al. 2000).

The MSCS scale was developed using a large representative number of young people in the US. The sample drew from 2501 pupils from Grades 5 to 12, equivalent to UK Y6 to Y13 to test, norm and validate the scale (Bracken 1992). Bracken offered strong empirical support for the factorial, concurrent, convergent and discriminant validity of the scale, based on the original sample, and has since provided further independent data to add further support (Crane and Bracken, 1994).

Diagram 2.3 The Bracken Multidimensional Self-Concept Scale (MSCS)



The Bracken (1992) Multidimensional Self-Concept Scale (MSCS)

Other scale test measures are also high; test retest procedures over a four week period were shown to be in the range of 0.73 to 0.90 and internal reliability estimates were reported to be in excess of 0.80 and sometimes 0.90 (see Crane and Bracken 1994; Keith and Bracken 1996).

The global component is reported to have greater stability than the sub-components with this global component representing the individuals generalised response pattern across multiple environments.

Refinements by Marsh and Associates

Herbert Marsh aided by numerous other self-concept researchers set about testing the multidimensional and hierarchical construct validity of the Shavelson Model, (e.g. Marsh, 1986; Marsh, 1990; Marsh, Byrne and Shavelson, 1988).

In pursuit of this, the Self-Description Questionnaire-I (SDQ-I, Marsh, 1992a) was designed to measure in pre-adolescent children (primary aged children) the multiple dimensions of the self-concept facets of Shavelson's model. It incorporated a total of eight scales; a general self-concept scale, four non-academic scales and three academic scales. The three academic scales measured verbal, mathematics and general academic self-concepts. Three other parallel scales were also developed to test the Shavelson model with older aged pupils. SDQ-II was developed for adolescent high school students (Marsh, 1992b) and the SDQ-III (Marsh 1992c) for late adolescents and young adults. The SDQ instruments were developed through a combination of theoretical and empirical research (Byrne 1996a) and provided particularly strong tests for the Shavelson model (Marsh and Craven, 1997). Central to the SDQ scales was that self-concept could be shown empirically to have a structure which was multidimensional and hierarchical in nature.

Structural Nature of Self-Concept

According to Marsh, Byrne and Shavelson (1988) construct validation requires a series of correlational relationships. The hierarchical nature of self-concept which was perceived to run in a vertical direction down through a pyramid structure needed to produce a consistent pattern of relationships amongst the different self-concept facets at different levels in the hierarchy. Correlations would be strongest when the levels are directly subordinate / super-ordinate to each other and correlations would be less strong between facets related through an intermediate facet of self-concept. Thus a correlation between the apex and an element at the second level, (e.g. global self-concept and academic self-concept) would be stronger than correlations between the top and the third level, (e.g. general self-concept and a subject specific self-concept like science). At the lower end of the hierarchy, a correlation between observed behaviours, like achievement, and a subject specific self-concept would be greater than that between observed

behaviours and academic self concept. The further the levels are apart then the smaller would be the correlations between them. Correlational research which has focussed relations toward the base of the pyramid has shown a meaningful link between academic self-concept measures in an individual subject and measurable classroom outcomes, such as test results.

Marsh and colleagues completed a series of studies in which they examined the factor structure and the convergent and discriminant validity of the SDQ-I as well as the construct validity of the multidimensional self-concept per se. The original research was mostly undertaken in Australia using a number of schools and pupils from a range of ages and social and economic backgrounds, (e.g. Marsh, 1985; 1992a). Later, validation exercises were also been carried out in numerous other countries, for example, in Britain (Smith and Marsh, 1985), in Canada (Byrne and Worth Gavin, 1996), in the Philippines (Watkins and Gutierrez, 1989) and Nigeria (Watkins and Akande, 1992). Exploratory and confirmatory factor analysis supported the SDQ-I scales and, taken as a whole, the studies provided very strong support for the SDQ-I instrument as a measure of pre-adolescent multidimensional self-concept. It was consistent with the Shavelson Model and provided confirmation of the multidimensional, hierarchical, and developmental nature of self-concept structure.

The Shavelson model through the SDQ instruments has been consistently shown to be a robust multidimensional model. Indeed the clutch of the three SDQ instruments have been extensively trialled, tested and evaluated, with researchers finding them to be amongst some of the best self-concept scales in terms of their psychometric properties and construct validation (Byrne, 1996a; Hattie, 1992; 1996). The SDQ instruments are amongst the most widely used and have been translated into numerous languages. The instruments have extended the capability to measure self-concept and as such have provided significantly greater opportunity to extend the theoretical and empirical knowledge of the structure and nature of self-concept.

Research with the SDQ instruments has also supported the differentiability of dimensions of self-concept. Like studies with the PCSC and SPPC studies, the correlation coefficients between scales of the SDQ instruments have not been so

high as to negate claims of indistinguishability, and therefore the SDQ instruments support the differentiability of multidimensional self-concept.

Although analysis of data from the SDQ-I instrument provided strong support for the multidimensional nature of self-concept the analysis also uncovered a serious anomaly in the results. The Shavelson model conceptualised Mathematics and Verbal self-concepts in terms of a single higher-order academic self-concept. This was in part due to the fact that they were perceived as substantially correlated. Whereas Marsh had found that mathematics and verbal *achievement* exhibited a strong correlation, there seemed paradoxically, to be a large and consistent body of research indicating that correlations between Verbal self-concept and Mathematics *self-concept* were substantially lower than expected and consistently lower than that between general academic self-concept and either Verbal or Mathematics self-concept. In fact in some studies the correlations between Verbal and Mathematics self-concept were near zero, making it indefensible to advocate that they be incorporated into a general academic self-concept as originally proposed. For example, correlations between Verbal and Mathematics self-concept ranged from -0.06 to 0.12 for Grades 5 – 6, compared with correlations between 0.25 and 0.56 between general academic self-concept and either Verbal or Mathematics self-concept (Byrne & Shavelson, 1986; Marsh and Hocevar, 1985; Marsh, Relich and Smith, 1983). These unexpected correlational findings led Marsh and Shavelson (1985) to conclude that it was not possible to combine the Mathematics and Verbal self-concepts to form a single higher-order academic self-concept and that the Shavelson Model needed to be adjusted to form a revised model. Other research results had also revealed unexpected correlational outcomes. These results related largely to the non-academic side of the pyramid structure and are of less interest to us here. None-the-less to explain the apparent contradictions a revision to the theoretical model was undertaken which has since become known as the Marsh/Shavelson model (Marsh and Shavelson, 1985). Of all the revision elements the separation of a single facet of academic self-concept to Verbal academic self-concept and Mathematics academic self-concept is probably the most significant revision imposed on the Shavelson Model.

The Internal – External Revision

In an attempt to explain the anomalous result relating to the unusually low correlation between Verbal and Mathematics self-concepts, Marsh (1986) proposed the Internal/External Frame of Reference Model (I/E Model).

According to the I/E model, self-concept formation is more complex than merely the relationship between previous performance and experience in a particular school subject and subsequent self-concept formation in that subject. The I/E model proposed that there were both internal and external comparisons contributing to the self-concept construction.

Marsh (1986) posited that when learners form their specific self-concepts in particular academic areas they engage in the external process of comparing self perceptions of their abilities against those of other learners around them in a normative comparison, *how do I perform in this subject compared with others in my class?* Therefore self-concept in a particular subject area is influenced by the individual's perception of (i) their ability in that particular subject together with, (ii) a comparison against other individuals' abilities in that subject included in their frame of reference. This external comparison process should, according to Marsh, have a strengthening effect on the correlation between Verbal and Mathematics self-concepts. The explanation being that Verbal and Mathematics achievement are highly positively correlated and therefore not only are self-concepts of both subjects being partly based on the same achievement index but comparisons against other individuals are set to reinforce that positive relationship.

The internal process occurs when individuals engage in an additional internal process of comparing their perception of ability in one subject e.g. mathematics with their perception of ability in another subject, e.g. Verbal. This is an ipsative comparison, *how do I perform in this subject compared to other subjects I study?* The difference between these two perceived abilities contributes to a greater self-concept in one subject to the detriment of the other subject, lowering the correlation coefficients between the self-concepts in those subjects. The greater an individual's perceived difference between their Verbal and Mathematics' abilities the greater is the negative correlation; meaning that a higher self-concept in one area was likely to result in a lower self-concept in the other area. In other words,

there is a direct negative interaction of Mathematics achievement on Verbal self-concept and of Verbal achievement on Mathematics self-concept. This leads to a negatively correlated relationship between Verbal and Mathematics self-concepts. The bottom line on this is that Mathematics self-concept is likely to be highest when achievement is mathematics is high (external) and when mathematics achievement is greater than Verbal achievement (internal).

The operation of the internal and external effects together allowed Marsh to explain the near zero correlations between the Verbal and the Mathematics self concepts; one effect moves to strengthen the correlation positively whilst the other effect serves to strengthen the correlation negatively. The I/E model has received much testing and validation (see for instance Marsh, 1986), and there is substantial empirical support for the I/E model from children aged 7 years to mature adults. Marsh (1986) reviewed a number of SDQ studies involving preadolescent, adolescent, university students and adult samples which provided strong support for the model. There is also strong evidence that the effect has cross-cultural validity in English speaking countries (see Marsh and Hau, 2004) as supporting research has been reported from Canada (Byrne and Worth Gavin, 1986), from the US (Marsh 1989), and from Australia (Marsh, 1986).

Marsh has taken the external comparison process one step further and introduced what has come to be called Big-Fish-Little-Pond-Effect (BFLPE). Primarily, the BFLPE (Marsh, 1987; Marsh and Parker 1984; Marsh et al., 2000) predicts that equally achieving students will have lower academic self-concepts when comparing themselves against highly achieving classmates and conversely, will have high academic self-concepts when comparing themselves against lowly achieving classmates. This means that highly achieving students will experience a reduction in their academic self-concepts whilst attending schools where the average achievement levels of their classmates is high, and conversely, experience an increase in their academic self-concepts whilst attending schools where the school-average attainment is low. The BFLPE seems to be an effect exclusively relating only to the academic components of self-concept. It does not appear to show itself strongly, if at all, in relation to general self-concept, (Marsh, 1987), or self-esteem measures.

The BFLPE emerges from social comparison theory (Festinger, 1954) whereby, in the absence of unambiguous objective criteria, individuals use comparisons with others in their frame of reference to make subjective estimates of their ability levels and self-worth. According to Festinger, individuals need to make comparisons with others in order to define the *self* and then pass judgements on that definition. The choice of individual and comparison group is a key contributor to the social comparison theory outcomes. The relative strength of the reference group's attributes can therefore have a significant effect on the evaluation of self-worth. It is also thought that that when individuals are faced with choice of relatively similar and dissimilar others, the individual is likely to choose similar others as the basis of social comparison.

There is considerable cross-cultural support for the BFLPE. There are consistent reports of the effect from Australia (Marsh and Rowe, 1996), the US (Marsh 1991), Germany (Jerusalem, 1984; Marsh, Koller and Baumert, 2001), Israel (Zeidner and Schleyer, 1999) and Hong Kong (Marsh, Kong and Hau, 2000). Marsh and Hau (2003) tested the cross-cultural generalisability of the BFLPE with 4000 15 year olds from 26 countries and found very strong support for internal validity, external validity, generalisability, and policy-practice implications for the BFLPE.

Three issues of special note about BFLPE are worth singling out and describing in turn. The first issue concerns BFLPEs related to gifted and talented pupils in special programmes. Marsh, Chessor, Craven, and Roche (1995) reported two matched comparison group studies into the effects on different facets of self-concept of participation in gifted and talented programs over time. There was clear evidence for negative BFLPEs in so much that the academic self-concept of pupils in the gifted and talented programmes declined over time in relation to the comparison group. They also reported BFLPEs were consistently large for Mathematics, Verbal, and Academic self-concepts but were small and largely non-significant for general self-esteem and four non-academic self-concepts.

The second issue concerns BFLPEs for less able pupils. If less able pupils attend schools where the average ability of the pupils is low, i.e. either because they attend special schools, or foundation schools where schools selection by ability still occurs, or are assigned to classes set strictly by ability, then their academic self-concept has the tendency to rise (Schwarzer, 1982 reported in Ludtke et al., 2005).

This is not the case however, if these low achieving pupils are returned to classes where in comparison to others they are more lowly achieving. In these instances their academic self-concept decreases. Similar findings have been found by Tracey et al. (2003) with their work on learners with mild intellectual disability (IM students). They found that IM students possessed significantly higher academic self-concepts when enrolled in a full-time IM support unit and suffered lower academic self-concepts (and felt socially excluded) when attending 'regular' classrooms.

The third issue concerns the impact on pupils' academic self-concept of attending an academically selective school. There can sometimes be an uncritical presumption from policy makers, teachers and parents that highly achieving pupils who attend a selective school will be academically benefited. Marsh and Rowe (1996) found that all pupils, regardless of achievement level suffered a lowering of academic self-concept as a consequence of attending a selective school. Coleman and Fults (1985) by contrast found there was still a potentially negative effect although they could only perceive a significant effect on pupils in the lower half of the academically selective classes. Dai (2004) is unconvinced by much of the BFLPE consequences, particularly in relation to the negative impact of selective schools. He feels that social comparison theory is being interpreted too simplistically and that other effects may be at work. He reports that upward social comparison can be self-enhancing as well as self-deflating in some circumstances, and that some people can display a self-enhancement or self-protection bias which might nullify the effect of BFLPE. None-the-less there is a possibility that such selective schools create an unintentional negative effect. The question of choice (outside of ideological considerations) then becomes a consideration of the balance of educational opportunities and motivational capabilities.

By way of contrast to the negative BFLPEs of being placed in an environment where relative comparisons against highly achieving individuals can be deflating, comes the potentially positive outcome associated with membership of a highly achieving group. This Basking in Reflected Glory Effect (BRGE) (Trautwein, Koller, Ludtke, and Braumert, 2005) is a positive influence on self-concept which goes some way to counter the negative outcome of the BFLPE. Is it better, from a self-concept enhancement perspective, to be *third violin* in the Royal Philharmonic, for instance, or *first violin* with the local orchestra? How do the opposing effects of

BFLPE and BRGE match up? Ludtke et al. (2005) distinguished between the two components of the BFLPE labelling them *contrast* and *assimilation* effects. The term contrast was used for the negative outcome of the BFLPE and occurred when the self judgement is moved away from the target background or context. Conversely, the term assimilation is used to describe the positive outcome when the self-judgement is moved toward the target background or context. The *assimilation,* or *reflected glory* or *labelling* effect results from the individual benefiting merely by virtue of being chosen to join a more elite group, i.e. *If I'm good enough to be chosen to be part of this group, with these highly achieving people, then I too must highly achieving.* There is much less research evidence for the presence or strength of the reflected glory effect, and indeed the effect is still contested. However research recently completed (Trautwein, et. al., 2004; Trautwein et al., 2005; Ludtke et al., 2005,) has tentatively concluded that any reflected glory effect is much smaller than that compared with total BFLPE outcomes, that the effect is temporally shorter and that assimilation measures are incorporated into (and swamped by) contrast effects of the general BFLPE measures.

Marsh and Hau (2004) interpret the ramifications of the BFLPE as extremely far reaching and feel that, unfortunately, policy makers are either, unaware of the effect and its consequences, or misunderstand its potential to have significant impact citing research showing that:

> equally able students attending higher ability high schools were likely to select less demanding coursework and to have lower academic self-concepts, lower educational and occupational aspirations, and lower school grades (p. 269).

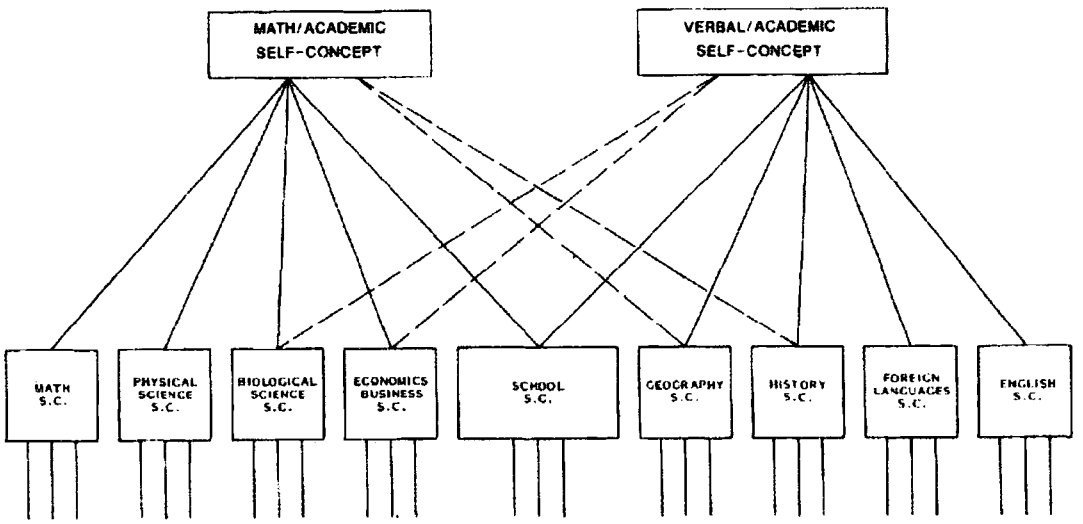
Ominously for Marsh, the BFLP effect appears to grow over time, be long lasting and influences the major educational outcomes.

Academic Self-Concept

Academic self-concept, as a facet of self-concept, has itself come to enjoy a considerable amount of research interest in its own right (Byrne, 1996a; Byrne, 1996b; Skaalvik and Rankin, 1990). Academic self-concept, like the general structure of self-concept, is also thought to be hierarchical and multidimensional in nature. One model places general academic self-concept at the apex, then differentiates the general academic self-concept into subject-specific academic self-concept facets, such as mathematics, verbal and science (Byrne and Shavelson, 1986, Shavelson and Bolus, 1982, Shavelson et al., 1976).

According to Marsh, Byrne and Shavelson (1988), construct validity of a multidimensional academic self-concept firstly, requires that academic achievement be more strongly correlated with academic components of self-concept than with general self-concept. Secondly, it requires that academic achievement in specific subjects be more strongly correlated to self-concept in the same subject than with different subjects. Marsh et al. (1988) set out to verify the framework with an array of self-concept instruments; the SDQ-III (Marsh, 1992c), the Self-concept of Ability Scale (SCA: Shavelson and Bolus, 1982) and the Affective Perception Inventory (API: Soares and Soares, 1979). Their 1988 study comprised a sample of Grade 11 and 12 Canadian students. The correlational outcomes were such for them to be able to state that 'the results provided remarkably strong support for the multidimensionality of self-concept and the content specificity of general, verbal, math, and school self-concepts' (p.376). This study, along with others (e.g. Marsh, 1990c) however, found that the claim of a general academic self-concept term could not be empirically justified. This was due in part to the lack of correlation between the mathematics and verbal self-concepts preventing the conflation of the two measures into the one scale. This was discussed above as being explained by the I/E effect and was incorporated into the Marsh/Shavelson revision. The revised model comprised at least two self-concept facets at the second order level of the hierarchy (mathematics and verbal self-concepts), with an array of specific first order facets beneath them and arranged relatively to each other in a spectrum bridging verbal/academic self-concept and mathematics/academic. It was hypothesised that specific factors arranged at the ends of the continuum would correlate only with Mathematics or verbal self-concepts where as those positioned nearer the centre would correlate to both mathematics and verbal self-concepts.

Diagram 2.4 Academic Self-Concept Structure, Marsh, Byrne and Shavelson



Academic Self-Concept Structure Marsh, Byrne and Shavelson (1988c)

The vertical lines on the model leading to each of the specific subjects indicate that each facet may...be defined by even more specific components, e.g. 'algebra, geometry, and calculus' or 'literature, composition and grammar'. (Marsh, 1990: 377).

The conceptual framework which separates academic self-concept into its more specific components has considerable empirical support. Marsh (1992) found that achievement in seven school subjects related more strongly to the self-concept measure in its own subject than to a measure in any of the other subject self-concept. Performance in science, for instance, correlated 0.702 with science self-concept, but only 0.453 with English, the next highest subject.

More recent research has focussed on an investigation of domain-specific self-concepts. Lau et al. (1999) for example have carried out a hierarchical and multidimensional study of English self-concept and found evidence to suggest that self-concept has a strong hierarchical nature at the specific subject level. Marsh et al. (1997) in their study of elite physical education students found similarly, that self-concept is remarkably domain specific. The exploration of the contributing subcomponents to science self-concept, or physical science self-concept or

biological science self-concept, or to however science may be structured, has up to now (to the author's knowledge), remained untested. A significant outcome of the research reported within this thesis will be an exploration of this area.

To satisfy construct validity requirements of the multidimensional model, with respect to relationships with academic achievement, two conditions must be met (Shavelson et al., 1976). Firstly, academic achievement must be correlated more positively with academic self-concept than with non-academic self-concept or with global self-esteem, and verbal and maths self-concepts must be correlated more highly with their corresponding achievement indicators than with the non-matching achievement indicators. The same pattern of correspondence must also be evident if additional subjects other than mathematics and verbal are also included.

Secondly, empirical data must demonstrate that measures of academic self-concept can be discriminated from measures of academic achievement. Unless there is a clear distinction between subject specific measures of academic self-concept and subject specific measures of achievement then it might be construed that that the academic self-concept measure is purely an alternative measure of achievement (Byrne and Shavelson, 1986; Shavelson and Bolus, 1982).

Additionally, construct validity requirements of the hierarchical model impose yet another set of conditions on the multilevel facets (Shavelson et al., 1976). Correlational relationships must be consistent with the structure pattern of the model, with correlational strengths varying in size depending on their relative position in the hierarchy whilst remaining consistent with the multidimensional requirements. This generates three conditions. These are that, correlational relationships between academic achievement and its corresponding first order self-concept facets should be, i) stronger than correlations between that subject's academic achievement and other first order self-concept facets within the same rubric, ii) even stronger than correlations between that subject's academic achievement and other first order self-concept facets within a different rubric, iii) even stronger still than that correlational relationships between that subject's academic achievement and non-academic self-concepts. In relation to science achievement the pattern of relations would look like this:

i) Sci achievement & Sci self-concept > Sci achieve & Maths self-concept

ii) Sci achievement & Maths self-concept > Sci achieve & Verbal self-concept

iii) Sci achieve & verbal self-concept > Sci achievement & Non-ac self-concept

The same conditional pattern could be expressed in an alternative way:

i) Sci self-concept & Sci achievement > Sci self-concept & Maths achievemt

ii) Sci self-concept & Maths achieve > Sci self-concept & Verbal achievement

iii) Sci self-concept & Verbal achieve > Sci self-concept & Non-ac achievement

There is still no agreement (see Byrne 1990; Marsh, 1990) as to how general academic self-concept should be defined or which components exactly constitute academic self-concept. What is known however, and what will be explored below, is that educational outcomes such as achievement and performance as well as other favourable outcomes like 'task choice, sustained effort, persistence in the face of difficulty and coursework selection' (Marsh, Walker and Debus, 1991) are affected favourably by self-concept and that these effects are more strongly felt at the specific level rather than at the general level (Marsh et al., 1988).

TIMSS and Self-Concept

Recently, there has been a growing trend toward implementing large scale international education projects. Examples of these include the Program for International Study Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Repeat Third International Mathematics and Science Study (TIMSS-R). The TIMSS-R study was the fourth in a series of 'world' studies into science and mathematics with the stated intention to, 'isolate the factors directly related to student learning', (Gonzalez and Miles, 2001:Sect. 1, p.5). The TIMSS-R study collected a host of academic achievement data, together with student background information, and of relevance here, data relating to self-concept and attitude constructs. TIMSS-R defines self-concept as 'confidence in ability', (Supplement 3, 2001, Section 1 – Student), and seems to use the term 'confidence' interchangeably with the term 'self-concept'. The nomenclature is not

justified within the project documentation and there seems little literature based evidence to support this usage. TIMSS does not present a model of self-concept and it is unclear upon what theoretical framework it is based. No empirical data is given for the validity for the TIMSS-R self-concept and no justification for the item selection of grouping rationale. The *positive attitude* variable, for instance, is not clearly demarcated from self-concept and indeed some of the attitude items may well have been included within the variable self-concept. If there is an acceptance of Shavelson et al.'s (1976) broad definition of self-concept as *perception of self* then clearly 'I would like science much more if it were not so difficult' would probably be excluded. However, given these reservations the TIMSS-R data along with its predecessors are of immense value because of their shear scale and richness of data. For this reason the TIMSS-R self-concept scale will be considered alongside the others scales discussed earlier.

Self-Concept and Academic Achievement

The possibility that academic self-concept has an influence on subsequent academic achievement has probably been one of the strongest motivating factors for self-concept research. It has certainly generated a great deal of research interest and academic achievement is one of the most frequently examined constructs in between network studies of self-concept. Models, principles and methodological strategies relating to the nature, structure and measurement of self-concept have made considerable advances in the last 30 years and therefore a consideration of studies reported in the 1970s and 1980s have sometimes little to add to the arguments being rehearsed today. However, they provide context and landscape from which the more recent and more robust studies have emerged. Early studies exploring between network research involving self-concept and academic achievement resulted in contradictory and disappointing outcomes which contributed little to our understanding of self-concept (e.g. West, Fish and Stevens, 1980). Much of the early work concerned itself with research relating to global self-concept or general academic self-concept at a level of specificity much broader than contemporary research which, we now know, has yielded more robust data. Consequently, reviews of this early research (e.g. Byrne, 1984) and meta-analyses (e.g. Hansford and Hattie, 1982) revealed null findings, contradictory evidence and wide discrepancies between researchers.

Advances in self-concept thinking have brought with it improvements in methodological strategies and conceptual frameworks. There is now widespread agreement that self-concept v academic achievement relationships cannot be understood outside the multidimensionality paradigm (e.g. Byrne, 1996a; Marsh, 1990a), and therefore much of the current research concerns relationships between subject specific academic self-concept indices and achievement in the corresponding subjects.

A significant number of studies, although not exclusively, have tested the self-concept v achievement relationship using the revised Marsh/Shavelson model through use of the associated Self Description Questionnaire (SDQ) instrument and its more academically focused stable mate, Academic Self Description Questionnaire (ASDQ: Marsh, 1999). As a consequence of this, much of the research reviewed on the relationship between achievement and academic self-concept is based on the SDQ or ASDQ instruments.

A great deal of the between-network research of academic self-concept and achievement has emphasised the importance of subject specific measures and has provided support for the structural validity of the multidimensional nature of self-concept. Shavelson and Bolus, (1982) reported that grades in mathematics, English and science were more highly correlated to corresponding areas of academic self-concept than to global self-concept. Hansford and Hattie (1982) reported that measures of academic performance and academic ability correlated with self-esteem and undifferentiated measures of general self-concept at 0.20, but at a higher correlation of 0.40 with measures of academic self-concept.
Marsh and Gouvernet (1989) used the Perceived Competence Scale for Children (PCSC; Harter 1982) and found that mathematics and reading attainment, as measured by standardised tests with Grade 7 to 9 Australian pupils, were more highly correlated to academic self concept than they were to other non-academic self concepts, i.e. social, physical and general self-concept facets. Marsh (1990) revisited this theme sometime later with the addition of two further self-concept test instruments, the Piers-Harris Children's Self-Concept Scale (PHSC; Piers, 1984) and the Self Description Questionnaire 1 (SDQ-1; Marsh, 1992a).

The results indicated that, in each of the scales, academic achievement in mathematics and reading correlated more highly with academic self-concept facets

than with non-academic self-concept facets. This provided further evidence of the validity of separating the academic from the non-academic scales of self-concept.

SDQ research has provided substantial data on the achievement-academic self-concept relationship. Marsh (1990a; 1992a) reviewed eleven SDQ studies conducted with Australian children in Grades 7 to 9. Mathematics and reading scores were collected by way of standardised tests and/or teacher ratings. Thirteen correlations were obtained between mathematics self-concept and mathematics achievement and seventeen correlations between reading self-concept and reading achievement. All correlations were significant and positive with the mean correlations of the two subjects were 0.35 for mathematics and 0.38 for reading. Marsh also found that the correlations between mathematics self-concept and reading achievement and vice versa were positive, small and non-significant. This is consistent with the I/E model. He also found that the correlations between reading and mathematics achievement and non-academic self-concept facets were also very small with only two significant from 136 correlations.

Marsh (1992d) extended this earlier work by focusing more particularly on the relationship between the specific components of academic self-concept of eight school subjects and academic performance in those subjects. Consistent with the predictions from structural theory, Marsh found the correlations to be substantially high and significant in matching areas of self-concept and achievement, ranging from 0.45 to 0.70 with a mean of 0.57, and lower in non-matching subjects. More recently, this correlational pattern was also seen by Marsh, Trautwein, Ludtke, Koller and Baumert (2005) in a study in German schools. They found large and systematic patterns of correlations between specific academic self-concepts from a range of different school subjects and their matching academic outcomes as measured by standardised test scores, school grades and coursework selection. For example, mathematics self-concept was correlated to mathematics test score by 0.59, to school grades by 0.71 and taking advanced mathematics course by 0.51. However, mathematics self-concept correlated to English test score by 0.01, school grades in German by 0.06, English by 0.11 and taking advanced courses in English by -0.27.

The evidence from a range of studies appears to indicate rather strongly that academic self-concept is substantially correlated to academic achievement and that the correlations are stronger when the measure of academic self-concept becomes more specific. It also shows that there is little or no correlation between academic achievement and non-academic self-concept measures or with global self-esteem.

Causal Effects of Self-Concept

To establish a correlational effect between self-concept and academic achievement is interesting and has potential implications for classroom practice and education policy. What the research does not say however, is anything about their temporal ordering or the causal relationships between the two constructs. Does academic achievement have a causal effect on academic self-concept; does academic self-concept have a causal influence on academic achievement; both or neither? The theoretical basis of the academic self-concept model *assumes* a causal link, in that academic achievement is one of the prior determinants of academic self-concept (Marsh, 1993). Correlational relationships do not imply causality and never should it be assumed that it might be the case. In the past, even the recent past, it was not possible to be able to investigate causality and researchers contented themselves with reporting correlational data. Attempts at causal ordering measuring have been reported although these early attempts have been heavily critiqued for methodological inadequacies (see Marsh, Byrne and Yeung, 1999). Recent advances in methodological design coupled with the much increased processing power of modern computers have allowed us to explore causality more robustly. Researchers however, would still be well advised to remain cautious and conservative with their claims.

Calsyn and Kenny (1977) proposed two models by which self-belief and achievement could be visualised and this has provided a framework through which the competing notions of the direction of causality can be viewed. The Skill Development Model posits that academic self-concept emerges as a consequence of academic achievement, that is, levels of academic achievement causes changes to academic self-concept, but not vice versa. By contrast, the Skill Enhancement Model implies that academic self-concept is a primary determinant of academic

achievement, that is, levels of academic self-concept directly causes changes in academic achievement, but there is not an effect in the opposite direction. Marsh (1990d) argued for a third way, a Reciprocal Effects Model, in which the causal influences would flow in both directions with prior self-concept affecting subsequent achievement and prior achievement affecting subsequent self-concept.

The theoretical explanation for the Skill Enhancement Model is based on the idea that individuals act in a way as to maintain consistency with their self-view (see Swan, 1997). The consequence of this is that individuals with a high self-view of themselves behave or attempt to perform in a way which is consistent with that self-view. So, individuals with a high academic self-concept apply themselves to their school work in such a way that attempts to preserve or protect that high academic self-concept which they perceive and value. According to Rosenberg (1979) this results in an individual having the capability to achieve more highly as a result of their motivational state. Rosenberg (ibid) also suggests that individuals with a high self-value might strive for academic achievement as a way of preserving their self-worth. Bandura (1997) proposes that self-efficacy also has an effect on achievement as high self-efficacy promotes exertion of effort, task persistence and selection of adaptive goals. However, Marsh, Walker and Raymond (1991) visualise self-efficacy as a dimension of self-concept and attribute positive behavioural outcomes not directly to the causal influence of high self-efficacy per se, but to its role in influencing subsequent action through the cognitive, affective and motivational mediation processes which the performance expectances instigate.

Valentine, DuBois and Cooper (2004) carried out a meta-analysis of longitudinal investigations into the causal relationships of self-beliefs and achievement. They considered that methodologically strong studies should control statistically for baseline levels of achievement and then predicted future achievement using self-beliefs, thus allowing enquiries into the possible contributions of self-beliefs to changes in achievement over time. Specifically, in their meta-analysis they reinterpreted past data to test the strength of relationship between self-belief and achievement after controlling for prior achievement. The analysis was undertaken on all the studies which met their inclusion criteria of i) a measure of self from any of self-concept, self-esteem, self-efficacy, self-perception or self-competence, ii) longitudinal studies with measures taken at least at two different time instances, iii)

enough data to compute or report later achievement, controlling for prior achievement, together with results in the form of a standardized regression or path coefficient. They filtered studies down to 55 reports on 60 independent samples, containing 282 separate effect sizes.

Their meta-analysis yielded an outcome which was consistent with the view that self-beliefs can influence future academic achievement, although the size of the effect was quite small. They report that the effect size, $\beta = 0.08$ which approaches the threshold for small effect sizes as defined by Cohen (1988) as $\beta = 0.1$. In summary Valentine et al. conclude:

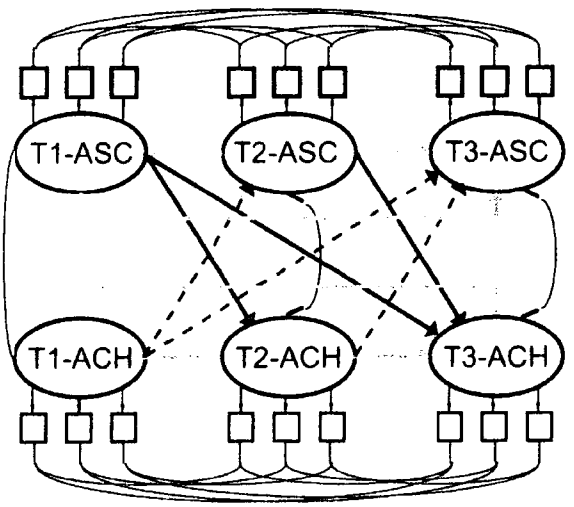
> Overall, results suggest that, among equally achieving students, having positive self beliefs confers a small but noteworthy advantage on subsequent achievement measures relative to students who exhibit less favorable self-beliefs (p.127).

From a theoretical perspective the result adds weight to those who view the self as a causal agent, (e.g. Bandura 1997; Deci and Ryan, 1985) and provides support for the Skill Enhancement Model of self-concept.

Marsh and Craven (2005) propose what they call 'a prototype for the idea causal modelling study' (p.22). The diagram represents a full forward multiwave-multivariable model in which multiple indicators of academic achievement (ACH) and academic self-concept (ASC) were recorded in three successive waves at times T1, T2 and T3. The small square boxes represent the multiple indicators; the ovals represent latent variables of ASC or ACH factors derived from the indicators and the straight lines with arrow heads represent possible causal paths and curved lines represent covariances. Each latent variable is connected to each other latent variable in subsequent waves.

For the Skills Model to be valid, i.e. ACH influencing ASC, data would indicate that only pathways from prior achievement to subsequent self-concept would be positive. For the Self Enhancement Model to be valid, i.e. ASC influencing ACH, only pathways from academic self-concept to academic achievement would be positive. For the Reciprocal Effects Model to be valid all pathways would be positive.

Diagram 2.5 Prototype Causal Ordering Model



Prototype causal ordering model to test self-enhancement, skill development and reciprocal effects models. Marsh and Craven, 2005.

Marsh et al. (2005) implemented such a model to undertake two tests of the reciprocal effects model on causal ordering of self-concept and achievement. They also extended the study to juxtapose self-concept against academic interest as causal influences. The studies were based on a representative sample Grade 7 (and Grade 8) German students. In the first study the sample comprised 5649 Grade 7 (13 year old) students. Mathematics self-concept, mathematics achievement and mathematics interest were measured on two occasions (T1 and T2). Mathematics self-concept was measured with a five item questionnaire and mathematics interest with a four item questionnaire. The mathematics achievement test items were taken from the First and Second International Mathematics Study and mathematics grades were self-reports from the end of the sixth Grade.

The second study was very similar except that it was longitudinal over two years (Grades 7 and 8) and so data collection waves were separated by a full academic year. This served to evaluate the replicability of results from Study 1 and the generalisability of results across two school years and with two different interest measures. The second study used a sample of 2264 students. Following SEM analysis of both studies Marsh et al. (2005) were able to report:

> Our results provide clear evidence that prior academic self-concept does predict subsequent academic achievement beyond what can be explained in terms of prior measures of academic interest, school grades, and standardized achievement test scores (p.412).

The effect was significant in both directions, supporting the idea that there is a reciprocal effect. Although the effect was reciprocal the data from both studies indicated that 'the effects of self-concept on achievement are stronger than the effects of achievement on self-concept', (p. 411). They also reported that academic self-concept was more highly correlated than academic interest with achievement and the causal effects to be much stronger. Indeed, the causal effects of academic interest on subsequent achievement were largely non-significant.

As previously mentioned there has been a plethora of research literature exploring the possible relationships between academic achievement and various aspects of self-concept, in fact Byrne (1990) reports that in academic self-concept research the majority has focussed on its relationship with academic achievement. Results from early work, as indicated above, were inconsistent and often contradictory. More recently, there has been a growing body of evidence that relationships between the two are real, genuine and significant, both in the statistical sense and in the impact as experienced by ordinary learners. There is also evidence that the influences between self-concept and achievement are reciprocal in nature.

However, once trapped within a particular paradigm, which is inescapable, it is important to guard against celebrating the clothed emperor. It is useful then to give careful attention to alternative points of view, even if the tone appears more cynical than critical, more pedantic than insightful, more contrary than constructive, and particularly if the case is over stated to make the point. There have been those which have sought to take the opposite view, and this has been seen by some researchers as an unwelcome distraction and annoyance, particularly from some researchers who see their life's work being undermined.

Baumeister, Campbell, Krueger and Vohs (2003) presented a scathing critique of many of the conclusions and implications from much published self-esteem research. Their analysis, commissioned for Psychological Science in the Public Interest, refutes many of the positive outcomes claimed by researchers into

measures of self and asserts strong reservations as to whether self-esteem has any influence on academic achievement, or indeed, any other significant educational outcome. In fact they speculated that high self-esteem may even prove to have an undesirable influence in many circumstances. Baumeister et al. conducted a thorough review of the self-esteem literature, reducing a possible 15000 articles to a relative handful through a filtering process operating strict criteria. In their cautiousness to reduce fragmentation and maintain a charged affective domain they decided to exclude all research articles which utilised domain specific self-measures and retain only those articles where measures relating to global self-esteem were used. The operationalisation of this choice brought with it some unfortunate consequences. Much of contemporary self-concept and self-efficacy research has argued for the importance of a multidimensional viewpoint of self measures and is cautious and sceptical about the usefulness of unidimensional views or global measures of self. The arguments for a multidimensional perspective have already been well rehearsed and compelling, and indeed cross-network studies have revealed that correlational relationships between self-concept measures and educational outcomes like achievement increase as the level of specificity of the self-concept domain increases (e.g. Choi, 2005; Marsh, Byrne and Shavelson, 1988; Marsh, Walker and Debus, 1991; Pietsch, Walker and Chapman, 2003).

Baumeister et al. (2003), not unexpectedly report inconclusive data and contradictory results in their cross-network review of self-esteem and achievement. A similar flaw appears in their critique of causal effect studies. Due to their selection criteria they review studies mainly carried out in the very early days of self-concept / self-esteem research when methodological technique was less robust than today and statistical analysis less powerful. In fact, the most recent study included in their review of causal influences was published in 1990, which was some 13 years old at the time of review publication whilst omitting a great deal of modern research representing contemporary thinking with methodological improvements. This is most unfortunate as legitimate critique is important yet this influential review missed an opportunity to make a valid and worthwhile contribution.

METHODOLOGY

Methodological Framework

The central purposes of this study were to explore the multidimensional and hierarchical nature of academic self-concept in science and to provide a model for the way in which academic self-concept in science may be conceived. This required an investigation of the internal structure and measurement qualities of a proposed psychological construct. In undertaking this task the simultaneous use of two methods was needed; namely those of path analysis and factor analysis. This combined technique is commonly referred to as Structural Equation Modelling (SEM).

Path analysis has been an important tool in theory and model testing for almost a century. It is a mathematical analytic technique first developed by geneticist Sewall Wright more than 80 years ago. The technique which has proved to be extremely powerful, has significantly widened the statistical landscape and revolutionized the process by which data can be analysed (Denis and Legerski, 2006).

Path analysis is an extension of the statistical technique of simple regression in which the aim is to estimate the magnitude and significance of any hypothesised causal connections between sets of variables. Goodness to fit indices are calculated by comparing the regression weights predicted by the model with those actually observed for the indicator variables. The goodness to fit indices allow different models to be directly compared by expressing the closeness of fit of the models against the data. Path analysis will not determine whether a theory or model is 'true', but it will assist in testing whether the relations in the data are consistent with theory. Path analysis can only be applied appropriately and successfully if there is first an explicit theoretical framework against which the data can be tested. The power of path analysis lies not in *generating* theory but in *testing* the proposed theory. The assistance it provides for the researcher is in helping to reject or modify inaccurate causal models.

Path analysis is often called "causal modelling" and indeed it is commonplace to refer to causal connections between variables in the model. However, as Brannick (2006) points out:

> the "causal" in "causal modeling" refers to an **assumption** of the model rather than a property of the output or consequence of the technique. That is, people assume some variables are causally related, and test propositions about them using the techniques. If the propositions are supported, it does **NOT** prove that the causal assumptions are correct (Brannick, 2006, p.1 Original emphasis).

The use of the term 'causal' is widespread in the modelling nomenclature. Denis et al. (2006) assert that the often used term of 'causal modeling' arose directly out of the context of Sewall Wright's work. It was Wright who first applied the term, although it was not that Wright was wrong in referring to causal linkages in his model. (It's important to know Wright isn't wrong!) In Wright's particular case it was reasonable to assume an underlying causality among the variables in the networks he proposed. He was actually working on the genetics of guinea pigs and as such was within a much more positivistic paradigm. However, the causality implicit in his works (and that of others, e.g. Duncan and Hodge, 1963) connected less to the notion of path coefficients, and more to the substantive claims of his research. Viewed in this way, the method of path analysis, (along with other statistical tools), is not 'causal' but is 'simply a calculating machine applied to a substantive problem of theoretical interest' (Denis et al., 2006, p. 1).

The assumptions under which path analysis can be successfully operated are the same as hold true for regression. Path analysis is particularly sensitive to the model specification as omitting significant variables or introducing extraneous variables can greatly affect the path coefficients. These path coefficients, in turn, are used to make calculations of the significance of 'causal' paths to each dependent variable. It is the outcome of this process which yields the opportunity to compare different models and evaluate model fit. When the variables in the model are *latent variables*, i.e. unobservable constructs formed from multiple observed indicators, then path analysis is termed structural equation modelling (SEM).

SEM is therefore an extension to path analysis and is a technique for estimating unknown parameters given a set of simultaneous equations. These equations are used to map out the interrelations among a pre-determined network of variables. By convention, the recognised difference between path analysis and SEM is whereas path analysis specifies relations among single indicator variables (observed variables), SEM, can in addition, allow for the estimation of latent variables (Denis et al., 2006).

An important component of SEM is *factor analysis* which utilizes the covariation among a set of observed variables in order to gather information about underlying (unobserved) latent constructs (or factors). Factor analysis can either be exploratory or confirmatory in approach. Exploratory factor analysis (EFA) is used where the connection between the observed variables and latent variables is unknown. The process proceeds in an exploratory manner to establish the existence and strength of connections between the observed variables and the hypothesized factors. The analysis determines to find the smallest number of factors which account for correlations or covariances among observed variables. The relationship between observed variables and latent variables are represented by factor loadings. A strong model would produce well differentiated factors whereby, the directly measurable items would exhibit high factor loadings on their related latent variable, and low factor loadings on unrelated latent variables. The procedure is exploratory in the fact that the researcher has no prior knowledge as to how many factors can most appropriately and effectively explain the covariance and map the variables to the factors.

In contrast to EFA, confirmatory factor analysis (CFA) is used when previous empirical experience or knowledge of underlying theory allows a researcher to propose in an a priori manner the expected relationship between the manifest (observed) and latent variables. A priori specifications allow items to freely load on one factor whilst restricting them to have zero loadings on all other factors. The advantage of this is that the exact form of a factor model can be specified and the statistical indices can be derived to determine the extent to which the model fits the empirical data.

Structural equation modelling as a statistical method brings together path analysis, through the application of multiple regression techniques, and confirmatory factor analysis simultaneously to test 'causal' theories involving experimental or non-experimental data. The term SEM conveys two important aspects of the procedure:

> (a) that the causal processes under study are represented by a series of structural (i.e. regression) equations, and (b) that these structural relations can be modeled pictorially to enable a clearer conceptualization of the theory under study (Byrne, 1994, p.3).

SEM has significant advantages in statistical processing over conventional methods and has been fully utilized in this study. As a family of techniques it is much more powerful than traditional statistical methods (Garson, 2006) such as principle component EFA procedures. The weakness in the EFA procedure is that the collected factors are imperfect representations of the hypothetical construct. The presence of unknown extraneous influences plus the variance in each of the items work together to make it difficult to interpret the true variance, and thus making inferences about the proposed theory extremely difficult. This study has made some use of EFA as an adjunct to the more rigorous CFA, although its use was limited. SEM, by contrast to EFA, is able to carry out clustering and multiple regression simultaneously thus helping to isolate and analyse measurement error, unexplained variance and 'true' variance at the same time.

Other advantages that SEM possesses over factor analytic and multiple regression are that they allow for the specification of regression structure between the latent variables (Byrne, 1998), where the impact of one latent variable on another can be hypothesized in the model structure. This particular feature was utilized extensively in this research such that the relationships between the new latent variables could be tested in the various different models. There is a key feature in the way in which parts of a model can be described, or the whole model divided. A *measurement model* depicts the links between latent variables and their associated observed measures (manifest variables), and a *structural model* depicts the relations between the different latent variables.

SEM is not without its limitations and not all researchers value its application. SEM is a statistical procedure which is not universally welcomed and indeed there are those who feel that it is a misemployed technique. A fuller critique of SEM will be carried out later in the study, particularly in relation to the use of modification indices.

In order to carry out SEM procedures a specialist statistical package was employed, and there was a choice of three different software packages. These packages were; LISREL 8, AMOS and EQS. Full reference information for the packages and a detailed review of each program can be found in 'Software Review' (1998). The package chosen here was LISREL 8, (LInear Structural RELations; Jöreskog and Sörbom, 1993). LISREL 8 is not an intuitive program to use, and in fact has been euphemistically called 'heavy going' by Diamantopoulos and Siguaw (2000).

However in its favour it has been found to be extremely powerful, versatile and has had widespread use in contemporary self-concept research.

Procedure

According to Garson, (2006) SEM is carried out as a two stage process. The first stage is to validate the measurement model through the use of confirmatory factor analysis and the second stage is to fit the structural model through path analysis with the latent variables. Diamantopoulos et al., (2000), identified an eight step process to use SEM in the model testing/validation and to achieve these two goals. The steps are:

1. Model conceptualization
2. Path diagram construction
3. Model specification
4. Model identification
5. Parameter estimation
6. Assessment of model fit
7. Model modification
8. Model cross-validation

These steps provided a methodological framework which guided this study and as such the sequence of procedures above were followed particularly closely. What follows next is a report of the SEM procedures carried out for this research from the model conceptualization stage to the model cross-validation stage. In order to aid continuity of the description and evaluation of the modelling process, where issues arise that are outside the modelling process but the modelling process relies on the information or resolution of the issue, then the issues will be discussed at that time. This should result in a tighter more focused analysis of how the modelling was carried out together with the rationale for the choices taken.

Model Conception

The modelling process began with the *model conceptualization* stage. In fact, a number of alternative models were hypothesized such that they could be later compared and judgments made about which, if any, provided the best fit to the data. This began with a visualization of a number of different models based on the self-

concept theory together with an understanding of the impact that learning science has on young people. This was probably the most important stage of the modelling process. A model which is poorly conceived in relation to relevant theory or inconsistent with other successful empirical studies cannot later be retrieved or rescued through any amount of skilled post-conceptualization work.

Appropriate self-concept facets were chosen and these were conceptualized as latent variables within the models. Relationships between the latent variables and the manifest (indicator) variables together with their structural features within the model were also conceived. The process of designing, refining and selecting these indicators will be discussed later in the thesis. Thought was given to the relationship between the latent variables and the number of indicators feeding into each latent variable. Good practice usually recommends that at least two indicators (Byrne, 1989) should be used to inform each latent variable. The relationship between indicators and latent variables was guided by methodological knowledge of other self-concept studies and self-concept theory, particularly lessons learnt from the Shavelson/Marsh model. Each latent variable was identified as being either exogenous (independent) or endogenous (dependent) and relationships between the latent variables, including their direction, was identified. Seven different conceptual models were eventually hypothesized and these are indicated below along with a brief description.

The design and selection of the latent variables was influenced by two criteria. First, the way in which science was defined and disentangled from being a unitary construct to being a multidimensional construct; second, the perceived curriculum experiences of young learners. Within this study 'science' is defined as 'the pursuit of better investigative strategies and more reliable information about the physical and biological world' (DES, 1988, p.A6). With this in mind, the curriculum was conceived to be located within three broad domains, that of, (i) conceptual knowledge and understanding of science, (ii) procedural knowledge and understanding of science, and (iii) ideas about the nature of science. From this, five facets of science self-concept were conceived; Physics, Chemistry; Biology; Scientific Enquiry and Nature of Science. As discussed above 'balanced science' was a statutory requirement in all maintained schools from 1989, however despite this, school departmental structures, resources/equipment and staff expertise have still often remained organised within the physics, chemistry, biology groupings, and it has continued to

remains a strong part of the mind set of science teachers. Self-concept by its very nature is shaped and influenced by the individual's experiences (Byrne 1996) and thus school pupils may well differentiate strongly between physics, chemistry and biology as broad descriptions of scientific knowledge and curriculum activity. It was therefore felt to be appropriate to use these distinctive curriculum elements to form meaningful self-concept facets. Scientific enquiry is a key aspect of school science and was something which was well known to pupils, probably under the label *investigations*. Undertaking investigations became the dominant pedagogical framework by which many pupils engaged with their practical laboratory based activity and other related facets of learning about science enquiry.

The final of the five new self-concept facets was what is termed in this study 'Nature of Science'. This was not a statutory feature of the 1999 science curriculum, although interestingly it was part of the Science National Curriculum at its inception in 1989 (DES, 1998). Some aspects of the Nature of Science have been re-introduced into the Key Stage 4 new science curriculum (introduced in September 2006) through an aspect of the Programme of Study called 'How Science Works' (DfES, 2006). Despite the fact that much of the Nature of Science is not a statutory part of the National Curriculum for science, it still embraces a view of science seen by some influential science educators as an important non-statutory element of young people's science education (Millar, 1993; Leach and Scott, 2003). In this context the *nature of science* was defined as being concerned with ideas relating to an understanding of:

- the purpose of scientific work;
- the nature and purpose of scientific knowledge;
- science as a social enterprise.

Given the importance of these ideas and the balance that its inclusion brings to the shape of the model, it was decided to include the Nature of Science as the fifth facet.

Path Diagram Construction

Path diagram construction allowed for the model conceptualizations and measurements schemes to be graphically represented. What follows are seven hypothesized models with varying degrees of parsimony. Seven different models will be presented. The first four models are consistent with the ideas of science and self-concept that have been rehearsed previously. However, for completeness, it was felt appropriate that a few selected other models possessing conceptually different structures to those hypothesized were also offered for testing such that they could be included or eliminated from the range of possible final 'best' models. These additional models are listed as Model 5 to Model 7.

Model 1: Indicative multidimensional and non-hierarchical model for the full-science self-concept

This was a simple multidimensional and non-hierarchical model of science self-concept. It was the *measurement model* component of the full structural equation model. There were five latent variables; Physics, Chemistry and Biology representing the knowledge and understanding aspects of science together with the procedural understanding aspects represented by Enquiry, and finally the Nature of Science term.

Diagram 3.1 Model 1

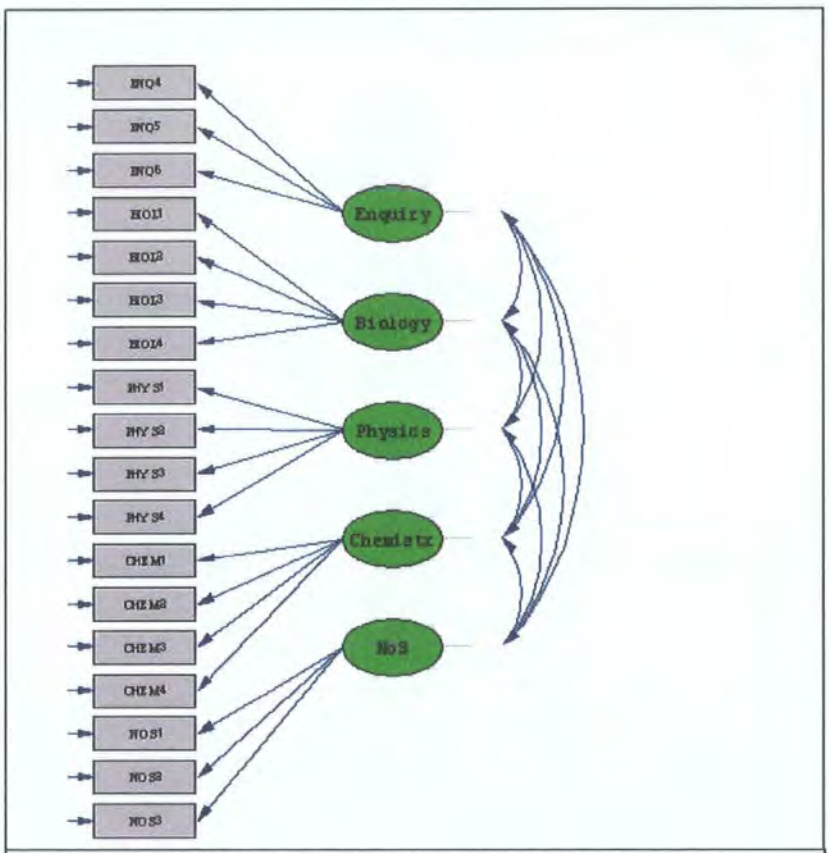


Diagram 3.1: Model of multidimensional (m/d) science self-concept. This specifies the measurement model for Models 2, 3 and 4.

The boxes represent manifest variables and the ovals represent latent variables. The arrows between the latent variables to the manifest variables represent the path coefficients for regression of observed variables onto unobserved factors. The

curved lines represent the correlation between factors. The arrows entering each manifest variable from the left indicates the measurement error.
For this multidimensional science model of self-concept to be successful it required that all five facets be distinct, with substantial factor loadings of the manifest variables onto the respective a priori constructs. It also required the correlations among these constructs to be substantial but distinctly different from one another. There was no hierarchical element with this model.

At this stage the number of manifest variables relating to each latent variable was purely *indicative*. The final decision about the numbers of manifest variables and how these related to the data collection instrument was decided at the stage of *model specification.*

Model 2: Indicative multidimensional model for science self-concept with an inferred (non-measured) hierarchical General Science factor

This model tested a multidimensional and hierarchical model of the full-science view of science self-concept. This model tested whether the five individual facets of science successfully combined to form a further inferred higher order facet of General Science. In order for the model to have supported the notion of a hierarchical structure two conditions needed to be satisfied. Firstly, that the five first order factors had to be well defined and separable from each other. Secondly, that these factors had to be substantially correlated to the inferred higher factor, and that a good proportion of the five variances should be explained. In other words there should be substantial path coefficients between the higher order and the first-order factors.

Diagram 3.2 Model 2

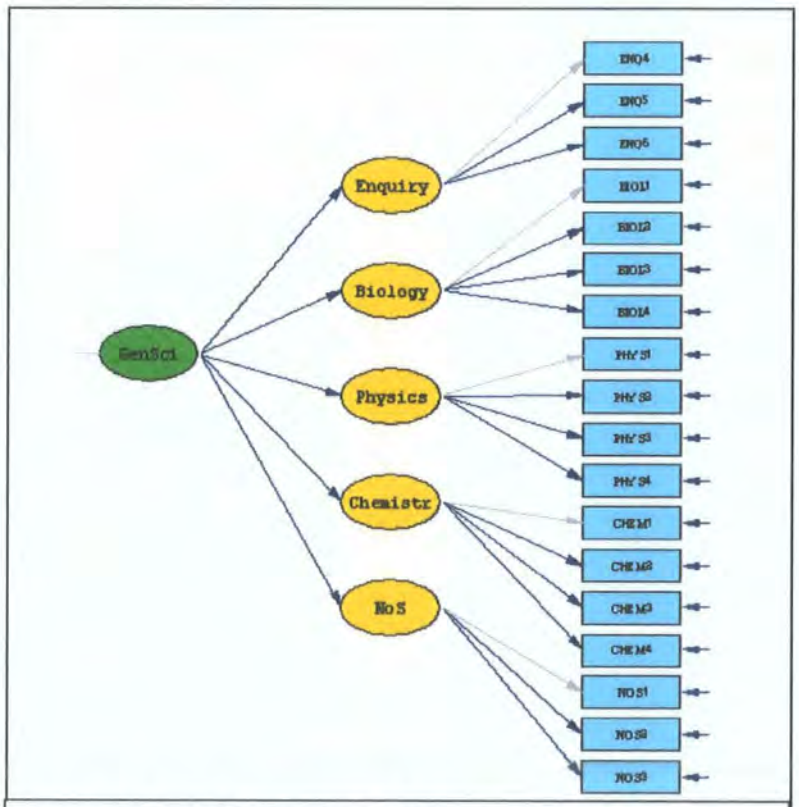


Diagram 3.2 Model 2 hypothesises that science self-concept is multidimensional and hierarchical. The GenSc higher order factor is an inferred (non-measured) term derived to explain the variances within the five other factors.

Model 3: Indicative hierarchical model for full-science self-concept

Model 3 was also a hierarchical model. It differs from Model 2 in the nature of the hierarchical factor. Whereas Model 2 had its higher order factor *inferred* from the five factors lower in the model, in Model 3 the higher order Global Science factor was constructed directly from the measured science manifest variables. This model hypothesized a relationship between the five first order factors of science and a measured higher order Global Science factor. Support for this model had similar criteria to Model 2 with the addition that there should be high correlations between the Global Science term and its manifest variables. The arrow entering into the latent variable of Global Science represents the measurement error, that is, the variance unaccounted for in the latent variable GlobalSc by the manifest variables SCI1 to SCI3.

Diagram 3.3 Model 3

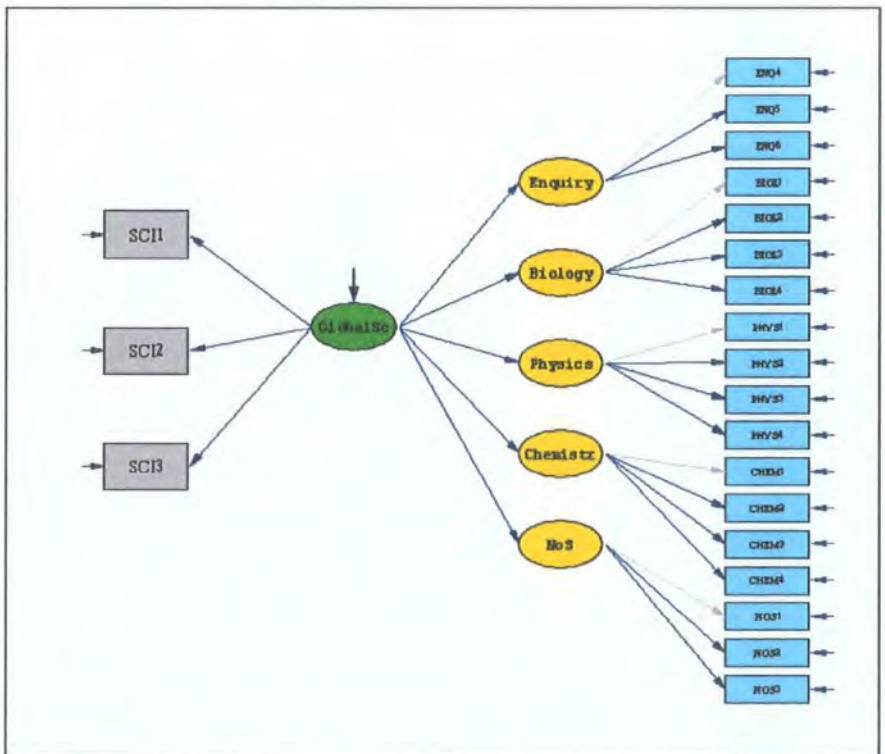


Diagram 3.3 Model 3 hypothesises that science self-concept is multidimensional and hierarchical. The Global Science higher order factor is derived directly from the science manifest variables. This Global Science term must explain the variances within the five lower order factors.

Model 4: Indicative relationship between the higher order and global science self-concept constructs (SKU model)

Model 4 was to determine whether the higher order science self-concept inferred from the first order facets was reflected in the global measure of science self-concept. In addition to the inferred higher order GenSci factor of Model 2, and the measured higher order GlobSc factor of Model 3, this model included both higher order factors together. This model examined the relationship between the global science 'measured term' and the general science 'inferred term'. A strong model required not only substantial path coefficients between the lower order and the higher order self-concept facets but also a high correlation between the global science self-concept and the general science self-concept reflecting their equivalence. The curved line connecting the latent variables Global Science and General Science indicates a correlation between the two.

Diagram 3.4 Model 4

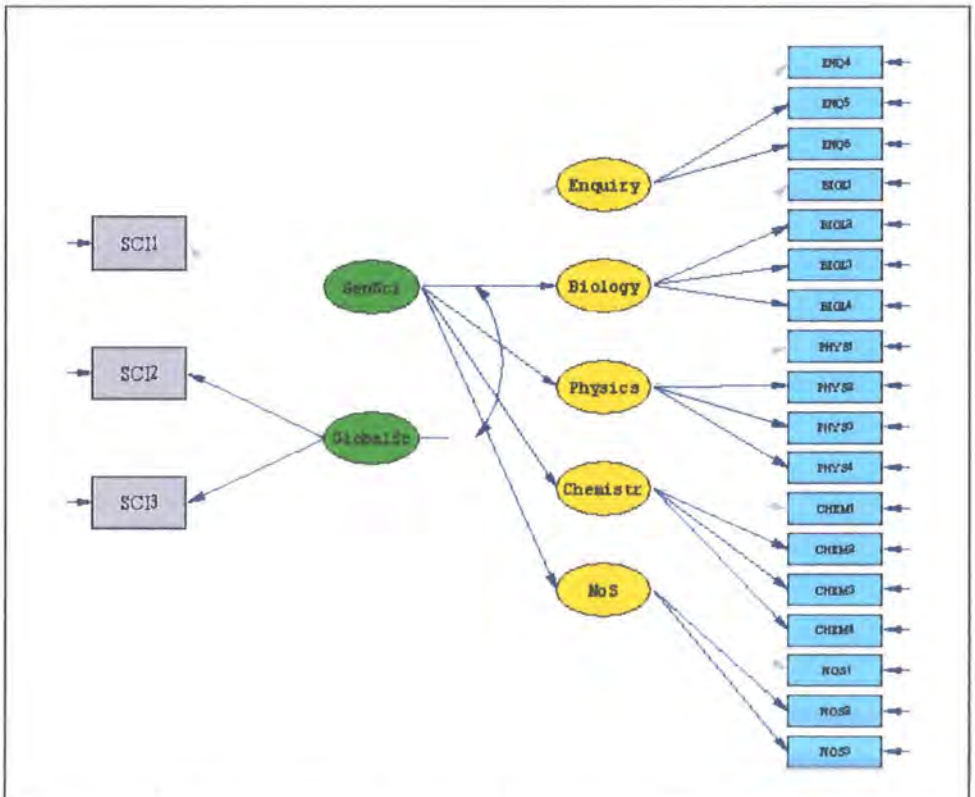


Diagram 3.4 Model 4 hypothesises that science self-concept is multidimensional and hierarchical. This model has both a derived and measured higher order science term. There should be a strong correlation between the two higher order terms.

Model 5: Indicative hierarchical model for PCB view of science self-concept

This was the simplest model proposed for self-concept in science. It consisted of only three latent variables all of which represented the knowledge and understanding aspects of science through the traditionally conceived facets of physics, chemistry and biology. In this model there was no attempt to incorporate self-perceptions of the procedural features of science, that is, aspects related to scientific enquiry or practical investigative work. This is a legitimate model, in terms of the views of science, although it does not include a full description of science in the way in which pupils experience science in the classroom, and importantly, as specified by the National Curriculum for science (DfEE, 1999, DfES 2006). This model's inclusion allowed this hypothesis to be tested through a comparison of model fit between this 'narrower view' of science and a 'full-conception' view of science.

Diagram 3.5 Model 5

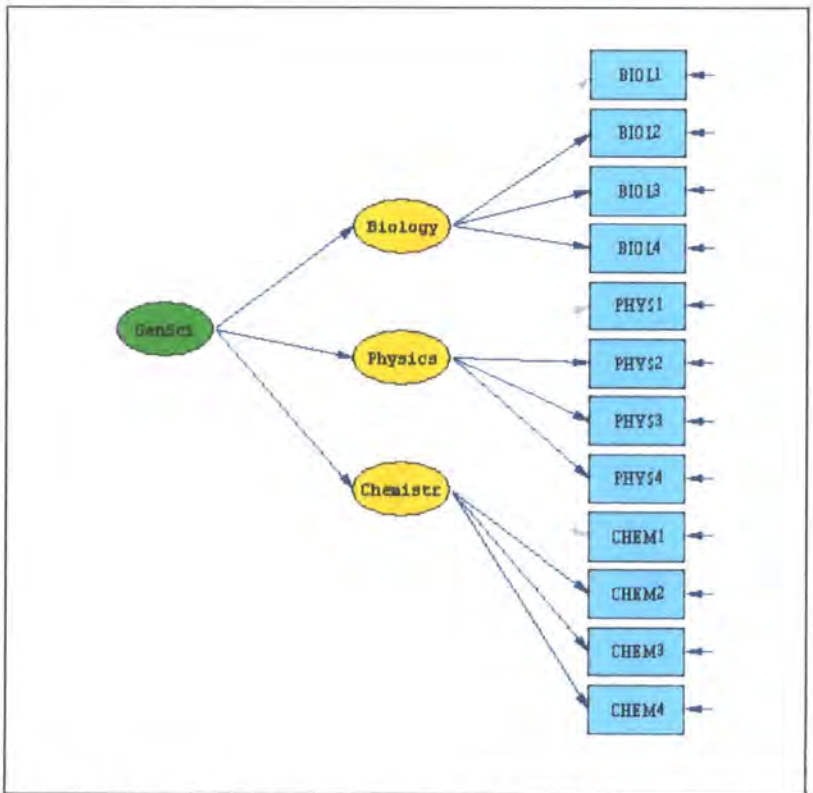


Diagram 3.5: Model 5 represents a more restricted, physics, chemistry, biology view of science, with an inferred higher order General Science factor.

Model 6: Indicative hierarchical model for the Enquiry - Knowledge view of science self-concept

The physical-biological model (Model 6) did not discriminate between physics and chemistry, but did include the latent variables of Enquiry and NoS. This again was a legitimate proposition, as there has been a good amount of curriculum cross-over within the physical sciences of physics and chemistry.

Diagram 3.6 Model 6

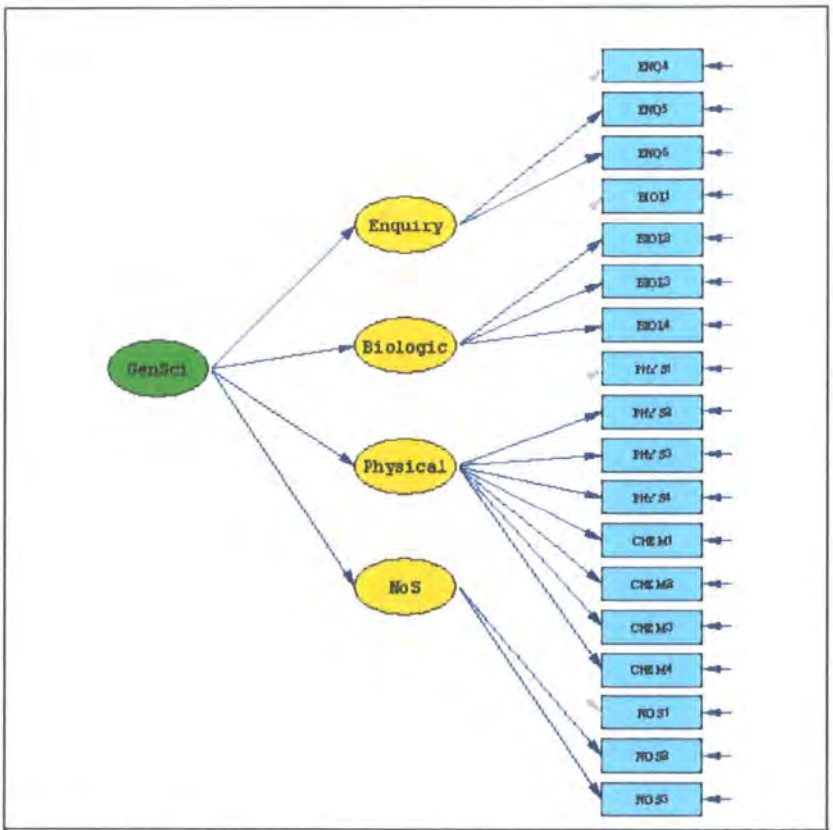


Diagram 3.6: Model 6 represents a view of science composed of physical science, biological science, enquiry and NoS. It has an inferred higher order General Science factor.

Model 7: Indicative hierarchical model for the Physical-Biological view of science self-concept

The enquiry-knowledge model (Model 7) proposed a model whereby science is separated into its most base and least distinct facets. The knowledge components of physics, chemistry and biology were grouped together and were only separated from the Enquiry and NoS variables. This represented science as knowledge-procedures-institution, which again was legitimate although was a less discriminating model.

Diagram 3.7 Model 7

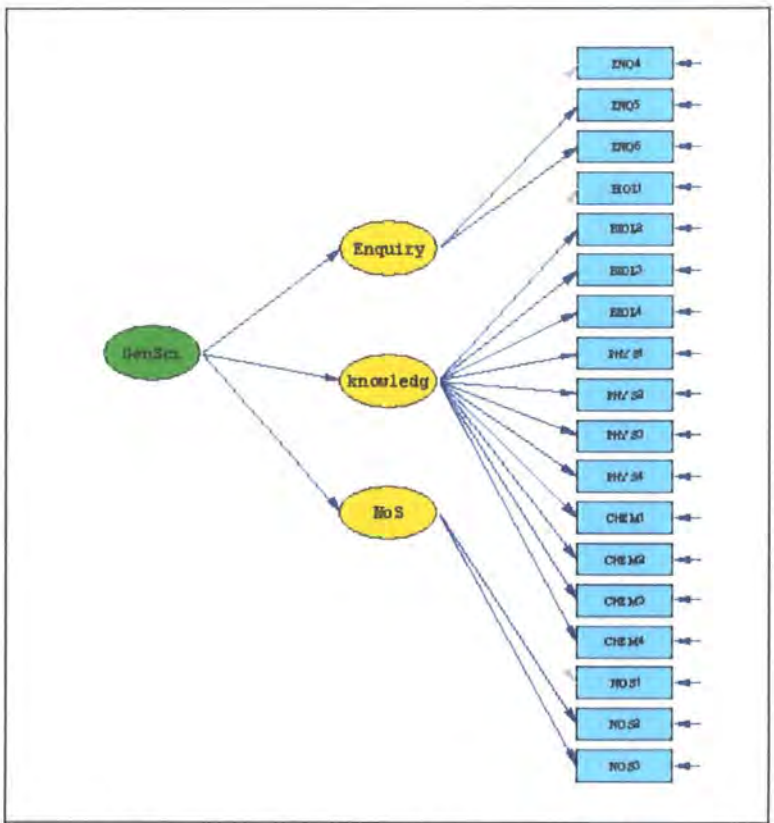


Diagram 3.7: Model 7 represents a view of science composed knowledge-procedures-institution. It has an inferred higher order General Science factor.

Model Specification

Completing the model specification required a description of the nature and number of parameters to be estimated. Normally inputting data straight into LISREL requires the information to be in a mathematical form. Fortunately, as an alternative to this, more recent versions of LISREL have an additional command language called SIMPLIS [1] which has been designed as an aid to LISREL operations and which allows the names of the variables and their specific relationship with other variables to be specified in words and symbols. This greatly aided the model specification procedures.

Before the model specification could progress further a decision needed to be taken about the nature and scale of the data gathering procedures. In particular, the response mode offered to respondents and other related issues concerned with information generation and collection needed to be finalised. This was required since the choice of instrumentation would affect the character of the data and hence the nature of the manifest variables. The manifest variables were the entry points for the data flow into the model and as such would impact on the model specification.

The choice of mode of self-concept measurement was guided by previous studies reported in the literature. The data collection types found in the literature could broadly be arranged within one of four categories and were (i) Self-report measures, (ii) Projective measures, iii) Interview measures and, (iv) Ratings by others.

The literature indicated that self-report instruments were the most frequently and most widely used instruments. Here data are collected primarily through questionnaires, inventories and interviews. These questions are asked directly to respondents about they way they feel or think about themselves. Questions could be either in an open form where respondents complete a sentence , e.g. 'I am good at ___' or 'I prefer ___', or in a closed form where they are asked to make judgments against pre-determined criteria. With pencil and paper testing, responses are usually on a Likert type scale, (see Edwards, 1953) and typically comprised a range of responses consisting either a 1-4, 1-5 or 1-7 scale. An even number scale forces

[1] SIMPLIS (SIMPle LISel) is a command language used to simplify both the creation of LISREL input files and the reporting of output files. Parameter specification is also substantially easier with SIMPLIS.

respondents to commit to either the positive or negative half, whereas an odd numbered scale allows respondents to remain neutral if they choose. An example of this type of question would be 'I enjoy science', with 'true for me', 'false for me' at the extremes of a five point scale.

An alternative to the Likert scale is the 'semantic differential'. Here bipolar statements are placed at either end of a continuum and respondents place themselves relative to the extremes. The scales could have various numbers of increments all providing a quantitative output. An example of this type of question is:

When learning science I am:

Fast	1	2	3	4	5	6	7	Slow
Happy	1	2	3	4	5	6	7	Sad
Strong	1	2	3	4	5	6	7	Weak
Confident	1	2	3	4	5	6	7	Scared

The semantic differential is easily adapted to different contexts with adjustment to the choice of descriptors. This adaptability has also proved to be its downfall in the past, since unless different researchers have chosen identical descriptors it has made comparability very difficult. Hattie (1992) has been critical of research into self-concept using semantic differentials. He feels that a justification or rationale for the choice of bipolar adjectives is rarely provided by researchers which cast further doubt over their claims for validity.

Questionnaire type data collection such as Likert scale and semantic differentials lend themselves quite readily to scrutiny by sophisticated statistical testing. Data collection is 'cheap' in so much that large numbers of individuals can be tested fairly quickly and easily with measurement items usually objective, quantifiable and standardized. This is an attractive advantage.

An alternative to pencil and paper response has been the use of a 'card-sort' activity. Q-methodology (Stephenson, 1953) has used such an approach, although it has mainly been used by clinicians. Q-methodology invites respondents to rank attitudes or judgments. The technique is more concerned with the choices of the individual,

and less concerned with making inferences about populations. It is operationalised by respondents placing cards bearing personality characteristic statements in a pre-determined number of piles such that clinicians can make psychological judgments. The technique is time consuming with the results being less generalisable than with large sample questionnaire data. It lends itself particularly well to small scale clinic use rather than larger scale research use and has found less favour with researchers.

An alternative to Q-methodology is Projective Measures which utilise an indirect measures approach and provides researchers with 'unconscious or unwilling' self-evaluations detail. It is thought that this detail does not often emerge using standard self-report techniques (Wells and Marwell, 1976). Such techniques include Draw-a-Person (DAP: Machover, 1949) where typically an individual is asked to draw a figure, then draw another of the opposite sex, and then are questioned about themselves and their drawing. A number of scoring systems have recently been developed to add quantification to the technique. One system is called the Human Figure Drawing Test (HFDT: Mitchel, Trent and McArthur, 1993) and another the Screening Procedure of Emotional Disturbance, (DAP: SPED; Naglieri, McNeish, & Bardos, 1991; Naglieri and Pfeiffer, 1992). An alternative to the DAP test is the House-Tree-Person Test (HTP) created by Buck (1948). Tests of this type, which according to Bekhit, Thomas and Jolley (2005) are more likely to be used by psychologists in the United States than those in the UK. The *house and tree drawings* are used to gather information relating to the individual's feelings of the home and environment, and the *person drawings* are used to gather information relating to issues of self-concept.

Projective Measures instruments are more useful than orthodox measures of self-concept measure in providing supplementary personal information, although they have been limited in their application outside clinical settings. Some self-concept researchers have used such measures, although Burns (1979) makes the point that there is questionable evidence that researchers are able to access additional information by tapping into the subconscious level that they would not already be able to access at the conscious level.

Interviews allow for a greater in depth and dynamic understanding of an individuals self-concept without the constraints imposed by less responsive means of data

gathering. They offer enhanced opportunity for respondents to provide meaning to their personal descriptions and evaluations. Additionally the interviewer is able to collect observational data in addition to asking for clarification or being able to insert supplementary questions. Interviews are used extensively in clinical settings but there is very little evidence available in the literature about the substantive use of interviews in research contexts. Anderson (1992) however, has reported some examples of the limited use of interviews to investigate self-concept.

'Ratings by others' are indirect methods of self-concept measurement. Two different methods of this have commonly been used. The first method is called public rating and consists of evaluations and judgments of a third party about the self-concept of the target individual. The second method is called inferred rating and self-concept information is inferred about an individual by a trained observer from their interactions with each other. Public ratings of self-concept are thought not to provide particularly consistent or valid outcomes (Wells and Marwell, 1976) and this despite the fact that self-concept does have a contribution from the reflected appraisals of significant others. Inferred ratings are generally used in two circumstances; firstly, when it is thought that the individuals being tested are unable to respond for themselves, for example, more elderly individuals or those suffering from learning difficulties (see Carsrud, 1986); secondly, when it was thought necessary to provide an adjunct to self-report ratings. Marsh and O'Neil (1984) found that contrary to other research in this area (e.g. Shrauger and Schoeneman, 1979) their studies have shown that there are 'remarkably good correlations between self-concepts as rated by the subject themselves and self-concepts as inferred by significant others' (p. 167).

The chosen measurement method for this study was governed by three criteria. First, in order that the proposed model for academic self-concept in science be credible the validity testing of the model needed to be robust and the model was required to be generalisable beyond the research sample. This necessitated quantitative data of sufficient sample size to provide secure statistical testing. Second, the relationship of self-concept with other important variables, such as, age and gender was to be explored which again required quantitative data. Third, the model of self-concept which lays claim to possess the strongest validity is the Shavelson/Marsh model. The measurement instrument used most extensively to test this model has been and remains to be the SDQ (Marsh, 1992a, 1992b, 1992c)

measurement scales. If the model of academic self-concept in science reported here is to provide a realistic extension to this Shavelson/Marsh model then it would seem most appropriate to measure self-concept and test the model using an instrument which is consistent in design to the SDQ instrument. For these reasons this research used a self-report self-concept instrument which utilized a five point Likert scale response mode.

Instrument Design

General structure of questionnaire

Model 4 was the most comprehensive of the proposed models with all other six models being a subset of Model 4. Model 4 was the most data hungry of the models and as such became the template against which the appropriacy of the range and scale of the data collection instrument was judged. In Model 4 each of the rectangular boxes represented a manifest (or indicator) variables and an item pair was chosen to inform each manifest variable.

The final format chosen for the instruments was a self-completion questionnaire. Two separate instruments were designed and completed, one for Key Stage 3 and one for Key Stage 4. This was necessary since some of the questionnaire items needed to relate specifically to the curriculum content of the Key Stage for its pupils and thus required a number of the items to be different between the two instruments. The instruments were written in parallel to maintain as much commonality as possible. In the end out of 60 items, only five were significantly different between the two.

Each questionnaire item was constructed from two component phrases; a prefix or question hanger component and a science content component. The choice of prefix components was influenced by, but not identical to, Marsh's (1990) ASDQ instrument. The protocol was to include items written in (i) absolute terms, (ii) internal terms, i.e. relative to other subjects and (iii) external terms, i.e. relative to other learners. These different forms take account of Marsh's internal/external revision.

The general pattern adopted for the questionnaire structure was this. For each of the five latent variables twelve questionnaire items were written and these twelve items were written at two levels of specificity. The pattern can be explained through an illustration of one of the latent variables.

The biology latent variable was informed by the items thus;

- *The lower specificity items* focused on the term 'biology';

    - 2 items written in absolute format, e.g. *biology is difficult for me,*
    - 2 items written in internal format, e.g. *I do better at biology than my other subjects,*
    - 2 items written in external format, e.g. *I do better at biology than others in my class.*

    Of the six items, half of them were written with a positive orientation and half with a negative orientation. This gave one positive sense and one negative sense item in each of the three formats. The top example of the three statements above is written in a negative sense. This ensured that the questionnaires, when being completed, did not have all the positive responses at one end of the scale.

- *The higher specificity items* focused on the biology curriculum. These six items would provide information specifically about the individuals' self-concept in relations to the actual curriculum content. Because each of the six items was individual and distinct it was not thought appropriate to write them in the different absolute, internal and external formats as with the items at lower specificity. For this reason all the items were written in absolute terms. However, half of the items were written with a positive orientation and half with a negative orientation. This is consistent with the lower specificity items above and this resulted in;

    - 3 items written in a positive sense, e.g. *I usually do well at understanding systems of the human body*
    - 3 items written in a negative sense, e.g. *I have poor knowledge about living things in their environment.*

Establishing the Science Content Component of Each Item

The instrument items were written at three levels of specificity and in total there were 60 items.

Level 1 the lowest level of specificity containing the word 'Science', (6 items);
Level 2 the middle level containing the four facets of Physics, Chemistry, Biology and Science Investigations, (6 x 4 = 24 items)[2]
Level 3 the highest level of specificity, consisting of statements which disentangle each of the Level 2 facets plus statements of the Nature of Science, (6 x 5 = 30 items).

The term 'science investigations' was chosen to represent the curriculum content relating to the procedural domain described by the National Curriculum Programme of Study (DfEE, 1999) for Scientific Enquiry. The term 'science investigation' was a classroom friendly label with pupils well tuned into what is understood by carrying out an investigation. Nature of Science did not appear in Level 2 as there was no appropriate generic term for what pupils could understand as *ideas relating to the nature of science*.

A provisional list of science content categories was arrived at after close reference to the National Curriculum for Science (DfEE, 1999), the Qualifications and Curriculum Authority schemes of work for science (QCA, 2000), and a number of GCSE science examination specifications (e.g. AQA, 2002). The science content items were refined after consultation with a panel of science teachers and Heads of Science from ten secondary schools together with a small group of science teachers undertaking Masters' degrees in Education at the University of Manchester. The process progressed iteratively until there was widespread agreement on the final science content components. The final list of items for the two key stages can be found in the next chapter in Tables 4.4a and 4.4b.

[2] 24 items were used in the pilot study, although as will be explained below, in the final instrument only 12 items were included at this level of specificity. The 'internal' and 'external' items for Physics, Chemistry and Biology were eventually dropped.

Question hangers

Five equivalent prefix statements were eventually chosen for each category of question, i.e. Absolute statements written positively and absolute statements written negatively; internal statements written positively and negatively; external statements written positively and negatively.

The choice was made after developing many more statements than was needed for each category and analysing these statements to test their internal consistency. This was done by administering all the statements through a questionnaire to approximately 100 pupils. The internal reliabilities of the statements were measured to determine Cronbach's α score and the five most internally consistent statements in each category were selected. The alpha scores were typically in the region of 0.85 – 0.90. This was thought strong enough to suppose that all the generated statements could be treated as equivalent items. In compiling the questionnaire items the prefix components were matched randomly to the science content components.

Although the use of non-identical items provides a possible source of methodological weakness, it was actively preferred to the alternative. The situation of having differently worded items was thought important in order that the repetition during questionnaire completion was reduced to a minimum. Had this not been carried out then the 60 items of the instrument would have been built upon only six differently phrased items. There was a balance to be found between the loss of validity through the use of differently worded items, and the loss of validity due to poor completion through lack of engagement. Hopefully this balance point was found in order to satisfy the two. The problem was also mitigated somewhat by the use of item-pairs which will be discussed directly below.

The validity issue was raised because of the assumption of equivalence between the differently worded statements. This assumption of equivalence does provide a possible challenge to the measurement validity, although this challenge was mitigated in three ways. First, items were never utilised singularly; the smallest unit of measurement was always the item-pair. The use of item-pairs, as will be discussed further below, is a well practiced technique which has been shown to mitigate against the effects of 'idiosyncratic wording of individual items' (Marsh and O'Neil, 1984, p. 157). Second, the vast majority of analysis was conducted at the

level of 'factors' or latent variables rather than at the level of item pairs. Each factor was constructed from either three item pairs or four item pairs comprising six or eight items in total. The final statements for the questionnaire were chosen from the *statement pool* in a way to maximise the homogeneity of question hangers across the different factors. The statements were chosen such that no less than three of the possible five statements from each pool were chosen for each of the factors. Each factor used a different three statements from the pool of five statements ensuring that there was significant commonality in the items employed to build each factor and adding to the high levels of comparability. Third, although pre-pilot items were selected on the basis of similar means, standard deviations and high alpha levels, in nearly all subsequent analysis, direct comparison of means of individual items was avoided. In the small number of cases when direct comparisons were made the statements carried this caveat.

Piloting

The questionnaire was piloted with 250 pupils. Teachers administering the questionnaires reported that the terms Physics, Chemistry and Biology caused some confusion amongst younger Key Stage 3 aged pupils who frequently asked the teacher to explain the meaning of the words. This was thought to reflect curriculum structures where teacher talk of physics, chemistry and biology topics/subjects to Year 7 and 8 pupils occurs less frequently in some schools. As a result of this, and in order to minimise its effect, 12 of the 18 items relating to the terms Physics, Chemistry and Biology were removed from the questionnaire. The Internal and External items were removed retaining only the six absolute items. The six items relating to science investigations were left unchanged.

Data Collection

In compiling the final questionnaire instrument 18 additional items were included. These additional items were made up of eight ASDQ 'science items', six ASDQ 'general school' items, taken from Marsh's academic self-concept questionnaire and four new 'importance of science' items. All 48 items for KS3, the 48 items for KS4 and the additional 18 items can be seen in the complete questionnaires which are located in Appendices 2 and 3. The ASDQ science items were included for reasons of concurrent validity. High correlations between the ASDQ items and the Level 1

science items would provide additional evidence of validity. There was also opportunity to compare the test-retest statistics of the new instrument with that of the ASDQ items as means of a comparative measure. The 'general school' and the 'importance of science' items were included to aid further lines of enquiry and will not be discussed as part of this present study.

The questionnaire response mode was in the form of a five point Likert rating scale where the pupils reported their level of agreement from 1 = Totally true for me to, 5 = Totally False for me. In the final coding, all negatively worded items were reversed such that a low score always represented a high self-concept rating.

Item Pairs

The next stage required the construction of item pairs. Each of the questionnaire items was combined with a similar item to create an item pair. This has been an often used technique and underpinned the data processing procedures of the SDQ instruments. The rationale for the use of this technique has been set out by Marsh and O'Neil (1984) and has been incorporated into many self-concept scale analyses since this time (see Marsh, 1992a, 1992b, 1992c). As explained by Marsh and O'Neil, the advantages of creating of item pairs are that the:

- ratio of the number of subjects to number of variables is doubled;
- reliability of items is increased;
- unique variance of each component is decreased;
- scores are less affected by the idiosyncrasies of the item wordings.

There are drawbacks with the use of item pairs which Marsh and O'Neil explain as being associated with:

- a loss of information about individual items;
- a necessity that items paired must be homogeneous with respect to the dimension of measurement;
- varying parameter estimates and factor scores depending on the choice of pair.

None of the drawbacks highlighted by Marsh and O'Neil are worrying in this context. The vast majority of analysis carried out within this research project was at the level of 'factors' so negating most aspects of the first concern. Items of close or near matching context were paired, thus satisfying the homogeneity requirement. Pairs were chosen after the model structure had been designed and pairings were chosen to fall within the same a priori factor thus ensuring that factor scores were not affected. Items to be paired were chosen against the following three criteria.

The first criterion was to match and combine positively and negatively worded items. This served two purposes; (i) many of the items were originally written as pairs with a positive and negative orientation and hence this combination was a natural choice, (ii) matching a positive and negative item went someway to reducing the method effects associated with negatively worded items as reported by some researchers (see DiStefano, 2006; Motl et al., 2002). This ensured that all items pairs had a balance of a positive and negative component.

The second criterion was to match items written within the absolute, internal and external frames. This ensured that the contributions made by these different frameworks were not diluted by the combination process.

The third criterion was to select items from matching areas of content. This was a relatively straight forward task with the low specificity physics, chemistry and biology items, and even the high specificity items did not pose any significant problems. The final list of item pairs can be found in the next chapter in Tables 4.4a and 4.4b.

Before the process of item pairing was completed an analysis of the individual items as sole entities was undertaken. The arithmetic means, standard deviations, alpha scores for internal reliability, skewness and kurtosis of both individual items and factors were examined. This was undertaken before the uniqueness of each item was lost due to the pairing procedures. Completing this also served to check that homogeneity of the dimension of measurement, as was required above, was satisfied. The outcome of these tests will be discussed further below.

Selecting the Sample

Eight schools were chosen for the main study. Schools were chosen for their character and nature. Schools were selected to be a part of the study on the basis that they were representative of the different types of school found in England. In this way the participating pupils would have been drawn from schools with differences in pedagogic, social, geographic and economic contexts. These school types were; a fee paying independent grammar school, a science specialist school, a technology college, a Roman Catholic School, an urban comprehensive (in special measures), a suburban comprehensive and a rural comprehensive. Two schools were situated within a poor social/economic area and one school had a high proportion of ethnic minority pupils. All of these schools were situated in the North West of England. The primary principle of selection for participation occurred at the level of school. Following this, classes were chosen on the basis of convenience for those schools participating, with all pupils in a particular class being asked to take part in the study. Schools were instructed to select participating classes to be representative of their 'middle-band' pupils through an oral explanation.

The questionnaire was administered to 1487 pupils within Y7 to Y11 in the eight different schools, (Y7 n=277, Y8 n=307, Y9 n=347, Y10 n=249, Y11 n=307). Test-retest data were collected from 192 pupils. There was a gap of between four and eight weeks between test and re-test sessions. The questionnaire was administered by the regular science teacher within science curriculum time. The purpose of the questionnaire was explained to the pupils prior to their completion along with an explanation of the mode of response. Parental permission was sought by means of nil return of an op-out reply slip and all pupils had a further opportunity to opt out of completing the questionnaire on the day. Pupils were asked to write their names on the questionnaires but confidentiality was assured to the pupils.

Model identification

Model Identification was a set of procedures carried out to check that there was sufficient information available from the data to be able to undertake parameter estimation. Structural models can be classified as one of three types; just-identified,

over-identified, or under-identified. A just identified model would have the same number of variances and co-variances (i.e. known features) as there would be estimable parameters (unknown features). There would be a one to one correspondence which would lead to an uninteresting (valueless) solution since there are no degrees of freedom, (Byrne, 1998). An over-identified model would have more data points than estimable parameters and therefore would possess positive degrees of freedom. If under these circumstances the model provided a good fit for the data it represents a positive outcome since there was never any guarantee that this would be the case. This suggests that the model would indeed be a reasonable representation of the self-concept construct. An under-identified model would be one where the number of parameters exceeded the number of variances and covariances and there would be insufficient information to uniquely identify each parameter. This would lead to an infinite number of possible solutions, (Byrne, 1998) and would be an undesirable outcome. An analogy can be used to help clarify the notion of model identification. It is connected to the idea of drawing a straight line in the correct position. Two dots drawn on a blank page will exactly specify where the line is to be drawn. The two dots are equivalent to a 'just identified' model. Two dots are enough to specify the position, although there is no safeguard available to check that the dots are themselves accurately plotted. A straight line will always pass through two dots and therefore it is a not an interesting outcome. Three or more dots on the page are equivalent to an 'over-identified' model. Now we have an automatic safeguard. A straight line can only pass through all three dots if the dots are arranged in a linear fashion. The more dots placed on the page then the more opportunities there are for one of the dots to be out of alignment and the line not to be able to pass through all the dots. With 20 dots on the page we would not expect a straight line to pass through them all, however, if it nearly managed this feat then it is indeed a worthy outcome, even if it wasn't necessarily a perfect fit. One dot on the page is equivalent to an under-identified model. The direction of the line is not specified in any sense, and an infinite number of lines could satisfy the model requirements of passing through a single point. This then leads to a trivial and less valued solution.

There are formal algebraic methods for calculating whether a model is over-identified (which is the desired state), however, this procedure can be sticky for the non-expert and therefore alternative procedures have been introduced by Bolen (1989) with his 'rules of thumb' approximation. The "Three Measure Rule" states that a factor-analytic measurement model will be *identified* if every latent construct is associated

with at least 3 measures. Alternatively, the "Two Measure Rule" states that a factor-analytic measurement model will be identified if every latent construct is associated with at least 2 measures AND every construct is correlated with at least one other construct. These rules do not provide a definitive answer, but they provide clues to whether identification is secure or borderline. In this case all seven of the hypothetical models are likely to be *identified* by reference to Three Measure Rule since all the latent factors in the measurement model are associated with three item pairs.

Parameter estimation

Parameter estimation is the first stage of the modelling process that can actually utilize the empirical data. It therefore provides the first opportunity to evaluate the quality of the hypothesized models. The model structure must firstly be programmed into a data analysis program like LISREL 8. The software then generates an implied, (model generated) covariance matrix for the models under examination. This model generated covariance matrix is an hypothetical data set consisting of a matrix of predicted associations between the variables based entirely on the theoretical model structure. This matrix of model predictions can then be matched against the observed (actual) covariance matrix generated from the empirical data, (Hayduk, 1987). A positive outcome for a model would be for the estimated covariance matrix to be as close as possible to the covariance matrix derived from the empirical data.

LISREL 8 uses seven different methods to generate model parameter estimates. These are Instrument Variables (IV), Two-Stage Least Squares (TSLS), Unweighted Least Squares (ULS), Generalized Least Squares (GLS), Maximum likelihood (ML), Generated Weighted Least Squares (WLS) and Diagonally Weighted Least Squares (DWLS). Each of the methods has been reported in the literature as producing parameter estimates that are very close to 'true' parameter values, provided there are no specification errors associated with the model (Diamantopoulos et al., 2000). The methods fall into two groups; limited-information techniques and full-information techniques. The limited-information techniques estimate each parameter equation separately and without reference to information from other equations in the model. It proceeds non-iteratively, is fast and fairly robust against model misspecification. IV and TSLS are examples of this technique. All the other procedures are full-

information techniques, which makes the parameter estimates using information gathered from the entire model system. This procedure is statistically more efficient but because the estimation of each parameter is dependent on the information from every other parameter then it was more susceptible to errors from misspecification from any equation in the model.

Parameter estimates are generated iteratively commencing with software generated starting values from which parameter estimates were calculated for all eight of the hypothesized models. This generated covariance matrix was then compared against the matrix from the empirical data. The software generated a residual matrix from the difference between the two and used this residual matrix to improve on the starting values. The model process was run again to generate a new improved set of estimates from which a new residual matrix was generated. This iterative process continued until it was no longer possible to change the parameter estimates to bring the implied covariant matrix any closer to the sample covariant, i.e. the residual matrix was as small as possible. When this match had been reached convergence had been achieved for the model.

Maximum Likelihood (ML), which was a full-information, iterative method, was the default procedure in LIRSEL 8 and is widely used in SEM practice, (see Diamantopoulos et al., 2000). It has an advantage in that it produces a wide range of additional statistical information that could be used to test the extent to which the model is consistent with the data. ML is a robust procedure and provides good estimates even if there are small departures from the multivariate normality assumption.

The LISREL output contains a wealth of other information as well as the ML parameter estimates. The LISREL output is divided into sections containing various tables and matrices for evaluation. The first section contains three pieces of information for each parameter. These are, (i) unstandardised parameter estimate, (ii) its standard error, and (iii) the relevant t-value. Taking these in turn, the magnitude of the unstandardised parameter estimate shows the resulting change in a dependent variable from a unit change in an independent variable, with all other independent variables being held constant. The polarity of the sign indicates whether the relationship is directly proportional in a positive or negative sense, with a negative sign indicating a decrease in the dependent variable for an increase in the

independent variable. The meaning of the unstandardised parameter estimate is very similar to the regression coefficient in conventional regression analysis. Being that the parameter estimates are calculated in an unstandardised format they provide absolute magnitudes of effect and as such direct comparisons can be made with similar models using other populations, provided the nature of the scale remains consistent.

The standard error provides an indication of the precision by which the parameter has been estimated. Smaller standard errors are generally more welcome as they indicate that the parameters have been well estimated. Excessively small standard errors however, (i.e. approaching zero) are unwelcome as they mean the test statistic had not be defined. Excessively large standard errors are also unwelcome as these indicate that the parameters cannot be determined (Jöreskog and Sörbom, 1993). The t-values are used to test whether the parameters are significantly different to zero in the population. A value of the t-statistic between -1.96 and +1.96 indicates that the parameter is not significantly different at the 5% level. This means that for parameter evaluation a positive outcome is a t-value outside of these values, i.e. t-value > modulus 1.96. The t-values can be calculated from the equation (t-value = value of parameter / standard error).

Additional to this the LISREL output contains two further pieces of information for each of the equations. Firstly, estimates of error variances are shown. For the measurement part of the model they indicated errors in measurement and for the structural part of the model they indicated residual terms. The estimates of error variance also have their accompanying standard errors and t-values displayed. Secondly, the squared multiple correlations, R^2 are also shown. As before, R^2 shows the amount of variance in the dependent variable accounted for by the independent variable. For example an R^2 of 0.82 would mean that 82% of the variance in the dependent variable is explained by the independent variable. This is analogous to the R^2 statistic in conventional regression. The R^2 statistic gives an indication of how free the manifest variables are from measurement error. The closer the statistic is to unity then the more the manifest variables show that it was an indicator of the latent variable. Higher values i.e. greater than 0.3 are considered positive results.

The next section of the output contains a covariance matrix of independent variables. This matrix presents the variances and covariances of the exogenous (independent)

latent variables in the model together with their respective standard errors and t-values. There is also an additional output which gives the covariant matrix of all the latent variables, and not just the exogenous latent variables. The t-values indicate whether the relationships between the independent latent variables are significant. They would be shown to be significant if the moduli of the t-values are greater than 1.96.

Assessment of Model Fit

Having completed the model estimation the next stage was the assessment of model fit. 'Of primary interest in structural equation modelling is the extent to which an hypothesized model "fits" or, in other words, adequately describes the sample data' (Byrne, 1998, p.103). This usually involves, as with parameter estimation, an examination of the closeness of the covariant matrix of the model compared with the covariant matrix of the empirical sample. A better term might actually be, assessment of misfit, as the procedures usually estimate the amount of disagreement between the two matrices. The outcome of the procedure results in the production of a fit index. There is not a single 'best-fit' index and the safest way to proceed is to recognize that, at best, the tests provide one means of fit estimation yielding some information on the appropriateness of the model in relation to the evidence from the data. Crucially however, it is only in relation to the sample of data. The fit indices vary so much because their different mathematical procedures are more or less influenced by sample size, model complexity, violations of underlying assumptions (e.g. multivariate normality) or variable independence, (Byrne, 1998). Also, a model fit index tells nothing of the usefulness or the plausibility of the model. Model fit testing is purely a number crunching exercise carried out independently of the context from which the model emerged and separated from the theoretical underpinnings of the model or its conceptual or philosophical roots. In fact Sobel and Bohrnstedt (1995) cautioned that although such indices provide an objective measure of fit, if they are used as the primary criterion for adequacy then 'scientific progress could be impeded', (p. 185).

A more secure way of utilizing model fit testing is in the spirit of Popper's falsification principle (Popper, 1963) in so much that 'success' is a rejection of an hypothesis, and theory can only be disproved, or verified, and never proved. The fit indices tells

something more definitive about the model when the fit index is low, i.e. the model can be rejected. If the fit index is high, it communicates that the model cannot be rejected on the basis of fit alone between model projections and empirical data. However, there might be alternative models for which the fit indices are just as good or indeed better. There will always be ambiguity in the interpretation of fit indices and they remain one tool, albeit a powerful tool, by which the researcher can make judgments as to the strength and appropriateness of the hypothesized model.

In working to establish goodness of model fit, Diamantopoulos et al., (2000) recommend the following sequence,

1. assessment of the model's global fit,
2. assessment of the measurement part of the model,
3. assessment of the structural part of the model.

Beginning with an assessment of global fit, the Chi square (χ^2) is the first statistic presented in the LISREL output. Its value is calculated by the use of formula $(N-1)F_{min}$ where N was the sample size and F_{min} was the minimum fit value function. It is in other words a test of perfect fit. Chi square has been the traditional measure for evaluating overall model fit between the unrestricted sample covariance matrix and the restrictive covariance matrix, (Byrne, 1998). The Chi square statistic is simply the test of probability of the null hypothesis, (i.e. that the model perfectly fitted the sample data). Unlike traditional statistical methods however, a non-significant Chi square result supports the adequacy of the hypothesis, and therefore indicates a fit between model and data has been achieved. Researchers are therefore looking for a non-significant outcome so as not to reject the null hypothesis since a statistically significant Chi square would cause rejection of the null hypothesis indicating a poor fit between model and data.

The Chi square statistic seems, on the face, to be a powerful and objective measure, however, in practical applications in real world situations model fit has proved to be more complex than a determination of whether a model's Chi square statistic is significant or not. The Chi square statistic has proved to be particularly sensitive to sample size. There is an increasing probability that the Chi statistic will reject the test model as the sample size increases (Bearden, Sharma and Teel, 1982). Chi square

has also been shown to be sensitive to (i) model complexity, (ii) degree of misspecification and (iii) assumption violations, (ibid). This has provided a tension between methods and assessment. SEM is based within asymptotic theory (reliant on large samples) whilst Chi square is disproportionately harsh on large sample data which mitigates against its use as a dichotomous test instrument used in this way (Hu and Bentler, 1995). The assumption that any model can fit perfectly is also extremely unrealistic, as it is improbable that any model being developed is anything other than an approximation. Some degree of misfit is therefore inevitable (MacCallum, 1995). In fact Jöreskog and Sörbom (1993) have suggested that Chi square should be thought of as a *badness* rather than goodness of fit statistic. That is, it is not so much a test statistic that is passed or failed, but more a measure of fit, (Jöreskog and Sörbom, 1993). The question should not be, *is the model a perfect fit?* A more realistic question to ask is how well does the model adequately represent the sample data (Browne and Cudeck, 1993)? The Non-Centrality Parameter (NCP, λ) attempts to address this question. Essentially the NCP measures the discrepancy between the population covariance matrix and the covariance matrix implied by the model. The greater the discrepancy between the two matrices then the greater will be the value of λ. Wheaton et al., (1977) proposed as an alternative to the straight Chi square statistic, a ratio test of (χ^2 / degrees of freedom). This has the added advantage that it takes account of model complexity (through the division by 'degrees of freedom'), although the issues of sample size remain because it is still inherent in the chi-square value. None the less, this has become a widely accepted test statistic with a value of between 2 and 3 being deemed acceptable.

In an attempt to overcome the particular limitations of the Chi square statistic a host of alternative model fit procedures have emerged. These newer procedures however, are themselves not without problems. LISREL 8 produces a total of 15 different goodness-to-fit measures and, within the literature, there has been a steady move away from the use of a single fit index towards multiple checks. The first of the alternative statistics presented by LISREL is the Root Mean Square Error of Approximation (RMSEA). The RMSEA is regarded by many as one of the most informative fit indices, (e.g. Browne and Cudeck, 1993; MacCallum, 1985). The RMSEA asks the question, 'how well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?' (Browne and Cudeck, 1993, p.137). The RMSEA discrepancy is measured per degree of freedom which makes it sensitive to model complexity. RMSEA values

below 0.1 are considered a reasonable fit and values below 0.05 are considered a very good fit with a cut off point being proposed at 0.08, (Browne and Cudeck, 1993; MacCallum, Browne and Sugawara, 1996). LISREL also presents the 90% confidence interval for the RMSEA. This is a useful perspective on the use of RMSEA. A wide confidence interval would indicate imprecision in the RMSEA value which most likely would cancel out the positive indication of a low RMSEA value itself. By contrast a narrow confidence interval would indicate much greater precision of the RMSEA in reflecting the model fit to the population (MacCallum, 1996) strengthening the evidence not to reject the model. The size of the confidence intervals however, are susceptible to sample size and model complexity, with small sample sizes and large numbers of parameter more likely to produce wide confidence intervals, (Byrne, 1998). LISREL also reports the *p*-value for closeness to fit for the RMSEA < 0.05. This tests the hypothesis that the error of approximation has a probability of less than 0.05. This is a 'significance test' which the researcher wishes to fail. The higher this value then the greater is the chance that the fit is not rejected. Jöreskog and Sörbom (1993) have suggested that the *p*-value should be greater than 0.5 to represent a good fit, with better values rising toward unity.

An alternative to the RMSEA is the Expected Cross-Validation Index (ECVI). This was proposed by Browne and Cudeck (1989) and is a measure designed to assess, in a single sample, the likelihood that the estimate model will be replicated with similar sized samples drawn from the population, i.e. is there predicted cross validation? At the heart of the index is a measure of the discrepancy between the covariant matrices of the analysed sample and that which would be expected in another sample of equivalent size. Hence this is a useful indicator of the model's overall fit. The ECVI is a relative, rather than absolute, measure and it assumes a comparison of other ECVI indices from alternative models under test. These alternative models could be competing hypothetical models, although in practice, alternative models are the independence model and the saturated model generated from the model under test automatically by LISREL. The independence model is a null-model where all observed variables are uncorrelated and represents the most restricted model. The saturated model is one where the number of estimated parameters equals the number of data points (c.f. just justified model) and is the least restricted. Therefore the ECVI value of the hypothesized model can be compared with those of the independent and saturated models. A good fitting model would have an ECVI value smaller than either of the comparison models. According to

Browne and Cudeck (1989) the index is a measure of the models relative predictive validity when compared against other models. The 90% confidence interval for the hypothesized model's ECVI value can also be used as a supporting additional comparison.

LISREL 8 also computes a number of indices which take account of model parsimony. Parsimony is the principle in which if a choice must be made between two models which are seemingly similar in other respects, then the preferred model is the one which is least complex. The first two of these indices were the Akaike's Information Criterion (AIC) and the Consistent Akaike's Information Criterion (CAIC). These indices are known as information criteria where the assessment of model fit takes account of the number of estimated parameters. The difference between the two is that CAIC adjusts for sample size effects. The AIC and CAIC are relative fit indices and are operationalised in a conceptually similar manner to ECVI where comparisons are made between two or more models with the smaller values representing the better fits.

The root mean square residual (RMR) is a measure which is based on the residual matrix, i.e. the difference between the model-implied (fitted) covariance matrix (Σi) and the sample covariance matrix (S), i.e. ($S- \Sigma i$). A good model will have small fitted residuals in comparison to the elements in S. The RMS index is calculated from the root mean square of the fitted residuals. Standardized RMR can sometimes provide a better estimate as this is based on standardized residuals which are not prone to variations relating to measurement size. Values of standardized RMR < 0.05 are considered to represent good fits (Byrne, 1998).

Whereas most of the indices considered up to now have been relative fit indices, there are a number of measures which produce an absolute fit index value. The Goodness-of-fit index (GFI) is a particularly strong example. The GFI algorithm is not based on a comparison between the sample matrix and baseline matrix rather it is a measure of the amount of variance and covariance accounted for by the model and as such gives an indication of how well the model performs in perfectly reproducing the covariance matrix of the empirical data. The parsimony goodness to fit index (PGFI) addresses the issue of the model complexity and adjusted goodness to fit index (AGFI) takes account of the number of degrees of freedom. Both of these latter measures tend to produces fit indices which are more modest. Acceptable

values of GFI are > 0.9 and PGFI > 0.5, (Mulaik et al., 1989). According to Diamantopoulos et al., (2000), the GFI index tends to give the most reliable measure of absolute fit in most circumstances.

Another family of relative measure indices is the non-normed fit index (NNFI) and the related measures PNFI and CFI. As with other algorithms reviewed previously the parsimonious and comparative fit versions usually produce more modest indices. Of the three indices in the family it is the NNFI which provides the most robust measure and is often referred to as the Tucker-Lewis Index (TLI).

The final fit measure produced by LISREL 8 is the Critical N statistic (CN). This measure is conceptually different from all other measures. Whereas in many of the previous measures sample size was either accounted for within the index, or problematic for the index, the Critical N provides an indication of the minimum sample size required by the model in order to produce an acceptable statistic. Many researchers (e.g. Byrne, 1989) quote CN > 200 as indicative of a model that adequately represents the sample data.

As can be seen, there is a huge choice facing the contemporary researcher using SEM techniques. Unfortunately, there is not widespread agreement as to which indices to use. Jaccard and Wan (1996) recommends that one index from three different categories be employed, whilst Kline (1998) recommends using at least four separate tests. Diamantopoulos et al., (2000) feels that for most practical purposes researchers should employ the Chi-square test in conjunction with the RMSEA, ECVI, standardized RMR, GFI and CFI indices to assess the overall model fit. What is important is that the tests are applied appropriately and consistently, with indices not being cherry picked because they reveal favourable outcomes, and at all costs the temptation to proceed on a fishing expedition to see what can be found must be avoided. Marsh, Balla and McDonald (1988) following their substantive investigation of more than 30 goodness-to-fit indices they concluded:

> On the basis of our research, we recommend at least one of the FFI2, LHRI2, χ^2/I2, TLI, and CAKI2 indices, as well as the examination of parameter estimates in relation to substantive issues and the examination of residual covariances. (p.408).

As implied by Marsh et al., (1988) in the second half of the quotation, the overall fit is only one piece of the jigsaw, albeit an important piece. Although measures of overall fit give us a strong steer on whether the model is to be rejected and how much the model misfits the data, it does not reveal anything about which parts the model or how parts of the model are inappropriate. To answer these questions we need to assess the measurement model and the structural model separately.

The assessment of the measurement model is an evaluation of the reliability and validity of the constructs. This is done in three ways. The first is to examine the magnitude and significance of the paths between each latent variable and its indicators. Factor loadings are significant at the 0.05 level if the t-value is greater than modulus 1.96 and the error variances being small but not approaching zero. These criteria were discussed as part of the parameter specification procedures. Secondly, the squared multiple correlation (R^2) of the indicators can be reviewed. High R^2 values indicate high reliabilities of the indicators as it means that a large proportion of the variance of the indicator is explained by the underlying latent variable. In other words, the unexplained variance (measurement error) is low. Thirdly, a composite reliability for each latent variable can be calculated. Unfortunately, LISREL 8 does not carry out this procedure automatically so the calculation needs to be performed manually. The formula below is employed.

$$\rho_c = (\Sigma\lambda)^2 / [(\Sigma\lambda)^2 + \Sigma(\theta)] \qquad \text{Equation 3.1}$$

where ρ_c = composite reliability

λ = indicator loadings

θ = indicator error variances (variances of the δ's and ε's)

Σ = summation over the indicators of the latent variables

A value for $\rho_c > 0.6$ is seen to be a desirable outcome (Bagozzi and Yi, 1988).

A second measure can also be calculated, which is the average variance extracted, ρv. This is 'the amount of variance that is captured by the construct in relation to the amount of variance due to the measurement error.' (Fornell and Larcker, 1981, p.45).

This value can be calculated by using the formula:

$$\rho_v = (\Sigma\lambda^2) / [\Sigma\lambda^2 + \Sigma(\theta)] \qquad \text{Equation 3.2}$$

A positive result is for $\rho_v > 0.5$, as this indicates that the underlying latent variables account for a greater proportion of the variance than the does the measurement error.

The assessment of the structural model sets out to test the nature and appropriateness of the relationships between then exogenous and endogenous latent variables. This is to test whether the proposed construction of the conceptual model (as described above) is borne out by the data and is achieved in three ways. First, the direction/polarity of the path linkages between the latent variables, as revealed by the empirical data, must be consistent with the original model, i.e. should all be positive in this instance. Second, the relationships between the latent variables should be strong, and at the very least should be significant. This is tested by inspecting the t-values which should have a value greater than modulus 1.96 to be significant at the 0.05 level. Third, that the amount of variance (as measured by R^2) of the endogenous latent variables accounted for by the exogenous latent variable should be high. This will show that the model has high explanatory powers. The standardized parameter estimates which are output by LISREL, give a clear indication of the relative impact that each independent variable has on each dependent variable. If standard parameter estimates are used as opposed to the original parameter estimates then the direct comparisons can be made as there are no issues associated with measurement scale. These data provide additions insights into the model structure and its operation.

Model Modification

It is highly unlikely that any 'true' model can exist to represent a population or indeed a sample (Browne and Cudeck, 1989). Therefore all models will have some degree of misfit. In order to address this situation *model modification* can be undertaken which attempts to make post priori changes. Model modification must be carried out with the utmost caution. The only route open to the researcher is a retrospective

adjustment of the conceptual model on the basis of the interpreted data. Adjustments are made to the model and then the model re-evaluated to establish if the adjustments have been beneficial in creating a better fit to the data. In this way the procedures being undertaken have now ceased to be confirmatory in nature and have become exploratory. The danger, and it is a real and substantial danger, is that the more the model is modified on the sole basis of the sample data and becomes aligned to this data, then the greater is the possibility that the modified model becomes 'susceptible to capitalization on chance in that the idiosyncrasies of the sample may influence the particular modifications that are performed.' (Diamantopoulos et al., 2000).

Bearing in mind this cautionary message, there are two primary procedures that may be undertaken on the model following an examination of the LISREL output data. LISREL produces two types of information relating to model misspecification. The first relates to 'residuals' and second to 'modification indices'.

As was rehearsed earlier, the residuals relate to the difference between the hypothesized model and the data covariance matrices. In LISREL the residuals are presented in a number of forms. There is a presentation of, (i) standardized residuals where values greater than 2.58 are considered large, (ii) lowest, highest and median residual value, (iii) stem-leaf residual plots. If the model is well-fitting then the stem-leaf plots will be symmetric with the residuals clustered around the zero, with most of the plots lying towards the centre and fewer lying towards the tails. A skewed distribution shows that the model has either systematically over estimated or systematically under estimated the covariances. Large positive residuals are indicative of an underestimated model which can be corrected by creating additional paths, usually achieved through the freeing of parameters. By contrast large negative residuals are indicative of an overestimated model and the model should be modified by deleting paths. This is achieved through the fixing of parameters relating to the appropriate covariance. This is necessary when the model has been overfitted. If there are high numbers of large residuals it is not clear from the stem-leaf diagrams how the model can best be modified, under these circumstances the modification indices can provide additional assistance. In LISREL, modification indices are calculated for every parameter set to zero. Modification indices (MI) are univariate versions of a Lagrange Multiplier which is a χ^2 statistic with one degree of freedom. The modification index 'approximates the amount by which the model's

overall χ^2 would decrease if a particular parameter were freely estimated' (Kline, 1998). This means that the greater the MI, then the more improvement the model would show if that parameter was added to the model. An MI > 3.84 is considered large, as this is the value for χ^2 with one degree of freedom to exceed critical significance at the 0.05 level. However, the freeing of parameters can never proceed in a mechanical way purely guided by the modification indices. Changes can only be made to the model specification if there is conceptual integrity in doing so. This process of modification should be done iteratively and singularly, as freeing one parameter may well have a knock-on effect and change other MIs making it less necessary that they should also be changed. Specification change should never be done purely on the modification indices alone. MI should always be used in conjunction with the Expected Parameter Change (EPC). The EPC is part of the LISREL output and shows the predicted changes for the fixed parameter. If the MI and the EPC are both excessively high and the change makes substantial (conceptual) sense then there may be a positive benefit from carrying out the procedure.

LISREL also produces a graphical output of the normal probability of the residuals. A perfect output would show a series of plotted points running in a vertical line parallel to the y-axis, whilst the worst possible plot would show a distribution of residuals running in a horizontal line running parallel to the x-axis. An acceptable plot would be a diagonal distribution following the y = x orientation. The steeper the gradient the better is the indicated fit. In all cases a non-linear plot is cause for concern as it indicates separation from normality, specification errors in the model or nonlinearity.

Cross-validation

The final stage of the methodology was to undertake a cross validation exercise. The data set was large enough for it to be separated into its subgroups and still retaining data sets large enough to operate successfully under SEM analysis (>200); hence the validation exercise was carried out using those different subsets of data. This served two purposes; first it was possible to check that the model was stable over two independent data samples thus providing evidence of its construct validity and generalisability to a wider population. Second, it allowed a check on parameter invariance across gender subgroups. This was particularly important as comparisons

were going to be made about the differential responses of gender and age related groups.

The first test was carried out by comparing the performance of each model in a tight, moderate and loose replication strategy. In these tests the two subgroup samples were fitted to the model simultaneously with various conditions imposed upon the fitting of the model parameters. In the tight replication strategy, all parameters were assumed equal across both groups which provided the most difficult conditions for model fit. In the moderate and loose replication strategies, certain parameters were allowed to vary by removing invariance constraints. The fit indices and a difference of Chi-square significance tests were carried to gain a judgment of the models' differential performance under hardening conditions of invariance constraints. The best performing model would have the least difference in fit characteristics as the replication strategy testing became more severe.

RESULTS AND ANALYSIS

The results presented within this chapter were collected through three empirical phases of the research. These phases were:

- Phase One: Pre-pilot
- Phase Two: Pilot
- Phase Three: Main study

Phase One: Pre-pilot

The pre-pilot stage comprised the generation and testing of the question hangers for each of the questionnaire items. The questions hangers consist of the first half of each questionnaire statements, and give each item its context. Recall from the previous chapter that the question hangers assigned to the statements their absolute, internal and external frame of reference. Each item within the pilot questionnaire asked the participants to issue a response in relation to 'science'. The final questionnaire by contrast, directed the participants to respond to the different *facets* of science.

This outcome of the pre-pilot procedure produced a bank consisting six groups of statements. Each group contained five statements from which the full questionnaire items were constructed. The groups consisted of a collection of differently worded, but associated, statements, and although they were not identical, they were conceived as being 'equivalent statements' for the purposes of questionnaire construction. The rationale for this decision was rehearsed in the previous chapter.

The results of the pre-pilot can be seen below. Before any statistical calculations were performed, incomplete entries were removed through pairwise deletion. The full list of statements can be seen in Table 4.1a. The pilot questionnaire contained 47 statements and was administered to 96 individuals. From the 47 possible statements 18 were eventually selected and used. The selection was made on the basis of their means, standard deviations and Cronbach alpha scores. These data can be seen in Table 4.1b.

Table 4.1a Pre-pilot Statements

	Absolutes
AP1*	**I am good at understanding about science**
AP2*	**I usually do well at science**
AP3*	**I am pretty good when it comes to science**
AP4*	**I have no trouble in learning about science**
AP5*	**I have good knowledge of science**
AP6	I can easily learn news things about science
AP7	I always seem to get the right answers in science
AP8	I feel happy with what I know about science
AN1*	**I have difficulty in understanding about science**
AN2*	**I usually do poorly at science**
AN3*	**I am pretty bad when it comes to science**
AN4	I have lots of trouble in learning about science
AN5*	**I have poor knowledge of science**
AN6*	**I have difficulty learning new things about science**
AN7	I always seem to get the wrong answers in science
AN8	I feel bothered by how little I know about science
	Internally referenced
IP1*	Science is my strongest subject science
IP2*	**I am better at science than my other subjects**
IP3	I find it easier to learn about science than my other subjects
IP4*	I get higher marks in science than most subjects
IP5*	**I enjoy science much more than my other subjects**
IP6*	I look forward to science more than my other subjects
IP7	I worry about my other subjects more than I worry about science
IP8	Science is my favourite subject
IN1*	**Science is my weakest subject**
IN2*	I am worse at science than my other subjects
IN3*	I find it harder to learn about science than my other subjects
IN4	I get lower marks in science than most subjects
IN5*	**I enjoy my other subjects much more than science**
IN6	I look forward to science less than my other subjects
IN7	I worry about science more than my other subjects
IN8*	Science is my least favourite subject
	Externally referenced
EP1*	I am one of the best in the class at science
EP2*	**I am better at science than my friends**
EP3*	I am usually quicker than my friends in learning about science
EP4*	It seems easier for me to learn science than my friends
EP5	I have a clearer understanding than my friends about science
EP6	My friends usually have to ask me when we do science
EP7*	**I know more about science than my friends**
EN1	I am one of the worst in the class at science
EN2	Others in my class are much better at science than me
EN3*	My friends are usually quicker than me in learning about science
EN4*	It seems harder for me to learn science than my friends
EN5*	**My friends have a clearer understanding than me about science**
EN6*	I usually need to ask my friends when we do science
EN7*	**My friends know more about science than me**
IN8*	Science is my least favourite subject

Table 4.1a contains the full set of question hangers from which the final questionnaire statements were selected and constructed. Items marked with an * were chosen to be used in the pilot study. Following the Pilot, some items were removed. Only items shown in bold were eventually used within the Main Study. An explanation of the coding procedure for items in described in Appendix Two.

Table 4.1b Reliability Analysis Scale (Alpha) for Pre-pilot Items specified in Table 4.1a

	Mean	Std Dev	Corrected item Total correlation	Alpha if Item deleted
AP1 *	**2.6225**	**0.9454**	**0.7706**	**0.9033**
AP2 *	**2.8052**	**1.0869**	**0.8085**	**0.8995**
AP3 *	**2.8053**	**1.0411**	**0.8108**	**0.8994**
AP4 *	**2.8417**	**0.9812**	**0.7334**	**0.9060**
AP5 *	**2.7699**	**0.9791**	**0.7302**	**0.9063**
AP6	2.9154	0.9902	0.6486	0.9128
AP7	3.1109	0.8973	0.6796	0.9104
AP8	2.5972	1.1028	0.6428	0.9144
No. of Cases = 83				
Alpha = 0.9173 Standardized item alpha = 0.9179				
AN1 *	**2.6796**	**1.0403**	**0.7379**	**0.8820**
AN2 *	**2.4091**	**1.1631**	**0.6953**	**0.8857**
AN3 *	**2.6178**	**1.1714**	**0.6932**	**0.8859**
AN4	2.4334	1.1215	0.7628	0.8791
AN5 *	**2.3833**	**0.9502**	**0.6920**	**0.8867**
AN6 *	**2.6671**	**1.0999**	**0.7707**	**0.8785**
AN7	2.5567	0.9814	0.6251	0.8919
AN8	2.6183	1.2330	0.5364	0.9021
No. of Cases = 82				
Alpha = 0.8994 Standardized item alpha = 0.9017				
IP1*	3.6420	1.0577	0.6868	0.8014
IP2*	**3.4943**	**0.9953**	**0.7264**	**0.7970**
IP3	3.3585	0.9069	0.5784	0.8171
IP4*	3.3338	1.0304	0.6598	0.8055
IP5*	**3.9381**	**1.0464**	**0.6020**	**0.8133**
IP6*	3.8142	0.9042	0.6961	0.8032
IP7	3.1728	0.9660	0.1397	0.8672
IP8	4.0122	1.1166	0.4968	0.8285
N of Cases = 82				
Alpha = 0.8372 Standardized item alpha = 0.8380				
IN1*	**2.6707**	**1.3349**	**0.7734**	**0.7604**
IN2*	2.9886	1.2147	0.5121	0.8089
IN3*	3.0865	1.1176	0.6666	0.7843
IN4	2.6469	1.1194	0.5534	0.8021
IN5*	**3.9505**	**1.1251**	**0.5760**	**0.7986**
IN7	2.5731	1.0596	0.3143	0.8360
IN8*	3.2570	1.3688	0.5790	0.7988
No. of Cases = 83				
Alpha = 0.8235 Standardized item alpha = 0.8207				

Items marked with an * were selected to be used in the pilot study. Following the pilot, some items were removed, and only items shown in bold were used within the main study.

Table 4.1b continued overleaf

Table 4.1b (continued) Reliability Analysis Scale (Alpha) for Pre-pilot Items

	Mean	Std Dev	Corrected item Total correlation	Alpha if Item deleted
EP1*	3.3781	1.0434	0.6844	0.8631
EP2 *	**3.1966**	**1.0294**	**0.6585**	**0.8665**
EP3 *	3.2075	0.9466	0.8201	0.8465
EP4 *	3.3419	0.9268	0.6857	0.8633
EP5	3.1218	0.9291	0.6592	0.8664
EP6	3.3412	1.1067	0.5012	0.8887
EP7 *	**3.2807**	**0.9406**	**0.7153**	**0.8596**
No. of Cases = 83				
Alpha = 0.8821 Standardized item alpha = 0.8855				
EN1	2.2651	1.0764	0.4765	0.8584
EN2	3.3623	1.1149	0.5442	0.8499
EN3*	2.8800	1.1761	0.6221	0.8393
EN4*	2.6145	1.1495	0.6917	0.8290
EN5*	**3.0128**	**0.9631**	**0.6208**	**0.8405**
EN6*	2.9396	1.2739	0.6378	0.8377
EN7*	**3.0606**	**1.1124**	**0.8015**	**0.8132**
No. of Cases = 84				
Alpha = 0.8586 Standardized item alpha = 0.8593				

Items marked with an * were selected to be used in the pilot study. Following the pilot, some items were removed, and only items shown in bold were used within the main study.

The Cronbach alpha values of internal reliability for the groups of statements were calculated and found to be between 0.917 (for the Absolute Positive statements) and 0.824 (for the Internal Negative statements). The high alpha scores indicated that the items were potentially measuring the same underlying construct and that the different items had a strong likelihood of eliciting consistent and reliable responses to each another. The items to be selected for use in the next stage of the research were selected by paying attention to their high Alpha scores, their closeness of means and their standard deviations. Items marked with an * in Table 4.1b were selected for use in the next phase of the research and were combined with the science content statements (shown in Tables 4.3a and b) to form the items from which the pilot questionnaires were constructed.

Phase Two: Pilot

The pilot questionnaires consisted of 60 items and were administered to 248 pupils across KS 3 and KS 4. The questionnaires required between 20 to 30 minutes to complete and resulted in minimal impact on the school. There were no problems reported with pupils not understanding the Likert scale response mode. Few problems were reported with administering the questionnaires although two teachers reported that the terms 'Physics', Chemistry' and 'Biology' caused some confusion amongst the younger pupils in KS 3. In these classes a number of pupils had asked the teacher administering the questionnaire to clarify the meaning of the words. This was thought to be an issue associated with Year 7 pupils who had been in their secondary school only a few weeks and had yet to come across the terms as part of the classroom discourse. To a lesser extent this may have also applied to a small number of Year 8 pupils. As a result of this, and in order to minimise this uncertainty, 12 of the 18 items relating to the terms Physics, Chemistry and Biology were removed from the questionnaire for the main study. The decision was taken to remove the Internal and External items, leaving only the six absolute items; two each for physics, chemistry and biology. The internal and external items relating to Science and Science Investigations were left unchanged. These were now the only items which contained the internally and externally related frameworks.

Following the completion of the pilot study, a number of additional items were added to the questionnaire. The additional statements were the eight items of Marsh's Academic Self Description Questionnaire (ASDQ: Marsh, 1990c) relating to science, the six ASDQ General School items and four newly written items about the Importance of Science. This resulted in the final questionnaire comprising a total of 66 items. The 18 additional items can be found in Table 4.2.

The final science content part of the items were selected after a process of referral to:

i) the National Curriculum for Science in England and Wales (DfES, 2000);
ii) the QCA Schemes of work for KS3 (QCA, 2006);
iii) AQA examination specifications for various GCSE science courses (AQA, 2006)
iv) Meetings and interviews with various Heads of Science

A list of the final science content statements can be found in Tables 4.3a and 4.3b. Table 4.3a specifies the statements for KS3 and Table 4.3b for KS4. Below in Table 4.2 are the *additional statements* referred to earlier. Note that the *additional statements* were not part of the newly formed self-concept instrument. The eight ASD statements were used to test concurrent validity whilst the four IOS and six GSM statements were included to aid cross-network research referred to in the final chapter.

Table 4.2 Additional Statements

Additional Statements		
I	IOSP1	It is important for me to gain high marks in science tests or exams
o	IOSP2	Doing well in Science is important for me
S	IOSP3	It is important for me to understand the work in Science
	IOSP4	It is important for me to make good progress in Science
	ASDN3	I am hopeless when it comes to Science
A	ASDP1	I get good marks in Science classes
S	ASDP2	Work in Science classes is easy for me
D	ASDP4	I learn things quickly in Science
Q	ASDP5	Compared to others of my age I am good at Science classes
II	ASDP6	I have always done well in Science classes
	ASDP7	It is important for me to do well I Science classes
	ASDP8	I am satisfied with how well I do in Science classes
S	GSMN1	I am hopeless when it comes to most school subjects
C	GSMP2	I learn things quickly in most school subjects
H	GSMP3	I have always done well in most school subjects
O	GSMP4	Compared to others my age I am good at most school subjects
O	GSMP5	Work in most school subjects is easy for me
L	GSMP6	I get good marks in most school subjects

Coding
IOS are importance of science statements
ASD are the science statements from Marsh's (1990b) ASDQ II instrument
GSM are general school statements from Marsh's (1990b) ASDQ II instrument

Table 4.3a Science Statements for Key Stage Three

Physics
what energy can do
electrical circuits
how forces can move and change things
planet Earth and the solar system
heating and cooling
light and seeing, sound and hearing
Chemistry
acids and alkalis
chemical reactions
solids, liquids and gases
atoms, elements and compounds
rocks and weathering
reactivity of metals
Biology
cells
how our body works, e.g. digestion, keeping healthy, reproduction
putting living things into groups and looking at differences between them
respiration and photosynthesis
living things in their environment, e.g. food chains and habitats
inheritance and selection
Enquiry
thinking up ideas to investigate
planning how to do an experiment
collecting the results of an experiment
recording results from an experiment
drawing and explaining graphs
explaining the results of an experiment
Nature of Science
why we need scientists
how scientists work
what scientists do
how science can help us
how we get science knowledge
where theories come from

Table 4.3b Science Statements for Key Stage Four

Physics
what energy can do and how we use it
how forces can move and change things
planet Earth and space
electricity and how it is made and used
light, sound and waves
calculations and formulae
Chemistry
periodic table
rates of reaction
useful products from oil and rocks
chemical reactions and equations
atoms, molecules and bonding
changes in the Earth and its atmosphere
Biology
cells
systems of the human body, e.g. digestion, circulation, breathing, respiration.
photosynthesis and the transport of substances in green plants.
living things in their environment e.g. adaptation, competition and food chains
variation and inheritance
evolution
Enquiry
thinking up ideas to investigate
planning how to do an experiment
collecting the data of an experiment
recording results from an experiment
drawing and explaining tables and graphs
evaluating an experiment
Nature of Science
why we need scientists
how scientists work
what scientists do
how science can help us
how we get science knowledge
where theories come from

Phase 3: Main Study

The results from the Main Study are separated into four sections. The first section reports on the data preparation procedures. The second section consists of a principal-components analysis of the factors. This was carried out as a preface to the more robust confirmatory factor analysis. The third section provides the descriptive statistics of the individual items and their factors. The fourth section reports information on the model testing and associated work, particularly parameter estimation and model fit analysis.

Section One: Data Preparation

In the main research a 66 item questionnaire was administered to 1488 school pupils within eight different schools. The pupils were between the ages of 11-16 inclusive, (Year 7 to Year 11). Completed questionnaires were collected from 770 boys and 718 girls.

Data preparation was the first stage of the preliminary analysis. This consisted of:

- screening for rogue submissions
- reverse scoring appropriate items
- dealing with missing data
- creating self-concept items pairs
- identifying multivariate outliers

The first stage of the data analysis consisted of a visual inspection of the data along with an examination of the actual instrument scripts. From this examination, spoilt papers and papers with large amounts of missing data were removed from the sample. In all 90 papers were removed at this stage, (59 boys and 31 girls). Spoilt papers consisted of any deliberately mis-scored items, which for instance included questionnaires where the same response score was inserted for every item, or nearly every item, or the respondents had created interesting patterns on the answer grid. Cases with more than 10% missing items were excluded from the sample. Where this occurred it was likely that either the participants had not reached the end of the

questionnaire, or had skipped an entire page from the centre (probably due to turning over two pages at once), or had a disproportionately large number of missing items peppered throughout the questionnaire.

The final sample, which was used for all further analysis consisted of 1398 participants of which 711 were boys and 687 were girls. The spoilt or missing data scripts were spread across all six schools and across all ages and did not seem to form any pattern, although they may well have been scripts from disaffected pupils. Number of spoilt scripts were proportionally small in number compared with the whole sample and further investigation into this was not carried out.

Items were reverse scored where necessary, ensuring that the responses to 'negatively' worded items were changed such that the scoring direction was consistent with the positively worded items. This resulted in all item responses being consistent with low numbers on the scoring grid representing higher self-concept scores.

The next stage required the construction of item pairs. Each of the questionnaire items was combined with a similar item to create an item pair. This has been an often used technique and underpins the data processing procedures of the SDQ instruments (Marsh, 1992b).

Before the process of item pairing was completed an analysis of the individual items as sole entities was undertaken. The arithmetic means, standard deviations, alpha scores for internal reliability, skewness and kurtosis of the individual items were examined. This was undertaken before the uniqueness of each item was lost due to the pairing procedures. Completing this also served to check that homogeneity of the dimension of measurement, as was required above, was satisfied. The outcome of these tests will be discussed further below.

Missing data were imputed using the Expectation-Maximization algorithm (EM algorithm). This procedure employed maximum likelihood estimators to hypothesize values for the missing data and was undertaken in SPSS 11.5. This was a particularly robust procedure which has become the procedure of choice for many researchers.

Table 4.4a Items pairs for Key Stage 4

Facet	Code	Final Questionnaire Statements	Item pair
S	GAP5	I have good knowledge of science	SCI1 = GAP5, GAN6
C	GAN6	I have difficulty learning new things about science	
I	GIP5	I enjoy science much more than my other subjects	SCI2 = GIP5, GIN1
E	GIN1	Science is my weakest subject	
N	GEP7	I know more about science than my friends	SCI3 = GEP7, GEN5
C	GEN5	My friends have a clearer understanding than me about science	
	GAP3	I am pretty good when it comes to science investigations	ENQ1 = GAP3, GAN2
	GAN2	I usually do poorly at science investigations	
E	GIP2	I am better at science investigations than my other subjects	ENQ2 = GIP2, GIN1
N	GIN5	I enjoy my other subjects much more than science investigations	
Q	GEP2	I am better at science investigations than my friends	ENQ3 = GEP2, GEN7
U	GEN7	My friends know more about science investigations than me	
I	EAN1	I have difficulty in evaluating an experiment	ENQ4 = EAP1, EAN2
R	EAN2	I usually do poorly at drawing and explaining tables and graphs	
Y	EAN6	I have difficulty in recording results from experiments	ENQ5 = EAP2, EAN6
	EAP1	I am good at planning how to do an experiment	
	EAP2	I usually do well at collecting the data of an experiment	ENQ6 = EAP4, EAN1
	EAP4	I have no trouble in thinking up ideas to investigate	
B	GAP2	I usually do well at biology	BIOL1 = GAP2, GAN5
I	GAN5	I have poor knowledge of biology	
O	BAN2	I usually do poorly at understanding systems of the human body	BIOL2 = BAP5, BAN5
L	BAN3	I am pretty bad when it comes to cells	
O	BAN5	I have poor knowledge of living things in their environment	BIOL3 = BAP2, BAN2
G	BAP2	I usually do well at photosynthesis and the transport of substances	
Y	BAP3	I am pretty good when it comes to learning about variation and inheritance	BIOL4 = BAP3, BAN3
	BAP5	I have good knowledge of evolution	
P	GAP1	I am good at understanding about physics	PHYS1 = GAP1, GAN3
H	GAN3	I am pretty bad when it comes to physics	
Y	PAN1	I have difficulty in understanding about light, waves and sound	PHYS2 = PAP1, PAN1
S	PAN2	I usually do poorly at understanding what energy can do and how we use it	
I	PAN3	I am pretty bad when it comes to planet Earth and space	PHYS3 = PAP2, PAN2
C	PAP1	I am good at understanding about electricity and how it is made and used	
S	PAP2	I usually do well at learning how forces can move and change things	PHYS4 = PAP3, PAN3
	PAP3	I am pretty good when it comes to learning about calculations and formulae	
C	GAP4	I have no trouble in learning about chemistry	CHEM1 = GAP4, GAN1
H	GAN1	I have difficulty in understanding about chemistry	
E	CAN1	I have difficulty in understanding about rates of reactions	CHEM2 = CAP1, CAN6
M	CAN5	I have poor knowledge of atoms, molecules and bonding	
I	CAN6	I have difficulty learning new things about useful products from oil and rocks	CHEM3 = CAP4, CAN1
S	CAP1	I am good at understanding about changes in the Earth and its atmosphere	
T	CAP4	I have no trouble in learning about the periodic table	CHEM4 = CAP5, CAN5
	CAP5	I have good knowledge of chemical reactions and equations	
	NAN3	I am pretty bad when it comes to knowing how science can help us	NoS1 = NAP3, NAN3
N	NAN5	I have poor knowledge of where theories come from	
o	NAN6	I have difficulty learning about how we get science knowledge	NoS2 = NAP4, NAN5
S	NAP3	I am pretty good when it comes to knowing why we need scientists	
	NAP4	I have no trouble in knowing how scientists work	NoS3 = NAP5, NAN6
	NAP5	I have good knowledge of what scientists do	
I	IOSP1	It is important for me to gain high marks in science tests or exams	IoS1 = IOSP1, IOSP2
o	IOSP2	Doing well in Science is important for me	
S	IOSP3	It is important for me to understand the work in Science	IoS2 = IOSP3, IOSP4
	IOSP4	It is important for me to make good progress in Science	

Table 4.4b Items pairs for Key Stage 3

Facet	Code	Final Questionnaire Statements	Item pair
S	GAP5	I have good knowledge of science	SCI1 = GAP5, GAN6
C	GAN6	I have difficulty learning new things about science	
I	GIP5	I enjoy science much more than my other subjects	SCI2 = GIP5, GIN1
E	GIN1	Science is my weakest subject	
N	GEP7	I know more about science than my friends	SCI3 = GEP7, GEN5
C	GEN5	My friends have a clearer understanding than me about science	
	GAP3	I am pretty good when it comes to science investigations	ENQ1 = GAP3, GAN2
	GAN2	I usually do poorly at science investigations	
E	GIP2	I am better at science investigations than my other subjects	ENQ2 = GIP2, GIN1
N	GIN5	I enjoy my other subjects much more than science investigations	
Q	GEP2	I am better at science investigations than others in my class	ENQ3 = GEP2, GEN7
U	GEN7	My friends know more about science investigations than me	
I	EAN1	I have difficulty in explaining the results of an experiment	ENQ4 = EAP1, EAN2
R	EAN2	I usually do poorly at drawing and explaining graphs	
Y	EAN6	I have difficulty in recording results from experiments	ENQ5 = EAP2, EAN6
	EAP1	I am good at planning how to do an experiment	
	EAP2	I usually do well at collecting the results of an experiments	ENQ6 = EAP4, EAN1
	EAP4	I have no trouble in thinking up ideas to investigate	
B	GAP2	I usually do well at biology	BIOL1 = GAP2, GAN5
I	GAN5	I have poor knowledge of biology	
O	BAN2	I usually do poorly at understanding how our bodies work, eg dig and repro	BIOL2 = BAP5, BAN1
L	BAN3	I am pretty bad when it comes to cells	
O	BAN5	I have poor knowledge of living things in their envir, eg food chains and habs	BIOL3 = BAP2, BAN2
G	BAP2	I usually do well at putting living things into groups and looking at differences	
Y	BAP3	I am pretty good when it comes to learning about inheritance and selection	BIOL4 = BAP3, BAN3
	BAP5	I have good knowledge of respiration and photosynthesis	
P	GAP1	I am good at understanding about physics	PHYS1 = GAP1, GAN3
H	GAN3	I am pretty bad when it comes to physics	
Y	PAN1	I have difficulty in understanding about light and seeing, sound and hearing	PHYS2 = PAP1, PAN1
S	PAN2	I usually do poorly at understanding what energy can do	
I	PAN3	I am pretty bad when it comes to planet Earth and solar system	PHYS3 = PAP2, PAN2
C	PAP1	I am good at understanding about electrical circuits	
S	PAP2	I usually do well at learning how forces can move and change things	PHYS4 = PAP3, PAN3
	PAP3	I am pretty good when it comes to learning about heating and cooling	
C	GAP4	I have no trouble in learning about chemistry	CHEM1 = GAP4, GAN1
H	GAN1	I have difficulty in understanding about chemistry	
E	CAN1	I have difficulty in understanding about reactivity of metals	CHEM2 = CAP1, CAN6
M	CAN5	I have poor knowledge of atoms, elements and compounds	
I	CAN6	I have difficulty learning new things about acids and alkalis	CHEM3 = CAP5, CAN5
S	CAP1	I am good at understanding about rocks and weathering	
T	CAP4	I have no trouble in learning about solids, liquids and gases	CHEM4 = CAP5, CAN5
	CAP5	I have good knowledge of chemical reactions	
	NAN3	I am pretty bad when it comes to knowing how science can help us	NoS1 = NAP3, NAN3
N	NAN5	I have poor knowledge of where theories come from	
o	NAN6	I have difficulty learning about how we get science knowledge	NoS2 = NAP4, NAN5
S	NAP3	I am pretty good when it comes to knowing why we need scientists	
	NAP4	I have no trouble in knowing how scientists work	NoS3 = NAP5, NAN6
	NAP5	I have good knowledge of what scientists do	
I	IOSP1	It is important for me to gain high marks in science tests or exams	IoS1 = IOSP1, IOSP2
o	IOSP2	Doing well in Science is important for me	
S	IOSP3	It is important for me to understand the work in Science	IoS2 = IOSP3, IOSP4
	IOSP4	It is important for me to make good progress in Science	

Section Two: Principal Components Analysis

Exploratory Factor Analysis (EFA) was employed as a preface to the more robust model testing undertaken through confirmatory factor analysis and structural equation modelling. The advantage of running EFA on the data set at this early stage was that it allowed an initial exploration of the data prior to the constraints imposed by the more structured rigidity of the SEM tests. Although the models had been decided in an a priori manner, this was the first time that data of the type (to the author's knowledge) had been collected about science self-concept. It seemed legitimate and prudent to first ascertain if the data, when freely allowed to load onto a defined number of factors, did so in a way which was consistent with the a priori model. This provided some, although a limited, validity check of the model.

Table 4.5a: All Pupils

Item - Pairs	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Uniq Var
SCI1	**0.585**	-0.020	0.084	0.212	0.042	0.055	0.313
SCI2	**0.774**	-0.040	-0.044	-0.015	0.070	-0.023	0.438
SCI3	**0.420**	0.041	0.128	0.058	-0.011	0.078	0.619
ENQ4	-0.071	**0.741**	0.020	0.051	-0.003	-0.055	0.499
ENQ5	0.051	**0.401**	0.036	0.304	-0.042	-0.090	0.628
ENQ6	0.035	**0.477**	-0.036	0.104	-0.017	0.201	0.527
BIOL1	0.206	0.102	**0.540**	-0.175	0.116	-0.065	0.529
BIOL2	0.016	0.159	**0.482**	0.156	-0.029	0.051	0.475
BIOL3	-0.038	-0.127	**0.762**	0.265	-0.120	-0.077	0.474
BIOL4	-0.096	0.059	**0.496**	-0.046	0.136	0.163	0.570
PHYS1	0.167	0.173	-0.078	**0.221**	0.174	0.106	0.585
PHYS2	0.011	0.093	0.025	**0.556**	0.028	-0.050	0.608
PHYS3	-0.008	0.061	0.037	**0.589**	0.062	0.052	0.484
PHYS4	0.030	-0.021	0.100	**0.532**	0.098	-0.008	0.574
CHEM1	0.154	0.058	0.001	0.058	**0.534**	0.018	0.467
CHEM2	-0.037	0.121	0.234	-0.018	**0.219**	0.175	0.653
CHEM3	-0.049	-0.087	0.009	0.356	**0.427**	0.036	0.596
CHEM4	0.038	-0.025	-0.023	0.089	**0.701**	-0.023	0.458
NOS1	0.108	-0.006	0.103	0.229	-0.092	**0.419**	0.559
NOS2	-0.045	-0.022	-0.031	-0.032	0.024	**0.837**	0.400
NOS3	0.118	0.008	0.032	0.066	0.018	**0.582**	0.448

Principal Components Factor Analysis with Standardized Promax Rotated Factor Loadings.

The EFA was carried out using LISREL 8 to produce principal component analysis using both Varimax-rotation and Promax-rotation, and with SPSS 11.5 using Direct Oblim rotation. The exploratory factor analysis was run a number of times, firstly with the whole data set and then with the KS3 and KS4 segments of the data sample separately. This was important as although the items in the two Key Stages were very closely related, they were not all completely identical in their wording.

The three rotation methods produced results with no noticeable difference in features. In fact, although the values of the factor loading were different for each of the extraction methods, the patterns of loadings were difficult to distinguish apart. To save repetition, only the Standardized Promax Factor loadings will be presented here. In each case the rotations were forced to reveal six factors, as was consistent with the original model.

Table 4.5b: Key Stage 3 Pupils

Item - Pairs	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Uniq var
SCI1	**0.506**	0.025	0.051	0.258	0.045	0.101	0.331
SCI2	**0.745**	-0.044	0.044	-0.016	0.051	-0.030	0.451
SCI3	**0.356**	0.075	0.058	0.102	-0.058	0.160	0.653
ENQ4	-0.068	**0.618**	0.126	0.046	-0.031	0.013	0.546
ENQ5	0.028	**0.498**	-0.050	0.293	-0.010	-0.076	0.567
ENQ6	0.015	**0.547**	0.045	0.023	-0.009	0.166	0.493
BIOL1	0.230	0.065	**0.511**	-0.006	0.061	-0.030	0.486
BIOL2	0.046	0.126	**0.525**	0.351	-0.102	-0.064	0.414
BIOL3	-0.084	-0.094	**0.391**	0.590	-0.131	0.007	0.543
BIOL4	-0.083	-0.022	**0.351**	0.087	0.177	0.233	0.575
PHYS1	0.156	0.204	0.222	**-0.046**	0.192	0.085	0.546
PHYS2	0.019	0.206	-0.011	**0.434**	0.117	-0.112	0.640
PHYS3	0.032	0.126	-0.037	**0.481**	0.142	0.069	0.502
PHYS4	0.108	-0.031	0.018	**0.537**	0.057	0.039	0.586
CHEM1	0.120	0.146	0.252	0.041	**0.291**	0.001	0.510
CHEM2	-0.049	0.056	0.217	0.025	**0.297**	0.157	0.633
CHEM3	-0.061	-0.010	0.032	0.330	**0.286**	0.065	0.683
CHEM4	0.052	-0.049	-0.052	0.047	**0.771**	-0.019	0.433
NOSI	0.123	0.129	-0.027	0.248	-0.091	**0.376**	0.563
NOS2	-0.002	-0.031	-0.024	-0.034	0.050	**0.767**	0.445
NOS3	0.052	0.120	0.032	0.077	-0.029	**0.580**	0.444

Principal Components Factor Analysis with Standardized Promax Rotated Factor Loadings.

For completeness, other numbers of factors were also run and the results examined to see if a larger or smaller number of factors would form cognate loadings. The number of factors which produced the most sensible arrangement of factor loadings was, reassuringly, six. Outputs with five factors or seven factors produced a pattern of results which was a 'distorted' version of the six factor model. In neither the five nor seven factor models were there clearly identifiable and distinctly coherent factor patterns. This was not the case with the six factor model which formed into the model anticipated which can be confirmed by examination of the results presented in Tables. Three tables are presented; Tables 4.5b and 4.5c show the KS3 and KS4 pupils shown separately, and Table 4.5a shows all pupils combined.

Table 4.5c: Key Stage 4 Pupils

Item-Pairs	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Uniq var
SCI1	**0.522**	0.002	0.131	0.157	0.117	0.063	0.328
SCI2	**0.755**	-0.016	-0.009	-0.002	0.091	0.004	0.355
SCI3	**0.382**	0.017	0.229	0.089	0.083	0.037	0.544
ENQ4	-0.030	**0.815**	0.000	-0.072	0.039	-0.044	0.427
ENQ5	0.026	**0.415**	0.113	0.137	0.048	-0.082	0.669
ENQ6	0.040	**0.412**	-0.087	0.061	-0.065	0.351	0.576
BIOL1	0.150	-0.002	**0.733**	-0.149	0.046	-0.098	0.494
BIOL2	-0.001	0.108	**0.479**	0.033	-0.131	0.292	0.493
BIOL3	-0.026	-0.029	**0.847**	-0.034	-0.039	-0.070	0.423
BIOL4	-0.101	0.003	**0.625**	0.096	0.010	0.100	0.508
PHYS1	0.225	0.005	-0.209	**0.739**	-0.068	0.033	0.407
PHYS2	-0.017	-0.004	0.018	**0.705**	-0.043	-0.012	0.549
PHYS3	-0.083	-0.068	0.055	**0.834**	0.043	-0.046	0.399
PHYS4	-0.044	0.136	0.070	**0.390**	0.232	-0.039	0.588
CHEM1	0.172	0.016	-0.101	-0.067	**0.710**	0.089	0.350
CHEM2	0.039	-0.016	0.204	0.208	**-0.013**	0.292	0.637
CHEM3	-0.054	0.019	0.022	0.113	**0.621**	0.049	0.494
CHEM4	0.017	0.029	0.019	0.003	**0.768**	0.015	0.340
NOS1	-0.036	-0.073	0.129	0.073	0.113	**0.548**	0.528
NOS2	-0.055	0.024	-0.058	-0.057	0.018	**0.860**	0.379
NOS3	0.135	-0.072	-0.008	-0.026	0.101	**0.669**	0.426

Principal Components Factor Analysis with Standardized Promax Rotated Factor Loadings.

The first point to note is the consistency between the factor loading presented in the Tables 4.5a, b & c. There are few differences between the factor loadings with the exceptions of PHYS1 for KS3 and CHEM2 for KS4. These will be discussed separately later below.

Each of the different factors had a very strong identity, with items loading very highly on the anticipated factor and loading very weakly on all other factors.

The most striking aspect of the results was the ways in which the terms from each section of the Programmes of Study loaded onto the same factor, for example, all the Biology terms loaded onto Factor 3 and all the Enquiry terms loaded onto Factor 2. Taking the Biology as an exemplar to explore more closely, we recall that biology comprised four item pairs. The subjects of the item pairs were:

- BIOL1: biology (positive); biology (negative)
- BIOL2: environment; evolution
- BIOL3: human body; photosynthesis
- BIOL4: cells; variation/inheritance

In relation to the result of all pupils, the loadings onto Factor 3 (biology) items pairs ranged from 0.762 to 0.482. For KS4 the loadings ranged from 0.847 to 0.497, and for KS3 the loadings ranged from 0.525 to 0.351. The fact that all the biology items loaded strongly on the same factor whilst loading weakly on all other factors provided evidence that, from a self-concept response perspective, all the items probably belong to the same group.

The outcome of the exploratory factor analysis demonstrated the clear possibility that pupils exhibit a distinct self-concept term in relation to biology, and that this self-concept is separate and independent from other science related self-concepts. Furthermore, this outcome is true for Physics, for Chemistry, for the Nature of Science, for Enquiry and for Science. Across the whole spectrum of the science experience, all, or very nearly all the items, from each different science facet loaded together onto the same factor. This may seem self-evident, but so far it has not been argued thus in the literature. Indeed, it does seem to provide evidence for the principle of separately measurable academic self-concepts across the different 'subject' areas within science. If this is so then the proposal here that there is need

for an extension to the Shavelson/Marsh structure to include an additional level is supported with empirical evidence. The outcome of this exploratory factor analysis did not appear to be a manifestation of the linguistics in which similar sounding terms became grouped together, nor did there seem to be any obvious pointers within the phrasing of the statements which pointed the participants to connected responses, nor were there subversive forces through the way the items had been grouped, organised or coded. The assertion being made here is that this was a measurably distinct way in which the curriculum had impacted and influenced pupils' self-concept about science.

There are other interesting outcomes shown in Tables 4.5a, b & c. It does appear that scientific processes and procedures, as well as aspects of knowledge and understanding, are also equally likely to be grouped together.

The factor loading for KS3 were only just slightly less strong than for KS4. This was an unexpected outcome as the literature has suggested that as pupils age so their self-concept becomes progressively more distinct, (see Marsh and Ayotte, 2003). It was therefore expected that the distinctiveness of the factors would have been superior for KS4 than for KS3. This proposed increase in distinctiveness could be due in part to the fact that pupils in Years 7 and 8 have yet to have opportunity to experience the full breadth of the science curriculum, and as such, may have less well formed self-concepts. It is also thought that as the pupils encounter more and more of the curriculum so their self-concepts become more fully formed and stable, and consequently, measured more reliably. Additionally, Marsh and Ayotte (2003) have reported that young people are susceptible to a maturation process whereby young people's self-concepts become increasingly more multifaceted with age. There does not however seem to be strong evidence of that process occurring here.

The exploratory factor analysis served to intuit a factor structure from the data, and to provide reassurance that, with items allowed to load on all factors, a comprehensible and cognate set of factors would still emerge. By contrast the next phase employed the use of confirmatory factor analysis which was used to model 'causal' relationships between the factors (latent variables) and measure the strength of the fit of the empirical data to the theoretical model.

Section Three: Descriptive Statistics

The next section reports the descriptive statistics for the full questionnaire data. It includes information on all raw, individual items together with information on derived variables including items pairs and factor scales. The statistical data reported here include information on arithmetic means, standard deviations, skewness, and kurtosis values for all raw and derived variables. Summary data for the whole sample is presented in Table 4.6a. Subgroup data is presented by sex in Table 4.6b and Key Stage data in Table 4.6c.

Checking the Data

The skewness and kurtosis values were examined to check for normality of distribution. This was the first stage of the preliminary analysis. Normality conditions must be satisfied if SEM procedures are to be successfully employed. Normality is assumed if the data demonstrates skewness values within the range of -1.0 to 1.0. Values outside this range should be treated with caution. Curran, West and Finch (1996) following Monte Carlo simulation research have suggested that values in the range of 2.0 to 3.0 should be considered moderately non-normal and values greater than 3.0 should be considered significantly non-normal. Similarly, kurtosis values falling within the range -1.0 to 1.0 indicate kurtotic normally. Similarly, Curran et al. have suggested that values of kurtosis in the range 7.0 to 21.0 should be considered moderately non-normal and values greater than 21.0 considered as kurtotically non-normal. An inspection of the data in Tables 4.6 a, b and c reveal that the vast majority of the data show no tendency toward either skewness of kurtosis. Values for all four of the Importance of Science items were at the upper edge of skewness acceptability with values around the 1.0 mark, which, although not extreme enough to be considered as even moderately skewed, were moving toward the limit of acceptable normality.

The Importance of Science distribution, interestingly, was caused by a disproportionate number of individuals indicating very positive scores on the Importance of Science items, and likewise disproportionately few numbers of individuals indicating very negative scores. This positive tendency is also manifest in the factor means which can be seen to be significantly large values when compared

against the other questionnaire items. The most skewed item on the questionnaire was Marsh's GSMN1 (I am hopeless when it comes to most school subjects) with a skewness value of 1.4, and which, incidentally also exhibited the greatest amount of kurtosis of all items. With the exception of this one item, (IOSP1), all items were safely contained within the safe kurtotic range of 'modulus one'. This can be confirmed by examination of Tables 4.6 a, b and c. In order to correct for the slight amount of non-normality in the items, the Satorra-Bentler Chi square correction (Satorra and Bentler, 1994) was employed in the creation of an asymptotic covariance matrix in the LISREL data preparation stage. This mitigated for non-normality in the data and produced less bias in the outcomes of statistical procedures. This was not an entirely necessary procedure, given the normality of the data, although it served no harm to take this precautionary measure.

The final stage of the preliminary analyses involved an examination of the newly constructed science self-concept scales. Statistical information about the scales for Global Science, Enquiry, Biology, Physics, Chemistry and Nature of Science, together with the additional Importance of Science scale was examined. This was carried out separately for boys and girls, and for Key Stage 3 and Key Stage 4 individuals, as indeed all the tests above were. The full listing of results can be found in Tables 4.7 a, b and c.

Table 4.6a Descriptive Statistics for whole sample

Individual Items	N		% missing	Mean	St Deviation	Skewness	Kurtosis
	Valid	Missing					
GAP1	1374	28	2.04	2.857	1.095	0.185	-0.499
GAP2	1381	21	1.52	2.729	1.130	0.338	-0.535
GAP3	1386	16	1.15	2.636	0.986	0.269	-0.243
GAP4	1378	24	1.74	2.690	1.130	0.251	-0.599
GAP5	1378	24	1.74	2.425	1.022	0.466	-0.154
GAN1	1385	17	1.23	2.635	1.165	0.274	-0.690
GAN2	1384	18	1.30	2.329	1.021	0.521	-0.244
GAN3	1374	28	2.04	2.643	1.163	0.361	-0.637
GAN5	1378	24	1.74	2.334	1.155	0.651	-0.338
GAN6	1376	26	1.89	2.390	1.081	0.495	-0.444
GEN5	1360	42	3.09	2.788	1.143	0.196	-0.614
GEN7	1344	58	4.32	2.879	1.107	0.114	-0.534
GEP2	1389	13	0.94	3.275	1.045	-0.093	-0.517
GEP7	1375	27	1.96	3.215	1.109	-0.105	-0.532
GIN1	1386	16	1.15	2.341	1.290	0.662	-0.643
GIN5	1382	20	1.45	3.336	1.254	-0.195	-0.954
GIP2	1391	11	0.79	3.405	1.092	-0.274	-0.508
GIP5	1375	27	1.96	3.372	1.208	-0.293	-0.786
EAN1	1358	44	3.24	2.526	1.144	0.384	-0.574
EAN2	1358	44	3.24	2.328	1.207	0.600	-0.616
EAN6	1383	19	1.37	2.390	1.166	0.374	-0.947
EAP1	1392	10	0.72	2.722	0.862	0.210	0.114
EAP2	1378	24	1.74	2.498	0.995	0.323	-0.328
EAP4	1378	24	1.74	2.900	1.092	-0.025	-0.685
BAN2	1391	11	0.79	2.300	1.128	0.608	-0.449
BAN3	1364	38	2.79	2.469	1.106	0.410	-0.498
BAN5	1361	41	3.01	2.160	1.086	0.746	-0.138
BAP2	1374	28	2.04	2.373	1.090	0.537	-0.313
BAP3	1359	43	3.16	3.005	1.132	0.103	-0.617
BAP5	1371	31	2.26	2.843	1.149	0.190	-0.667
CAN1	1384	18	1.30	2.818	1.096	0.061	-0.604
CAN5	1378	24	1.74	2.655	1.203	0.317	-0.797
CAN6	1353	49	3.62	2.557	1.127	0.366	-0.549
CAP1	1393	9	0.65	2.894	1.067	0.131	-0.592
CAP4	1369	33	2.41	2.104	1.168	0.871	-0.134
CAP5	1385	17	1.23	2.748	1.081	0.211	-0.499
PAN1	1386	16	1.15	2.434	1.120	0.436	-0.624
PAN2	1389	13	0.94	2.465	1.025	0.397	-0.389
PAN3	1392	10	0.72	2.283	1.173	0.641	-0.501
PAP1	1398	4	0.29	2.343	1.012	0.506	-0.209
PAP2	1363	39	2.86	2.380	1.028	0.440	-0.251
PAP3	1388	14	1.01	2.417	1.095	0.510	-0.328
NAN3	1382	20	1.45	2.496	1.116	0.383	-0.620
NAN5	1361	41	3.01	2.827	1.099	0.191	-0.466
NAN6	1377	25	1.82	2.559	1.019	0.171	-0.467
NAP3	1361	41	3.01	2.508	1.110	0.381	-0.486
NAP4	1380	22	1.59	2.855	1.077	0.088	-0.579
NAP5	1383	19	1.37	2.810	1.078	0.182	-0.589
IOSP1	1386	16	1.15	1.822	0.956	1.066	0.588
IOSP2	1377	25	1.82	2.002	1.096	0.987	0.255
IOSP3	1373	29	2.11	1.854	0.984	1.058	0.590
IOSP4	1365	37	2.71	1.955	1.070	0.935	0.092
ASDN3	1149	23	2.00	2.153	1.209	0.778	-0.412
ASDP1	1152	20	1.74	2.393	0.975	0.448	-0.120
ASDP2	1154	18	1.56	2.744	1.032	0.277	-0.308
ASDP4	1141	31	2.72	2.672	1.104	0.220	-0.594
ASDP5	1155	17	1.47	2.712	1.022	0.226	-0.366
ASDP6	1157	15	1.30	2.564	1.020	0.230	-0.505
ASDP7	1162	10	0.86	1.976	1.063	0.965	0.231
ASDP8	1142	30	2.63	2.344	1.128	0.575	-0.406
GSMN1	1160	12	1.03	1.751	1.033	1.344	1.001
GSMP2	1151	21	1.82	2.146	0.973	0.539	-0.279
GSMP3	1163	9	0.77	2.089	0.980	0.696	0.049
GSMP4	1150	22	1.91	2.368	1.021	0.414	-0.297
GSMP5	1149	23	2.00	2.392	0.998	0.384	-0.294
GSMP6	1168	4	0.34	2.022	0.918	0.683	0.152

Table 4.6b Descriptive statistics for sex subgroups

Individual Items	Girls						Boys					
	N		%	St			N		%	St		
	Valid	Missing	missing	Deviation	Skewness	Kurtosis	Valid	Missing	missing	Deviation	Skewness	Kurtosis
GAP1	680	8	3.074	1.067	0.058	-0.430	692	20	2.645	1.083	0.344	-0.410
GAP2	680	8	2.734	1.083	0.370	-0.435	699	13	2.724	1.176	0.316	-0.633
GAP3	681	7	2.761	0.969	0.153	-0.253	703	9	2.515	0.990	0.401	-0.121
GAP4	682	6	2.826	1.108	0.180	-0.557	694	18	2.555	1.137	0.349	-0.579
GAP5	682	6	2.569	1.005	0.307	-0.196	694	18	2.282	1.021	0.661	0.085
GAN1	684	4	2.765	1.133	0.199	-0.603	699	13	2.508	1.183	0.374	-0.714
GAN2	680	8	2.375	1.005	0.472	-0.261	702	10	2.283	1.036	0.577	-0.206
GAN3	677	11	2.802	1.174	0.263	-0.699	695	17	2.488	1.133	0.462	-0.530
GAN5	679	9	2.312	1.124	0.684	-0.197	697	15	2.353	1.187	0.624	-0.460
GAN6	679	9	2.474	1.083	0.384	-0.541	695	17	2.306	1.075	0.617	-0.282
GEN5	675	13	2.889	1.142	0.135	-0.609	683	29	2.688	1.139	0.262	-0.596
GEN7	672	16	2.954	1.111	0.098	-0.539	670	42	2.804	1.100	0.129	-0.534
GEP2	681	7	3.464	1.024	-0.183	-0.573	706	6	3.089	1.031	-0.008	-0.398
GEP7	675	13	3.410	1.049	-0.083	-0.563	698	14	3.026	1.135	-0.057	-0.583
GIN1	684	4	2.510	1.309	0.500	-0.847	700	12	2.173	1.250	0.847	-0.311
GIN5	679	9	3.591	1.197	-0.422	-0.767	701	11	3.088	1.260	0.031	-0.942
GIP2	684	4	3.598	1.017	-0.372	-0.266	705	7	3.217	1.131	-0.126	-0.659
GIP5	681	7	3.595	1.157	-0.480	-0.581	692	20	3.150	1.218	-0.111	-0.837
EAN1	675	13	2.590	1.112	0.348	-0.544	681	31	2.461	1.174	0.438	-0.585
EAN2	672	16	2.260	1.181	0.629	-0.621	684	28	2.395	1.230	0.568	-0.626
EAN6	678	10	2.381	1.176	0.409	-0.895	703	9	2.403	1.158	0.337	-0.999
EAP1	683	5	2.773	0.846	0.158	0.033	707	5	2.670	0.875	0.273	0.222
EAP2	681	7	2.577	0.990	0.187	-0.480	695	17	2.419	0.995	0.468	-0.097
EAP4	679	9	3.080	1.060	-0.181	-0.519	697	15	2.723	1.096	0.144	-0.698
BAN2	687	1	2.262	1.103	0.644	-0.357	702	10	2.338	1.154	0.567	-0.538
BAN3	678	10	2.524	1.088	0.347	-0.451	684	28	2.415	1.123	0.480	-0.515
BAN5	675	13	2.110	1.062	0.757	-0.128	684	28	2.208	1.108	0.736	-0.156
BAP2	679	9	2.324	1.038	0.535	-0.167	693	19	2.423	1.139	0.515	-0.473
BAP3	665	23	3.050	1.126	0.055	-0.602	692	20	2.962	1.138	0.151	-0.622
BAP5	678	10	2.973	1.145	0.082	-0.689	691	21	2.713	1.139	0.307	-0.575
CAN1	681	7	2.900	1.059	-0.016	-0.513	701	11	2.735	1.123	0.147	-0.651
CAN5	677	11	2.756	1.207	0.244	-0.858	699	13	2.559	1.192	0.388	-0.716
CAN6	671	17	2.663	1.110	0.300	-0.535	680	32	2.451	1.136	0.453	-0.517
CAP1	685	3	2.982	1.033	0.059	-0.464	706	6	2.806	1.093	0.221	-0.661
CAP4	680	8	2.071	1.122	0.890	-0.004	687	25	2.135	1.212	0.846	-0.267
CAP5	679	9	2.887	1.086	0.123	-0.510	704	8	2.614	1.061	0.296	-0.440
PAN1	685	3	2.508	1.119	0.391	-0.585	699	13	2.359	1.117	0.489	-0.638
PAN2	685	3	2.562	0.998	0.316	-0.359	702	10	2.369	1.044	0.502	-0.342
PAN3	686	2	2.383	1.146	0.477	-0.636	704	8	2.186	1.193	0.811	-0.284
PAP1	688	0	2.538	1.013	0.305	-0.343	708	4	2.151	0.974	0.743	0.241
PAP2	675	13	2.519	1.018	0.343	-0.232	686	26	2.242	1.021	0.567	-0.157
PAP3	684	4	2.520	1.096	0.453	-0.288	702	10	2.315	1.086	0.582	-0.331
NAN3	679	9	2.533	1.132	0.334	-0.695	701	11	2.458	1.102	0.435	-0.534
NAN5	671	17	2.897	1.069	0.103	-0.396	688	24	2.757	1.126	0.285	-0.493
NAN6	680	8	2.622	0.960	0.136	-0.230	695	17	2.496	1.072	0.234	-0.639
NAP3	675	13	2.622	1.060	0.266	-0.423	684	28	2.395	1.148	0.527	-0.448
NAP4	680	8	2.982	1.072	-0.080	-0.573	698	14	2.729	1.070	0.260	-0.453
NAP5	683	5	2.930	1.036	-0.018	-0.472	698	14	2.693	1.107	0.386	-0.549
IOSP1	683	5	1.818	0.923	1.006	0.510	701	11	1.823	0.989	1.117	0.635
IOSP2	678	10	2.037	1.114	0.962	0.186	697	15	1.967	1.079	1.016	0.335
IOSP3	676	12	1.818	0.985	1.174	0.901	695	17	1.888	0.985	0.952	0.326
IOSP4	680	8	1.928	1.055	1.005	0.335	683	29	1.981	1.085	0.874	-0.110
ASDN3	571	9	1.576	1.222	0.596	-0.673	576	14	2.431	1.183	0.982	-0.001
ASDP1	572	8	1.399	0.965	0.356	-0.238	578	12	2.076	0.980	0.553	0.068
ASDP2	577	3	0.520	1.040	0.161	-0.341	575	15	2.609	1.001	0.391	-0.165
ASDP4	567	13	2.293	1.103	0.132	-0.602	572	18	3.147	1.073	0.309	-0.537
ASDP5	574	6	1.045	0.999	0.140	-0.356	579	11	1.900	1.028	0.339	-0.286
ASDP6	573	7	1.222	0.965	0.210	-0.301	582	8	1.375	1.067	0.289	-0.641
ASDP7	576	4	0.694	1.049	0.917	0.128	584	6	1.027	1.076	1.016	0.346
ASDP8	572	8	1.399	1.179	0.452	-0.615	568	22	3.873	1.054	0.676	-0.150
GSMN1	576	4	0.694	1.025	1.519	1.659	582	8	1.375	1.040	1.188	0.460
GSMP2	572	8	1.399	0.944	0.494	-0.355	577	13	2.253	1.002	0.571	-0.249
GSMP3	578	2	0.346	0.908	0.797	0.359	583	7	1.201	1.037	0.577	-0.218
GSMP4	574	6	1.045	0.967	0.453	-0.141	574	16	2.787	1.072	0.379	-0.435
GSMP5	573	7	1.222	0.957	0.345	-0.294	574	16	2.787	1.034	0.399	-0.332
GSMP6	579	1	0.173	0.856	0.680	0.107	587	3	0.511	0.966	0.643	0.067

Table 4.6c Descriptive statistics for Key Stage subgroups

Individual Items	Key Stage 3						Key Stage 4					
	N		%	Std.			N		%	Std.		
	Valid	Missing	missing	Deviation	Skewness	Kurtosis	Valid	Missing	missing	Deviation	Skewness	Kurtosis
GAP1	876	16	2.817	1.082	0.248	-0.399	498	12	2.928	1.115	0.073	-0.623
GAP2	882	10	2.773	1.130	0.307	-0.546	499	11	2.651	1.128	0.396	-0.493
GAP3	884	8	2.535	1.017	0.367	-0.269	502	8	2.813	0.903	0.198	-0.033
GAP4	879	13	2.581	1.140	0.362	-0.559	499	11	2.882	1.087	0.095	-0.519
GAP5	879	13	2.358	1.047	0.540	-0.182	499	11	2.543	0.967	0.379	0.019
GAN1	885	7	2.586	1.171	0.289	-0.699	500	10	2.720	1.149	0.260	-0.666
GAN2	882	10	2.268	1.073	0.672	-0.189	502	8	2.436	0.913	0.229	-0.265
GAN3	871	21	2.580	1.137	0.456	-0.436	503	7	2.753	1.200	0.199	-0.876
GAN5	879	13	2.358	1.153	0.609	-0.392	499	11	2.291	1.159	0.730	-0.218
GAN6	874	18	2.324	1.102	0.555	-0.484	502	8	2.506	1.036	0.428	-0.281
GEN5	867	25	2.752	1.160	0.204	-0.675	493	17	2.852	1.112	0.197	-0.491
GEN7	854	38	2.874	1.135	0.090	-0.638	490	20	2.890	1.057	0.169	-0.324
GEP2	883	9	3.262	1.082	-0.087	-0.623	506	4	3.298	0.977	-0.089	-0.317
GEP7	873	19	3.212	1.112	-0.099	-0.524	502	8	3.221	1.106	-0.117	-0.539
GIN1	882	10	2.271	1.263	0.702	-0.560	504	6	2.462	1.327	0.589	-0.781
GIN5	879	13	3.265	1.279	-0.158	-0.994	503	7	3.461	1.200	-0.238	-0.895
GIP2	887	5	3.349	1.141	-0.248	-0.633	504	6	3.504	0.993	-0.250	-0.340
GIP5	877	15	3.348	1.222	-0.274	-0.846	498	12	3.414	1.182	-0.324	-0.668
EAN1	866	26	2.460	1.172	0.479	-0.560	492	18	2.642	1.084	0.237	-0.504
EAN2	868	24	2.321	1.245	0.629	-0.666	490	20	2.339	1.138	0.536	-0.532
EAN6	880	12	2.434	1.207	0.361	-0.990	503	7	2.314	1.088	0.358	-0.950
EAP1	885	7	2.712	0.895	0.266	0.112	507	3	2.740	0.802	0.090	0.041
EAP2	880	12	2.442	1.014	0.365	-0.400	498	12	2.596	0.954	0.282	-0.128
EAP4	877	15	2.811	1.127	0.085	-0.762	501	9	3.056	1.010	-0.182	-0.417
BAN2	889	3	2.245	1.118	0.644	-0.391	502	8	2.396	1.141	0.548	-0.529
BAN3	870	22	2.454	1.107	0.358	-0.635	494	16	2.496	1.105	0.504	-0.261
BAN5	868	24	2.100	1.093	0.804	-0.118	493	17	2.266	1.065	0.670	-0.096
BAP2	875	17	2.311	1.082	0.591	-0.212	499	11	2.483	1.096	0.452	-0.437
BAP3	870	22	3.072	1.152	0.030	-0.667	489	21	2.885	1.086	0.219	-0.461
BAP5	879	13	2.688	1.149	0.340	-0.563	492	18	3.120	1.096	-0.025	-0.599
CAN1	882	10	2.874	1.102	0.026	-0.616	502	8	2.719	1.079	0.118	-0.559
CAN5	873	19	2.641	1.203	0.340	-0.772	505	5	2.679	1.202	0.279	-0.834
CAN6	862	30	2.498	1.174	0.433	-0.639	491	19	2.662	1.032	0.288	-0.282
CAP1	886	6	2.955	1.116	0.094	-0.690	507	3	2.787	0.967	0.121	-0.479
CAP4	877	15	1.766	1.005	1.461	1.846	492	18	2.705	1.196	0.182	-0.863
CAP5	882	10	2.700	1.074	0.239	-0.492	503	7	2.833	1.090	0.159	-0.491
PAN1	885	7	2.371	1.173	0.521	-0.680	501	9	2.545	1.012	0.319	-0.415
PAN2	884	8	2.429	1.054	0.439	-0.436	505	5	2.529	0.972	0.342	-0.265
PAN3	884	8	2.276	1.223	0.645	-0.593	508	2	2.295	1.082	0.630	-0.340
PAP1	890	2	2.209	1.003	0.697	0.108	508	2	2.579	0.985	0.239	-0.352
PAP2	868	24	2.262	1.027	0.588	-0.055	495	15	2.588	0.998	0.235	-0.320
PAP3	887	5	2.110	0.949	0.701	0.274	501	9	2.960	1.127	0.062	-0.624
NAN3	879	13	2.510	1.163	0.373	-0.765	503	7	2.471	1.031	0.384	-0.330
NAN5	865	27	2.822	1.127	0.218	-0.509	496	14	2.835	1.051	0.135	-0.392
NAN6	878	14	2.526	1.043	0.232	-0.552	499	11	2.617	0.973	0.067	-0.251
NAP3	866	26	2.404	1.120	0.498	-0.404	495	15	2.691	1.068	0.221	-0.464
NAP4	879	13	2.790	1.086	0.179	-0.534	501	9	2.970	1.053	-0.064	-0.577
NAP5	881	11	2.719	1.087	0.273	-0.546	502	8	2.970	1.044	0.049	-0.561
IOSP1	884	8	1.781	0.950	1.148	0.817	502	8	1.894	0.964	0.938	0.280
IOSP2	873	19	1.977	1.105	1.010	0.243	504	6	2.046	1.079	0.957	0.311
IOSP3	869	23	1.822	0.970	1.068	0.578	504	6	1.909	1.007	1.041	0.599
IOSP4	870	22	1.899	1.064	1.071	0.426	495	15	2.055	1.074	0.719	-0.340
ASDN3	878	14	2.124	1.213	0.819	-0.367	271	9	3.321	1.200	0.693	-0.479
ASDP1	880	12	2.317	0.969	0.483	-0.163	272	8	2.941	0.970	0.395	0.046
ASDP2	884	8	2.682	1.030	0.348	-0.268	270	10	3.704	1.022	0.128	-0.281
ASDP4	873	19	2.570	1.091	0.308	-0.492	268	12	4.478	1.101	0.027	-0.656
ASDP5	883	9	2.642	1.025	0.297	-0.347	272	8	2.941	0.997	0.088	-0.282
ASDP6	881	11	2.519	1.025	0.280	-0.451	276	4	1.449	1.004	0.129	-0.575
ASDP7	886	6	1.916	1.050	0.995	0.221	276	4	1.449	1.082	0.916	0.268
ASDP8	874	18	2.184	1.089	0.744	-0.111	268	12	4.478	1.131	0.261	-0.597
GSMN1	883	9	1.785	1.073	1.303	0.805	277	3	1.083	0.933	1.397	1.353
GSMP2	876	16	2.094	0.987	0.619	-0.264	275	5	1.818	0.932	0.390	-0.168
GSMP3	885	7	2.038	0.980	0.755	0.092	278	2	0.719	0.971	0.590	0.057
GSMP4	877	15	2.319	1.025	0.458	-0.268	273	7	2.564	1.005	0.333	-0.300
GSMP5	878	14	2.341	1.006	0.445	-0.263	271	9	3.321	0.970	0.271	-0.270
GSMP6	891	1	1.970	0.928	0.864	0.630	277	3	1.083	0.887	0.275	-0.766

Table 4.7a Item pair and factor scale statistics for whole sample

Item pair and factor statistics for whole sample n = 1398				
	Mean	Std. Deviation	Skewness	Kurtosis
Science	2.7554	0.77504	alpha =	0.753
SCI1	2.4086	0.87640	0.355	-0.185
SCI2	2.8566	1.05003	0.226	-0.614
SCI3	3.0011	0.90706	-0.003	-0.246
Enquiry3	2.5615	0.65930	alpha =	0.685
ENQ1	2.8727	0.99210	0.294	-0.470
ENQ2	3.0775	0.84922	0.021	-0.139
ENQ3	2.4814	0.84923	0.289	-0.036
ENQ4	2.5260	0.79790	0.335	-0.222
ENQ5	2.4456	0.85083	0.155	-0.468
ENQ6	2.7128	0.87476	0.153	-0.243
Biology	2.5273	0.68641	alpha =	0.755
BIOL1	2.5324	0.98998	0.415	-0.260
BIOL2	2.5026	0.88449	0.282	-0.276
BIOL3	2.3368	0.87510	0.444	-0.087
BIOL4	2.7372	0.86252	0.219	-0.115
Physics	2.4780	0.65641	alpha =	0.743
PHYS1	2.7499	0.98282	0.218	-0.326
PHYS2	2.3884	0.82948	0.252	-0.207
PHYS3	2.4232	0.81862	0.272	-0.068
PHYS4	2.3505	0.85375	0.312	-0.330
Chemistry	2.6380	0.68051	alpha =	0.735
CHEM1	2.6645	0.99993	0.189	-0.394
CHEM2	2.7252	0.85461	0.182	-0.117
CHEM3	2.4616	0.86029	0.311	-0.080
CHEM4	2.7008	0.92554	0.223	-0.257
Nat of Sci	2.6766	0.69969	alpha =	0.752
NOS1	2.5025	0.88810	0.226	-0.301
NOS2	2.8416	0.85977	0.145	-0.076
NOS3	2.6858	0.81882	0.046	-0.320
Imp of Sci	1.9084	0.81819	alpha =	0.824
IOS1	1.9127	0.91114	0.897	0.120
IOS2	1.9042	0.86303	0.904	0.453

Missing data were imputed before the item pairs were constructed.
Hence there is no 'missing data' within these values.

Science = mean of SCI1, SCI2 and SCI3
Enquiry3 = mean of ENQ4, ENQ5 and ENQ6
Biology = mean of BIOL1, BIOL2 and BIOL3
Physics = mean of PHYS1, PHYS2 and PHYS3
Nat of Sci = mean of NOS1, NOS2 and NOS3
Imp of Sci = mean of IOS1 and IOS2

Table 4.7b Item pair and factor scale statistics for Key Stage subgroups

Key Stage 3 n = 891					Key Stage 4 n = 507				
Item	Mean	Std. Deviation	Skewness	Kurtosis	Item	Mean	Std. Deviation	Skewness	Kurtosis
GlobalSc	2.712	0.761	alpha =	0.732	GlobalSc	2.832	0.794	alpha =	0.784
SCI1	2.342	0.885	0.422	-0.219	SCI1	2.525	0.849	0.273	-0.010
SCI2	2.811	1.032	0.184	-0.660	SCI2	2.937	1.077	0.276	-0.596
SCI3	2.982	0.906	0.011	-0.253	SCI3	3.034	0.910	-0.029	-0.220
Enquiry3	2.531	0.728	alpha =	0.702	Enquiry3	2.615	0.716	alpha =	0.648
ENQ1	2.811	0.986	0.292	-0.420	ENQ1	2.982	0.994	0.302	-0.575
ENQ2	3.068	0.869	0.042	-0.229	ENQ2	3.095	0.815	-0.015	0.049
ENQ3	2.402	0.887	0.471	0.125	ENQ3	2.622	0.759	-0.004	-0.211
ENQ4	2.518	0.831	0.375	-0.244	ENQ4	2.540	0.736	0.247	-0.260
ENQ5	2.439	0.885	0.185	-0.487	ENQ5	2.458	0.788	0.092	-0.505
ENQ6	2.637	0.906	0.239	-0.246	ENQ6	2.846	0.800	0.080	-0.188
Biology	2.501	0.722	alpha =	0.742	Biology	2.573	0.735	alpha =	0.784
BIOL1	2.566	0.986	0.387	-0.235	BIOL1	2.473	0.995	0.473	-0.274
BIOL2	2.398	0.898	0.378	-0.290	BIOL2	2.687	0.830	0.208	-0.063
BIOL3	2.279	0.853	0.450	-0.073	BIOL3	2.439	0.904	0.410	-0.140
BIOL4	2.762	0.859	0.199	-0.083	BIOL4	2.693	0.867	0.257	-0.148
Physics	2.383	0.669	alpha =	0.708	Physics	2.645	0.718	alpha =	0.777
PHYS1	2.699	0.955	0.324	-0.095	PHYS1	2.840	1.025	0.036	-0.585
PHYS2	2.291	0.836	0.398	-0.061	PHYS2	2.560	0.789	0.051	-0.195
PHYS3	2.348	0.825	0.337	-0.013	PHYS3	2.556	0.790	0.205	-0.065
PHYS4	2.194	0.823	0.452	-0.058	PHYS4	2.625	0.838	0.101	-0.444
Chemistry	2.577	0.668	alpha =	0.709	Chemistry	2.746	0.714	alpha =	0.772
CHEM1	2.586	0.993	0.214	-0.406	CHEM1	2.802	0.998	0.150	-0.354
CHEM2	2.727	0.874	0.199	-0.105	CHEM2	2.722	0.821	0.145	-0.168
CHEM3	2.323	0.790	0.436	0.431	CHEM3	2.705	0.924	-0.011	-0.445
CHEM4	2.670	0.907	0.256	-0.205	CHEM4	2.755	0.955	0.159	-0.330
Nat of Sci	2.630	0.704	alpha =	0.734	N of Sci	2.751	0.684	alpha =	0.779
NOS1	2.458	0.915	0.303	-0.317	NOS1	2.581	0.834	0.108	-0.213
NOS2	2.807	0.872	0.202	-0.032	NOS2	2.902	0.836	0.051	-0.120
NOS3	2.624	0.826	0.186	-0.230	NOS3	2.794	0.795	-0.198	-0.301
Imp of Sci	1.870	0.804	alpha =	0.819	Imp of Sci	1.975	0.840	alpha =	0.829
IOS1	1.880	0.899	0.892	0.097	IOS1	1.971	0.931	0.901	0.135
IOS2	1.861	0.847	0.937	0.513	IOS2	1.979	0.886	0.848	0.364

Table 4.7c Item pair and factor scale statistics for Sex subgroups

Girls n = 687					Boys n = 711				
Item	Mean	Std. Deviation	Skewness	Kurtosis	Item	Mean	Std. Deviation	Skewness	Kurtosis
GlobalSc	2.906	0.748	alpha =	0.728	GlobalSc	2.609	0.774	alpha =	0.759
SCI1	2.521	0.877	0.249	-0.321	SCI1	2.300	0.863	0.469	0.055
SCI2	3.051	1.036	0.141	-0.641	SCI2	2.668	1.030	0.327	-0.529
SCI3	3.147	0.867	-0.032	-0.268	SCI3	2.861	0.924	0.074	-0.201
Enquiry3	2.611	0.704	alpha =	0.711	Enquiry3	2.514	0.726	alpha =	0.662
ENQ1	3.053	0.967	0.221	-0.519	ENQ1	2.699	0.986	0.409	-0.337
ENQ2	3.207	0.846	0.032	-0.221	ENQ2	2.953	0.835	-0.001	-0.081
ENQ3	2.566	0.831	0.199	-0.007	ENQ3	2.400	0.859	0.394	0.022
ENQ4	2.519	0.779	0.305	-0.259	ENQ4	2.533	0.817	0.359	-0.200
ENQ5	2.480	0.866	0.143	-0.478	ENQ5	2.412	0.835	0.160	-0.462
ENQ6	2.833	0.847	0.054	-0.219	ENQ6	2.597	0.886	0.278	-0.160
Biology	2.537	0.694	alpha =	0.732	Biology	2.518	0.762	alpha =	0.775
BIOL1	2.525	0.963	0.505	-0.136	BIOL1	2.540	1.016	0.339	-0.362
BIOL2	2.543	0.886	0.233	-0.317	BIOL2	2.464	0.882	0.332	-0.215
BIOL3	2.294	0.828	0.437	-0.006	BIOL3	2.379	0.917	0.425	-0.201
BIOL4	2.786	0.869	0.201	-0.138	BIOL4	2.690	0.855	0.234	-0.081
Physics	2.612	0.686	alpha =	0.739	Physics	2.348	0.673	alpha =	0.727
PHYS1	2.936	0.963	0.140	-0.284	PHYS1	2.570	0.969	0.321	-0.265
PHYS2	2.522	0.829	0.137	-0.345	PHYS2	2.259	0.810	0.370	0.066
PHYS3	2.540	0.804	0.205	-0.067	PHYS3	2.310	0.817	0.363	0.026
PHYS4	2.452	0.846	0.176	-0.534	PHYS4	2.253	0.851	0.459	-0.017
Chemistry	2.731	0.677	alpha =	0.750	Chemistry	2.548	0.696	alpha =	0.713
CHEM1	2.796	0.979	0.152	-0.339	CHEM1	2.537	1.005	0.252	-0.408
CHEM2	2.821	0.841	0.203	-0.023	CHEM2	2.632	0.858	0.182	-0.204
CHEM3	2.486	0.840	0.425	0.139	CHEM3	2.438	0.880	0.224	-0.277
CHEM4	2.820	0.932	0.166	-0.297	CHEM4	2.586	0.905	0.271	-0.184
Nat of Sci	2.764	0.676	alpha =	0.751	N of Sci	2.591	0.712	alpha =	0.745
NOS1	2.576	0.878	0.143	-0.305	NOS1	2.431	0.892	0.315	-0.242
NOS2	2.940	0.836	0.131	-0.055	NOS2	2.747	0.873	0.189	-0.066
NOS3	2.777	0.766	-0.059	-0.150	NOS3	2.598	0.858	0.181	-0.394
Imp of Sci	1.901	0.836	alpha =	0.856	Imp of Sci	1.914	0.801	alpha =	0.791
IOS1	1.929	0.910	0.875	0.127	IOS1	1.897	0.913	0.921	0.126
IOS2	1.874	0.878	0.999	0.739	IOS2	1.933	0.848	0.814	0.182

Table 4.8 ITEM-PAIR '1' Scores for physics, chemistry and biology and ITEM-PAIR '2 - 4' average scores.

Item	Girls n = 687	Boys n = 711	KS3 n = 891	KS4 n = 507	All pupils n = 1398
PHYS1	2.936	2.570	2.699	2.840	2.750
CHEM1	2.796	2.537	2.586	2.802	2.665
BIOL1	2.525	2.540	2.566	2.473	2.532
SCIENCE	2.906	2.609	2.712	2.832	2.755
Average PHYS 2-4	2.505	2.274	2.278	2.580	2.378
Average CHEM 2-4	2.709	2.552	2.573	2.727	2.629
Average BIOL 2-4	2.541	2.511	2.480	2.606	2.526

Table 4.8 shows the ITEM-PAIR 1 scores and the average ITEM-PAIR 2-4 scores. Notice that the scores for PHYS1 have the highest values of the scores in the upper table (i.e. least positive), whilst PHYS 2-4 has the lowest value of the scores in the lower table (i.e. most positive).

N.B. ITEM-PAIR 1 refers to the most generic item-pair of each factor, i.e. items which used the words Physics, Chemistry and Biology. ITEM-PAIR 2-4 refers to the mean value of the three higher specificity items-pairs in each factor.

Table 4.9 Effect size between KS4 and KS3 for generic statements and high specificity statements

Outcome measure	Key Stage 4			Key Stage 3			Effect Size between scores
	mean	n	SD	mean	n	SD	
SCI 1-3	2.83	507.00	0.95	2.71	981.00	0.94	**0.13**
NOS 1-3	2.76	507.00	0.87	2.63	981.00	0.87	**0.15**
BIOL 2-4	2.61	507.00	0.87	2.48	981.00	0.87	**0.15**
PHYS 2-4	2.58	507.00	0.81	2.28	981.00	0.83	**0.37**
CHEM 2-4	2.73	507.00	0.90	2.57	981.00	0.86	**0.18**

Table 4.9 shows the effect size between KS4 and KS3 for the generic statements was 0.13 and 0.14, and the effect size for the specific statements was 0.15, 0.37 and 0.18. There appears not to be a change in pattern between the general and the specific which might indicate the absence of a curriculum-exposure effect. Table 4.9 clearly shows that the KS3 pupils remained slightly more positive than the KS4 pupils and that the differences between their positiveness remained about the same across all elements with a size effect of around 0.15.

Six self-concept scales and one Importance of Science scale were constructed from the item pairs. The descriptive statistics for the item pairs together with those of the newly formed scales are shown in Table 4.7a. Not expectedly, the skewness and kurtosis values moved closer to normality following the averaging process. The formation of scales from items, as expected, had the tendency to move distributions away from extremes. Table 4.7a, as indeed do all the following tables, show data after missing values had been imputed using the EM algorithm and as such the data set now includes full entries, either real or predicted, for all 1398 pupils. The reliability coefficients (Cronbach Alpha) for the newly formed scales are shown in the same table. All Alpha coefficients are high; Nunnaly (1978) has given a value greater than 0.70 as a bench mark to achieve a satisfactory reliability outcome. Values can be seen in Tables 4.7 a, b and c. All but one of the Alpha values was clearly in excess of 0.73 suggesting the scales to be unidimensional in nature, which was a very positive result. The exception to this was the Enquiry scale which fell to a value around 0.65 with KS4 pupils and with boys. This just clips beneath Nunnaly's threshold but does not give cause for concern. Enquiry scale was constituted from only three of the item pairs and not the entire six items as might be assumed from the table. This arose because the SEM calculations carried out as part of the model testing procedures to be reported later showed excessively high correlations (> 0.95) between Global Science and ENQ1, ENQ2 and ENQ3. These three items asked for responses to 'science investigations'. The high correlations indicated that the pupils were not discriminating between items which asked about 'science' and items which asked about 'science investigations'. For this reason it was thought best to remove items pairs ENQ1, ENQ2 and ENQ3 as they made no positive contribution and introduced unacceptable error covariances. Only the remaining enquiry item pairs ENQ4, ENQ5 and ENQ6 were used to form the Enquiry scale, which from now on will be known as Enquiry3. These items contained statements which had a higher level of specificity that the ENQ1-3 items and as such did not create the same difficulties. Tables 4.7b and 4.7c show the item pair and whole scale statistics for the subgroups of Key Stage and Sex respectively. Consistency within and across subgroups was important as self-concept responses have been shown to vary both for age (Crain and Bracken, 1994; Marsh, Barnes, Cairns and Tidman, 1984) and for sex (Marsh, 1985). Indeed, the assumption that comparisons of self-concept scores can be made across subgroups assumes that the measurement and factor structure of self-concept (observed through the factor loadings and factor covariances) are

equivalent. If these assumptions were not fulfilled then the comparisons would not have been valid (see Byrne and Shavelson, 1987; Marsh, Smith and Barnes, 1985). The Cronbach alpha coefficients for each of the scales are shown separately for boys and girls in Table 4.7c. The coefficients provide evidence for the internal consistency for each of the scales for younger KS3 pupils (α = 0.70 to 0.82) and older KS4 pupils (α = 0.65 to 0.83), and likewise, for boys (α = 0.66 to 0.79) and for girls (α = 0.71 to 0.86). All α-coefficients were reassuringly high providing evidence for their applicability across subgroups. Further sub-group analysis will be carried out as part of the cross-validation testing below.

Reliability of measurement was also evaluated by undertaking and estimating test-retest reliabilities. 192 pupils or 13.7% of the sample took part in the test re-test reliability procedures. Table 4.10 summarises the results.

Table 4.10 Test – Retest Statistics for all items

		Mean of all science items
		All re-test sample
Mean Science score Total 1	Pearson Correlation Sig. (2 tailed) N	**.778**** .000 191
Total N		192
		Mean of science ASDQ items
Mean ASDQ score Total 1	Pearson Correlation Sig. (2 tailed) N	**.764**** .000 189
Total N		192
**Correlation is significant at the 0.01 level (2-tailed)		

The test-retest correlation for the mean of the all science items was 0.778 and significant at the 0.01 level. By way of comparison, the Marsh ASDQ science items were test and retested at the same instances with a correlation of 0.764, also significant at the 0.01 level. The time between test and re-test for the sample was between four and eight weeks. The new science items compared favourably with those of the well established ASDQ items which indicated a positive sign for the reliability measures of the scales.

Item Values and Comparison between Subgroups

Individual item pair scores together with combined factor scores are represented in Charts 4.1 and 4.2 respectively. The accompanying tabular scores for sex and Key Stage subgroups are to be found in Tables 4.11 and 4.11b. Tables 4.12a and 4.12b show, amongst other measures, a 2-tailed T-test for difference between subgroup means, together with effect sizes of subgroup mean differences. The results shown in Charts 4.1 and 4.2 and Tables 4.11 a, b and 4.12 a, b are discussed below.

There are statistically significant differences between subgroup factor scores. A comparison between boys and girls shows that at the level of factors, five of the six factors showed significant differences. However, although statistically significant, the differences between means showed weak to modest effect sizes. Three factors exhibited the largest differences, these were Global Science and Physics with a mean scale difference of 0.30 and 0.26 respectively, representing a size effect for each of 0.39, and Chemistry with a mean scale difference of 0.18 representing a size effect of 0.27. To put these size effect in context, Cohen (1988) quotes that an effect size of 0.2 equates to the difference in heights between 15 year old and 16 year old girls in the US, and an effect size of 0.5 equates to a difference in heights between 13 year old and 18 year old girls. In a different context Coe (2000) calculates that for mathematics and English GCSE grades, an effect size of 0.6 equates to a difference of around one grade.

The physics and chemistry scores are not unexpected although the Global science grade was not such a predicted outcome. An examination of the individual item pairs in Chart 4.1 (and Table 4.11a) clearly indicates that all three components of Global Science show large differences with the boys being statistically significantly more positive in the absolute, internal and externally framed questions. This is perhaps the most surprising outcome as it was not expected that the effect size of the Global Science factor would rival the physics factor for sex group differences.

Chart 4.1 Item-pair values for gender and age subgroups

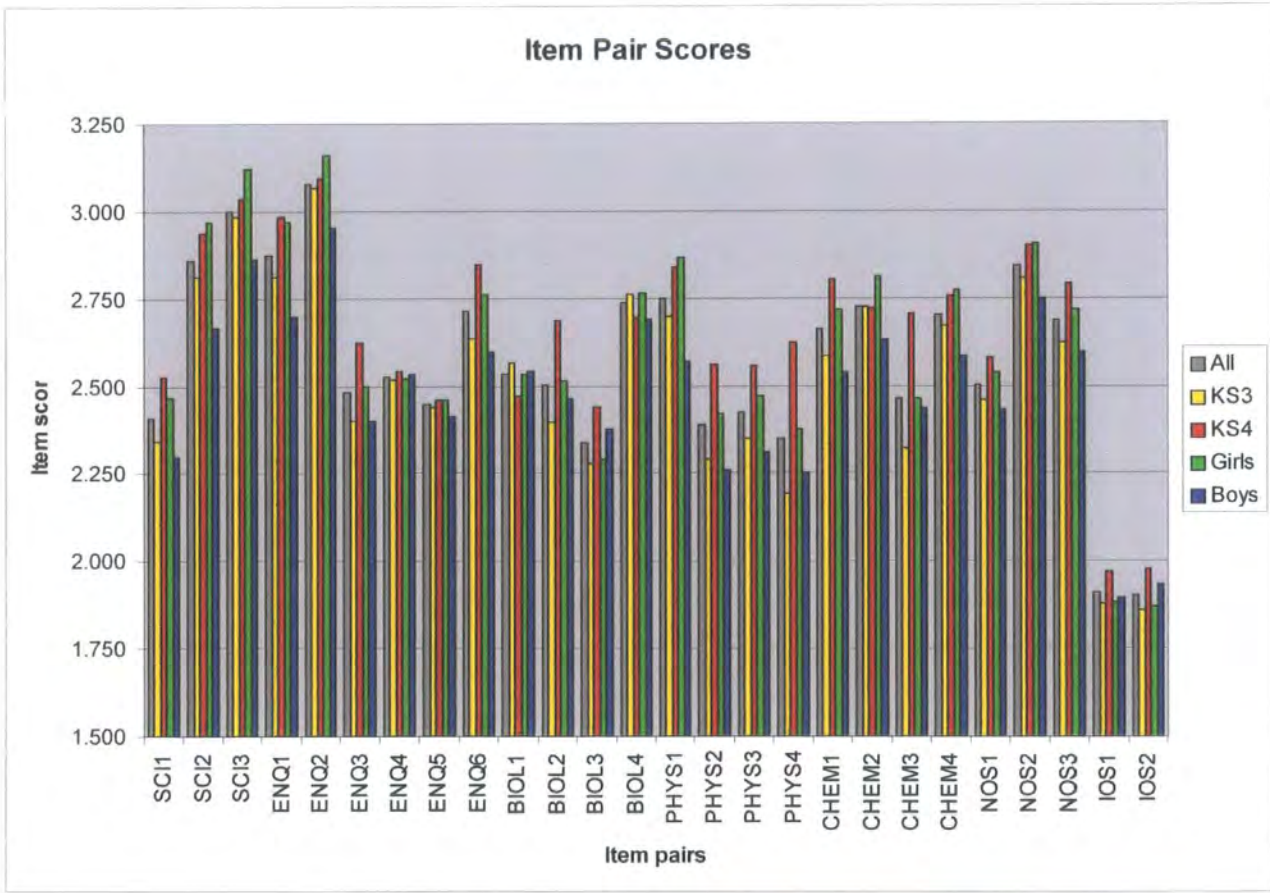


Table 4.11a

Item-pair values for gender and age subgroups

Item-pair	All	KS3	KS4	Girls	Boys
SCI1	2.409	2.342	2.525	2.468	2.300
SCI2	2.857	2.811	2.937	2.967	2.668
SCI3	3.001	2.982	3.034	3.120	2.861
ENQ1	2.873	2.811	2.982	2.970	2.699
ENQ2	3.078	3.068	3.095	3.161	2.953
ENQ3	2.481	2.402	2.622	2.498	2.400
ENQ4	2.526	2.518	2.540	2.522	2.533
ENQ5	2.446	2.439	2.458	2.459	2.412
ENQ6	2.713	2.637	2.846	2.761	2.597
BIOL1	2.532	2.566	2.473	2.533	2.540
BIOL2	2.503	2.398	2.687	2.512	2.464
BIOL3	2.337	2.279	2.439	2.290	2.379
BIOL4	2.737	2.762	2.693	2.766	2.690
PHYS1	2.750	2.699	2.840	2.865	2.570
PHYS2	2.388	2.291	2.560	2.420	2.259
PHYS3	2.423	2.348	2.556	2.471	2.310
PHYS4	2.350	2.194	2.625	2.376	2.253
CHEM1	2.664	2.586	2.802	2.716	2.537
CHEM2	2.725	2.727	2.722	2.811	2.632
CHEM3	2.462	2.323	2.705	2.462	2.438
CHEM4	2.701	2.670	2.755	2.773	2.586
NOS1	2.502	2.458	2.581	2.538	2.431
NOS2	2.842	2.807	2.902	2.906	2.747
NOS3	2.686	2.624	2.794	2.717	2.598
IOS1	1.913	1.880	1.971	1.884	1.897
IOS2	1.904	1.861	1.979	1.873	1.933

Chart 4.2 Factor values for gender and age subgroups

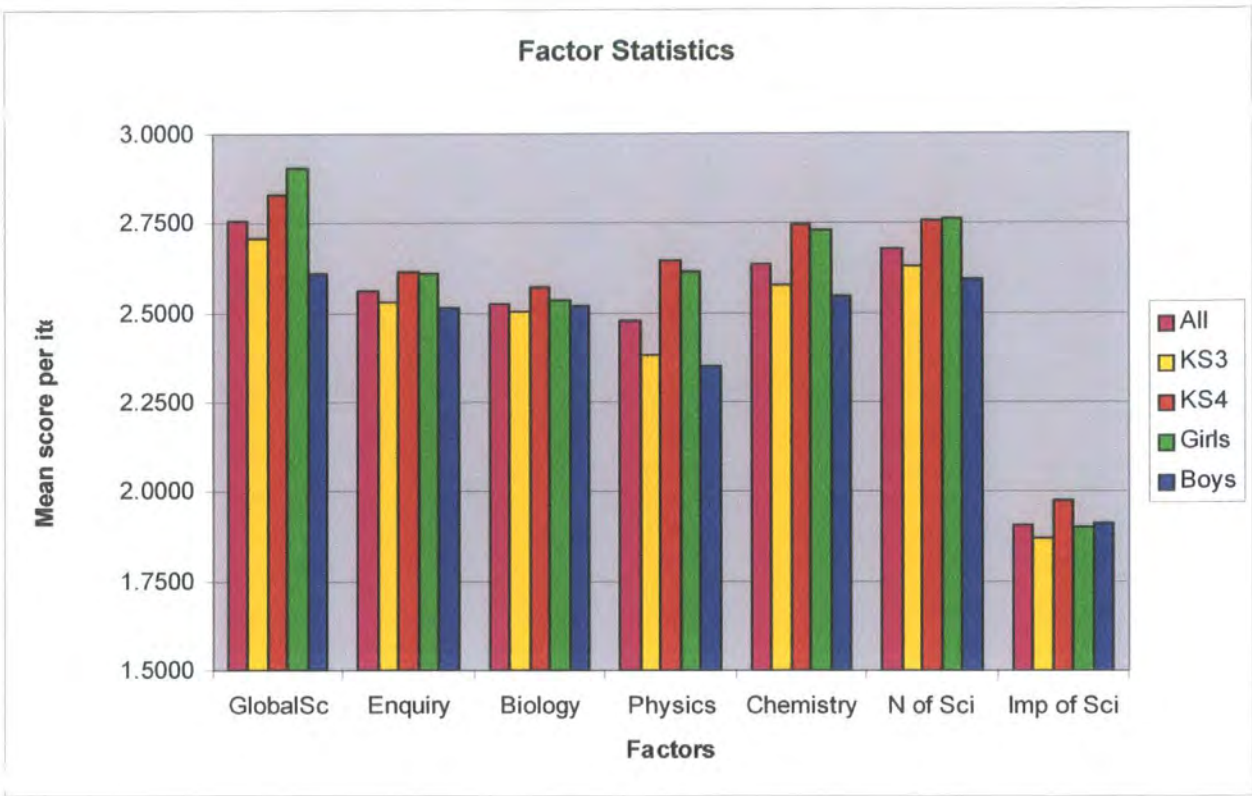


Table 4.11b

Factor values for gender and age subgroups

Factors	All	KS3	KS4	Girls	Boys
Global Sci	2.7554	2.712	2.832	2.906	2.609
Enquiry	2.5615	2.760	2.899	2.942	2.683
Biology	2.5273	2.414	2.533	2.454	2.461
Physics	2.4780	2.446	2.652	2.666	2.379
Chemistry	2.6380	2.545	2.743	2.701	2.535
Nat of Sci	2.6766	2.630	2.759	2.764	2.591
Imp of Sci	1.9084	1.870	1.975	1.901	1.914

Table 4.12a **Effect size of item and factor difference Subgroup: SEX**

Outcome measure	DATA ENTRY: Treatment group mean	Treatment group n	Treatment group SD	Control group mean	Control group n	Control group SD	RAW DIFFERENCE: pooled standard deviation	p-value for difference in SDs	Mean Difference	p-value for mean diff (2-tailed T-test)	Confidence Interval for Difference lower	Confidence Interval for Difference upper	STANDARDISED EFFECT SIZE: Effect Size	Bias corrected (Hedges)	Standard Error of E.S. estimate	Confidence Interval for Effect Size lower	Confidence Interval for Effect Size upper	Effect Size based on control gp SD
GlobalSc	2.906	687	0.748	2.609	711	0.774	0.76	0.18	0.30	0.00	0.22	0.38	0.39	**0.39**	0.05	0.28	0.50	0.38
SCI1	2.521	687	0.877	2.300	711	0.863	0.87	0.34	0.22	0.00	0.13	0.31	0.25	**0.25**	0.05	0.15	0.36	0.26
SCI2	3.051	687	1.036	2.668	711	1.03	1.03	0.44	0.38	0.00	0.27	0.49	0.37	**0.37**	0.05	0.26	0.48	0.37
SCI3	3.147	687	0.867	2.861	711	0.924	0.90	0.05	0.29	0.00	0.19	0.38	0.32	**0.32**	0.05	0.21	0.42	0.31
Enquiry	2.611	687	0.704	2.514	711	0.726	0.72	0.21	0.10	0.01	0.02	0.17	0.14	**0.14**	0.05	0.03	0.24	0.13
ENQ1	3.053	687	0.967	2.699	711	0.986	0.98	0.30	0.35	0.00	0.25	0.46	0.36	**0.36**	0.05	0.26	0.47	0.36
ENQ2	3.207	687	0.846	2.953	711	0.835	0.84	0.36	0.25	0.00	0.17	0.34	0.30	**0.30**	0.05	0.20	0.41	0.30
ENQ3	2.566	687	0.831	2.400	711	0.859	0.85	0.19	0.17	0.00	0.08	0.25	0.20	**0.20**	0.05	0.09	0.30	0.19
ENQ4	2.519	687	0.779	2.533	711	0.817	0.80	0.10	-0.01	0.74	-0.10	0.07	-0.02	**-0.02**	0.05	-0.12	0.09	-0.02
ENQ5	2.480	687	0.866	2.412	711	0.835	0.85	0.17	0.07	0.14	-0.02	0.16	0.08	**0.08**	0.05	-0.02	0.18	0.08
ENQ6	2.833	687	0.847	2.597	711	0.886	0.87	0.12	0.24	0.00	0.15	0.33	0.27	**0.27**	0.05	0.17	0.38	0.27
Biology	2.537	687	0.694	2.518	711	0.762	0.73	0.01	0.02	0.63	-0.06	0.10	0.03	**0.03**	0.05	-0.08	0.13	0.02
BIOL1	2.525	687	0.963	2.540	711	1.016	0.99	0.08	-0.02	0.78	-0.12	0.09	-0.02	**-0.02**	0.05	-0.12	0.09	-0.01
BIOL2	2.543	687	0.886	2.464	711	0.882	0.88	0.45	0.08	0.10	-0.01	0.17	0.09	**0.09**	0.05	-0.02	0.19	0.09
BIOL3	2.294	687	0.828	2.379	711	0.917	0.87	0.00	-0.09	0.07	-0.18	0.01	-0.10	**-0.10**	0.05	-0.20	0.01	-0.09
BIOL4	2.786	687	0.869	2.690	711	0.855	0.86	0.33	0.10	0.04	0.01	0.19	0.11	**0.11**	0.05	0.01	0.22	0.11
Physics	2.612	687	0.686	2.348	711	0.673	0.68	0.31	0.26	0.00	0.19	0.34	0.39	**0.39**	0.05	0.28	0.49	0.39
PHYS1	2.936	687	0.963	2.570	711	0.969	0.97	0.44	0.37	0.00	0.26	0.47	0.38	**0.38**	0.05	0.27	0.48	0.38
PHYS2	2.522	687	0.829	2.259	711	0.81	0.82	0.27	0.26	0.00	0.18	0.35	0.32	**0.32**	0.05	0.22	0.43	0.32
PHYS3	2.540	687	0.804	2.310	711	0.817	0.81	0.34	0.23	0.00	0.14	0.32	0.28	**0.28**	0.05	0.18	0.39	0.28
PHYS4	2.452	687	0.846	2.253	711	0.851	0.85	0.44	0.20	0.00	0.11	0.29	0.23	**0.23**	0.05	0.13	0.34	0.23
Chemistry	2.731	687	0.677	2.548	711	0.696	0.69	0.23	0.18	0.00	0.11	0.26	0.27	**0.27**	0.05	0.16	0.37	0.26
CHEM1	2.796	687	0.979	2.537	711	1.005	0.99	0.24	0.26	0.00	0.15	0.36	0.26	**0.26**	0.05	0.16	0.37	0.26
CHEM2	2.821	687	0.841	2.632	711	0.858	0.85	0.30	0.19	0.00	0.10	0.28	0.22	**0.22**	0.05	0.12	0.33	0.22
CHEM3	2.486	687	0.84	2.438	711	0.88	0.86	0.11	0.05	0.30	-0.04	0.14	0.06	**0.06**	0.05	-0.05	0.16	0.05
CHEM4	2.820	687	0.932	2.586	711	0.905	0.92	0.22	0.23	0.00	0.14	0.33	0.25	**0.25**	0.05	0.15	0.36	0.26
N of Sci	2.764	687	0.676	2.591	711	0.712	0.69	0.09	0.17	0.00	0.10	0.25	0.25	**0.25**	0.05	0.14	0.35	0.24
NOS1	2.576	687	0.878	2.431	711	0.892	0.89	0.34	0.15	0.00	0.05	0.24	0.16	**0.16**	0.05	0.06	0.27	0.16
NOS2	2.940	687	0.836	2.747	711	0.873	0.86	0.13	0.19	0.00	0.10	0.28	0.23	**0.23**	0.05	0.12	0.33	0.22
NOS3	2.777	687	0.766	2.598	711	0.858	0.81	0.00	0.18	0.00	0.09	0.26	0.22	**0.22**	0.05	0.11	0.32	0.21
Imp of Sci	1.901	687	0.836	1.914	711	0.801	0.82	0.13	-0.01	0.77	-0.10	0.07	-0.02	**-0.02**	0.05	-0.12	0.09	-0.02
IOS1	1.929	687	0.91	1.897	711	0.913	0.91	0.47	0.03	0.51	-0.06	0.13	0.04	**0.04**	0.05	-0.07	0.14	0.04
IOS2	1.874	687	0.878	1.933	711	0.848	0.86	0.18	-0.06	0.20	-0.15	0.03	-0.07	**-0.07**	0.05	-0.17	0.04	-0.07

Table 4.12b **Effect size of item and factor difference Subgroup: KEY STAGE**

DATA ENTRY							RAW DIFFERENCE						STANDARDISED EFFECT SIZE					
Outcome measure	Treatment group KS 4			Control group KS 3			pooled standard deviation	p-value for difference in SDs	Mean Difference	p-value for mean diff (2-tailed T-test)	Confidence Interval for Difference		Effect Size	Bias corrected (Hedges)	Standard Error of E.S. estimate	Confidence Interval for Effect Size		Effect Size based on control gp SD
	mean	n	SD	mean	n	SD					lower	upper				lower	upper	
GlobalSc	2.832	507	0.794	2.712	891	0.761	0.77	0.14	0.12	0.01	0.04	0.20	0.16	**0.16**	0.06	0.05	0.26	0.16
SCI1	2.525	507	0.849	2.342	891	0.885	0.87	0.15	0.18	0.00	0.09	0.28	0.21	**0.21**	0.06	0.10	0.32	0.21
SCI2	2.937	507	1.077	2.811	891	1.032	1.05	0.14	0.13	0.03	0.01	0.24	0.12	**0.12**	0.06	0.01	0.23	0.12
SCI3	3.034	507	0.91	2.982	891	0.906	0.91	0.46	0.05	0.30	-0.05	0.15	0.06	**0.06**	0.06	-0.05	0.17	0.06
Enquiry	2.615	507	0.716	2.531	891	0.728	0.72	0.34	0.08	0.04	0.01	0.16	0.12	**0.12**	0.06	0.01	0.23	0.12
ENQ1	2.982	507	0.994	2.811	891	0.986	0.99	0.42	0.17	0.00	0.06	0.28	0.17	**0.17**	0.06	0.06	0.28	0.17
ENQ2	3.095	507	0.815	3.068	891	0.869	0.85	0.05	0.03	0.57	-0.07	0.12	0.03	**0.03**	0.06	-0.08	0.14	0.03
ENQ3	2.622	507	0.759	2.402	891	0.887	0.84	0.00	0.22	0.00	0.13	0.31	0.26	**0.26**	0.06	0.15	0.37	0.25
ENQ4	2.54	507	0.736	2.518	891	0.831	0.80	0.00	0.02	0.62	-0.07	0.11	0.03	**0.03**	0.06	-0.08	0.14	0.03
ENQ5	2.458	507	0.788	2.439	891	0.885	0.85	0.00	0.02	0.69	-0.07	0.11	0.02	**0.02**	0.06	-0.09	0.13	0.02
ENQ6	2.846	507	0.8	2.637	891	0.906	0.87	0.00	0.21	0.00	0.11	0.30	0.24	**0.24**	0.06	0.13	0.35	0.23
Biology	2.573	507	0.735	2.501	891	0.722	0.73	0.33	0.07	0.08	-0.01	0.15	0.10	**0.10**	0.06	-0.01	0.21	0.10
BIOL1	2.473	507	0.995	2.566	891	0.986	0.99	0.41	-0.09	0.09	-0.20	0.01	-0.09	**-0.09**	0.06	-0.20	0.02	-0.09
BIOL2	2.687	507	0.83	2.398	891	0.898	0.87	0.02	0.29	0.00	0.19	0.38	0.33	**0.33**	0.06	0.22	0.44	0.32
BIOL3	2.439	507	0.904	2.279	891	0.853	0.87	0.07	0.16	0.00	0.06	0.26	0.18	**0.18**	0.06	0.07	0.29	0.19
BIOL4	2.693	507	0.867	2.762	891	0.859	0.86	0.41	-0.07	0.15	-0.16	0.03	-0.08	**-0.08**	0.06	-0.19	0.03	-0.08
Physics	2.645	507	0.718	2.383	891	0.669	0.69	0.04	0.26	0.00	0.19	0.34	0.38	**0.38**	0.06	0.27	0.49	0.39
PHYS1	2.84	507	1.025	2.699	891	0.955	0.98	0.04	0.14	0.01	0.03	0.25	0.14	**0.14**	0.06	0.03	0.25	0.15
PHYS2	2.56	507	0.789	2.291	891	0.836	0.82	0.07	0.27	0.00	0.18	0.36	0.33	**0.33**	0.06	0.22	0.44	0.32
PHYS3	2.556	507	0.79	2.348	891	0.825	0.81	0.14	0.21	0.00	0.12	0.30	0.26	**0.26**	0.06	0.15	0.37	0.25
PHYS4	2.625	507	0.838	2.194	891	0.823	0.83	0.33	0.43	0.00	0.34	0.52	0.52	**0.52**	0.06	0.41	0.63	0.52
Chemistry	2.746	507	0.714	2.577	891	0.668	0.69	0.05	0.17	0.00	0.09	0.24	0.25	**0.25**	0.06	0.14	0.36	0.25
CHEM1	2.802	507	0.998	2.586	891	0.993	0.99	0.45	0.22	0.00	0.11	0.32	0.22	**0.22**	0.06	0.11	0.33	0.22
CHEM2	2.722	507	0.821	2.727	891	0.874	0.86	0.06	0.00	0.92	-0.10	0.09	-0.01	**-0.01**	0.06	-0.11	0.10	-0.01
CHEM3	2.705	507	0.924	2.323	891	0.79	0.84	0.00	0.38	0.00	0.29	0.47	0.45	**0.45**	0.06	0.34	0.56	0.48
CHEM4	2.755	507	0.955	2.67	891	0.907	0.92	0.10	0.09	0.10	-0.02	0.19	0.09	**0.09**	0.06	-0.02	0.20	0.09
N of Sci	2.751	507	0.684	2.63	891	0.704	0.70	0.23	0.12	0.00	0.04	0.20	0.17	**0.17**	0.06	0.06	0.28	0.17
NOS1	2.581	507	0.834	2.458	891	0.915	0.89	0.01	0.12	0.01	0.03	0.22	0.14	**0.14**	0.06	0.03	0.25	0.13
NOS2	2.902	507	0.836	2.807	891	0.872	0.86	0.14	0.10	0.05	0.00	0.19	0.11	**0.11**	0.06	0.00	0.22	0.11
NOS3	2.794	507	0.795	2.624	891	0.826	0.81	0.17	0.17	0.00	0.08	0.26	0.21	**0.21**	0.06	0.10	0.32	0.21
Imp of Sci	1.975	507	0.84	1.87	891	0.804	0.82	0.14	0.11	0.02	0.02	0.19	0.13	**0.13**	0.06	0.02	0.24	0.13
IOS1	1.971	507	0.931	1.88	891	0.899	0.91	0.19	0.09	0.07	-0.01	0.19	0.10	**0.10**	0.06	-0.01	0.21	0.10
IOS2	1.979	507	0.886	1.861	891	0.847	0.86	0.13	0.12	0.01	0.02	0.21	0.14	**0.14**	0.06	0.03	0.25	0.14

pairs exhibit non-significant differences, although the strength of the third pair, with a size effect of 0.27 tips the balance. The biology scale shows no significant difference and a very weak size effect of 0.03. The biology factor was the most positively scored by the girls whereas for the boys, with the exception of physics which was by far and away the most positive response, all the other boys' factor scores were very similar. In fact, there were no significant differences between Enquiry3, Biology, Chemistry and Nature of Science for the boys at the 0.05 level.

The Importance of Science scale showed no significant difference between the sexes with an extremely weak size effect of 0.02. In absolute terms the Importance of Scale was scored substantially more positive than any other single factor. The effect size for boys between IoS and Physics (the next most positive) was 0.59 and the size effect for girls between IoS and Biology (the next most positive for girls) was 0.83. These are effect sizes far in excess of any other difference found in the results tables. It is unsafe to draw too much from these data, since the IoS items were different in nature to the self-concept items and the importance of other subjects were not measured. It is not possible then to gauge the importance that individuals placed on science compared with their other curriculum subjects. Even given this caveat, the IoS items were the only items to secure average scores more positive than 2.0 on the 1 to 5 Likert scale and the differences are so marked that it is difficult not to conclude that both sexes perceive science, at the very least, to be an important school subject.

Differences in means between the Key Stage 3 and Key Stage 4 subgroups across the factors were generally speaking slightly more closely matched than for the sex subgroups. The two exceptions to this were Physics where the effect size between the age groups was almost identical to that which was seen between the sex groups at 0.38, and Chemistry where the effect size was slightly smaller at 0.25 (reduced from 0.27). All factors except for Biology showed significant differences at the 0.05 level, even given the reduced effect sizes. Enquiry3, as with the comparison between sex subgroups, remained perched on the borderline with a significance *p value* = 0.04. In a copycat to the sex group differences two of the three Enquiry3 item pairs were non-significant with the same item pair responsible for the difference in values. A re-examination of the original single items showed that in both subgroup cases it was the individual item EAP4 that was responsible for the excessive and statistically significant difference between the groups. This item was 'I have no

trouble in thinking up ideas to investigate'. Clearly girls and older pupils perceive they have relatively more trouble than their counterparts. All other effect sizes were less than 0.18 which, although were statistically significantly differences, were modest in size.

In summary, there were significantly statistical differences between the extracted subgroups of sex and age (Key Stage). This outcome is not unexpected as the sample size was large enough to ensure that small statistical differences in value were likely to be significant. The effect sizes between subgroups were weak to modest in size. The largest effect sizes were for physics and chemistry with boys having a more positive disposition than girls and younger pupils more positive than older pupils. Enquiry, Biology and Importance of Science showed modest and sometimes non-significant differences. One noteworthy outcome was that boys showed a much more positive attitude to Global Science, which was rather more unexpected than other outcomes.

Section Four: Model Testing

(a) Parameter Estimation

As was explained in the previous chapter, parameter estimation, in part, comprises the comparison of a matrix of covariances associated with the captured empirical data, S with a matrix of hypothesized covariances estimated from the specifics of the model structure, $\hat{\Sigma}$. The closer the match between the actual matrix and the implied covariant matrices then the better is the model. The default method by which LISREL produces parameter estimates is Maximum likelihood (ML) procedures, and this was adopted for all subsequent analysis.

Each of the FOUR hypothesized models was analyzed in turn in order to undertake the parameter estimation procedure. En route to the matrix comparison a number of mathematical procedures were undertaken in order to determine that the models had been 'well estimated'. Directly below follows a full analysis of the parameter estimation procedures for Model 4. Model 4 was the most complex and least parsimonious of all the models and as such had greatest opportunity for misspecification or problematic parameter estimation. All other models were a subset, or near sub-set of Model 4 and so it made sense to scrutinize Model 4 fully and thoroughly since if Model 4 'passed the test' the likelihood would be that the others would perform similarly. Each stage of the analytical procedures has been explained in some detail such that there is transparency in the reporting and analysis of results. To avoid undue repetition, evaluations of the six other models, although carried out with equal rigour, has been reported in summary form only.

Firstly, and importantly, it can be reported that the model converged and LISREL was able to produce a full data output set for the hypothesized models. This may seem to be a small success to celebrate, but there are enough bear traps at each stage of the process to halt the progress of even the most experienced researcher. LISREL is one of the most unfriendly of analytical programmes with the error messages seemingly designed to be frustratingly unhelpful. When it is being correctly used it provides a wealth of model related data and these will be presented and evaluated systematically over the remainder of the chapter.

The *path diagram* for Model 4 is shown in Diagram 4.1. The path diagram represents, in pictorial form, the relationships between the exogenous latent variables (shown in green), the endogenous latent variables (shown in yellow), the manifest variable (shown in rectangles), whether they be correlations (shown as curves), parameter estimates (shown as straight arrows between variables) or measurement errors (shown pointing to the manifest variables). Diagram 4.1 shows non-standardized parameter estimates, and these can be matched against the same parameter estimates that are represented in the structural equations in Tables 4.13a and b.

The equations displayed within Tables 4.13a and b are from a portion of the LISREL output for Model 4. The first segment of the LISREL output shows estimates in equation form and these equations are grouped in two sections. The upper section shows the *Measurement Equations are* can be found in Table 4.13a. Here, each manifest variable is expressed as a linear function of the underlying latent variable. There are 21 equations in total. Table 4.13b contains a further five *Structural Equations*. These show how each dependent latent variable is expressed as a linear function of the independent latent variables. Beneath the structural equations in Table 4.13c is a correlation matrix which shows the correlations between the exogenous latent variables of General Science (GenSci) and Global Science (GlobalSc). Table 3.13d shows the correlation matrix between all the various latent variables.

Diagram 4.1 Model 4 Path diagram with unstandardised parameter estimates

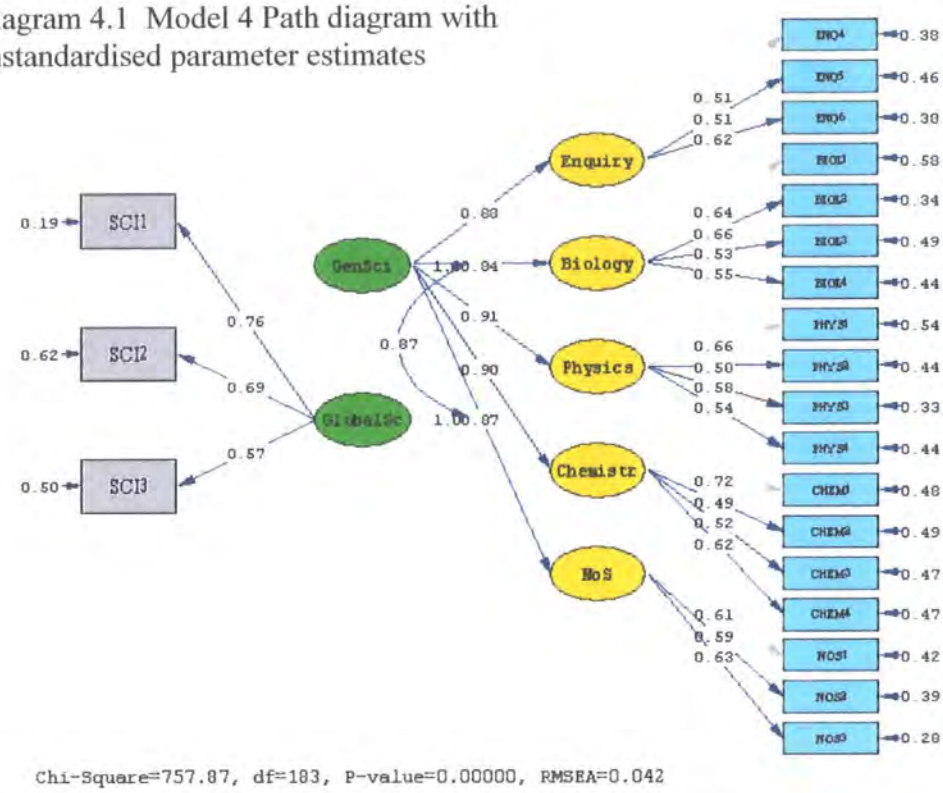


Diagram 4.1 shows the path diagram with unstandardized parameter estimates for Model 4. The green ellipses are exogenious latent variable; the yellow ellipses are the endogenous latent variables; the rectangles are the manifest variables; the curved lines are the correlations, the straight linking lines are the parameter estimates and the input arrows are the measurement errors. Note that the correlation curve is drawn in a concave manner rather than convex to avoid overdrawing the parameter estimates.

Table 4.13a LISREL Estimates (Robust Maximum Likelihood) for Measurement Equations

Measurement Equations

```
ENQ4 = 0.51*Enquiry, Errorvar.= 0.38  , R² = 0.41
                               (0.020)
                                18.99

ENQ5 = 0.51*Enquiry, Errorvar.= 0.46  , R² = 0.36
      (0.029)                  (0.021)
       17.78                    22.40

ENQ6 = 0.62*Enquiry, Errorvar.= 0.38  , R² = 0.50
      (0.030)                  (0.021)
       20.79                    18.45
```

```
BIOL1 = 0.64*Biology, Errorvar.= 0.58  , R² = 0.41
                                (0.028)
                                 20.58

BIOL2 = 0.66*Biology, Errorvar.= 0.34  , R² = 0.56
       (0.030)                  (0.020)
        22.42                    17.17

BIOL3 = 0.53*Biology, Errorvar.= 0.49  , R² = 0.37
       (0.028)                  (0.024)
        19.18                    20.59

BIOL4 = 0.55*Biology, Errorvar.= 0.44  , R² = 0.40
       (0.028)                  (0.021)
        19.64                    20.67
```

```
PHYS1 = 0.66*Physics, Errorvar.= 0.54  , R² = 0.44
                                (0.026)
                                 20.67

PHYS2 = 0.50*Physics, Errorvar.= 0.44  , R² = 0.36
       (0.026)                  (0.019)
        18.93                    22.88

PHYS3 = 0.58*Physics, Errorvar.= 0.33  , R² = 0.50
       (0.026)                  (0.017)
        22.43                    19.06

PHYS4 = 0.54*Physics, Errorvar.= 0.44  , R² = 0.40
       (0.028)                  (0.021)
        19.33                    20.49
```

```
CHEM1 = 0.72*Chemistry, Errorvar.= 0.48  , R² = 0.52
                                  (0.026)
                                   18.21

CHEM2 = 0.49*Chemistry, Errorvar.= 0.49  , R² = 0.33
       (0.027)                    (0.023)
        18.16                      21.62

CHEM3 = 0.52*Chemistry, Errorvar.= 0.47  , R² = 0.37
       (0.027)                    (0.023)
        19.43                      20.25

CHEM4 = 0.62*Chemistry, Errorvar.= 0.47  , R² = 0.45
       (0.027)                    (0.022)
        23.25                      21.40
```

```
NOS1 = 0.61*NoS, Errorvar.= 0.42  , R² = 0.47
                           (0.022)
                            18.95

NOS2 = 0.59*NoS, Errorvar.= 0.39  , R² = 0.47
       (0.029)              (0.020)
        20.17                19.41

NOS3 = 0.63*NoS, Errorvar.= 0.28  , R² = 0.59
       (0.026)              (0.017)
        23.89                16.02
```

```
SCI1 = 0.76*GlobalSc, Errorvar.= 0.19  , R² = 0.76
       (0.020)                   (0.017)
        37.55                     10.90

SCI2 = 0.69*GlobalSc, Errorvar.= 0.62  , R² = 0.43
       (0.025)                   (0.029)
        28.06                     21.89

SCI3 = 0.57*GlobalSc, Errorvar.= 0.50  , R² = 0.39
       (0.025)                   (0.022)
        22.76                     23.03
```

Table 4.13b LISREL Estimates (Robust Maximum Likelihood) for Structural Equations

Structural Equations

```
Enquiry = 0.88*GenSci, Errorvar.= 0.23  , R² = 0.77
          (0.042)                 (0.042)
           20.76                   5.48

Biology = 0.84*GenSci, Errorvar.= 0.29  , R² = 0.71
          (0.038)                 (0.043)
           22.17                   6.72

Physics = 0.91*GenSci, Errorvar.= 0.17  , R² = 0.83
          (0.038)                 (0.038)
           23.62                   4.59

Chemistr = 0.90*GenSci, Errorvar.= 0.18  , R² = 0.82
           (0.033)                 (0.029)
            27.36                   6.28

    NoS = 0.87*GenSci, Errorvar.= 0.24  , R² = 0.76
          (0.036)                 (0.034)
           24.07                   7.01
```

Table 4.13c Correlation Matrix for Exogenous Latent Variables

Correlation Matrix of Independent Variables		
	GenSci	GlobalSc
GenSci	1.00	
GlobalSc	0.87 (0.01) 63.81	1.00

Table 4.13d Correlation Matrix for All Latent Variables

Correlation Matrix of Latent Variables							
	Enquiry	Biology	Physics	Chemistry	NoS	GenSci	GlobalSc
Enquiry	1.00						
Biology	0.74	1.00					
Physics	0.80	0.77	1.00				
Chemistr	0.79	0.76	0.82	1.00			
NoS	0.76	0.73	0.79	0.79	1.00		
GenSci	0.88	0.84	0.91	0.90	0.87	1.00	
GlobalSc	0.77	0.74	0.79	0.79	0.76	0.87	1.00

The equations report three pieces of information for each free[1] parameter, (i) the unstandardized parameter estimate, (ii) its standard error and (iii) the relevant *t*-value. The magnitude of the unstandardized parameter indicates the resulting change in the dependent variable for a unit change in the independent variable when all other independent variables were held constant.

The equations provide considerable data about the measurement and structural aspects of Model 4. The first stage of the analysis was to ensure the feasibility of

[1] For each latent variable the variances of one of the indicator variables is set to unity. This is a scaling procedure which ensures that comparison between latent variables is more meaningful. Fixed parameters do not have standard errors or *t*-values.

parameter estimates. An inspection of the parameter values shows that there were no unreasonable or inappropriate results. Such results would have included negative values for either the error variances or the variances of the latent variables. All error variances were positive and had values ranging from between 0.19 (for SCI1) to 0.62 (for SCI2). When inspecting the model equations, the numbers contained within the parentheses are the values of the standard errors. The size of the standard errors ranged from 0.020 (for SCI1) to 0.042 (for ENQUIRY). If the standard error had approached zero it would have indicated that the test parameter could not be defined (Bentler, 1995); had the standard error been over large then it would have indicated that the parameters could not be determined (Jöreskog and Sörbom, 1989). All standard errors presented were neither excessively small nor excessively large indicating an unproblematic fit situation. The numbers below the standard errors figures in the equations are the *t*-values for the parameter estimates. For the parameter estimate to be significantly different from zero the *t*-values would have been greater than modulus 1.96, i.e. greater that +1.96 or smaller than -1.96. An inspection of the *t*-values in Table 3.13a shows that the *t*-values range from a maximum value of 23.03 (for variable SCI3), to a minimum value of 10.90 (for variable SCI1). Clearly, none of the *t*-values were within modulus 1.96 confirming that all values were indeed significant. The next procedure carried out was a check of the polarity of the sign for the parameter estimates to ensure that they were consistent with the hypothesized relationships predicted by the model. When the model was originally specified it was hypothesized that the correlational relationship between the different variables would be in the positive directional sense. That is, higher scores on the manifest variables would relate to higher values for the latent variables. An inspection of the parameter estimates shows that they were all of the appropriate sign (positive), which is a positive outcome for the model.

The R^2 values within the equations are analogous to values obtained in conventional regression analysis and as such show the amount of variance in the dependent variable accounted for by the independent variable(s) in the equation. Variance not accounted for is due to the measurement error and thus R^2 provides an indication of how free the manifest variables were from measurement error. The closer R^2 moves toward 1.0 then the smaller was the measurement error within the parameter estimate and the better the manifest variables were as indicators of the latent variables. All R^2 values were moderate to high in size, ranging from 0.33 (for CHEM2) to 0.76 (for SCI1). These values indicated that the manifest variables were

reasonably successful as measures of the latent variables in the model. Higher values, i.e. closer to 1.0 would have been welcome, but these results were not unsatisfactory. The R^2 values for the structural equations, that is, those equations expressing relationships between the endogenous and exogenous latent variables were substantially large, (target value greater than 0.5, see Diamantopoulos, 2000). The values in Model 4 were reassuringly high and range from a minimum of 0.71 for BIOLOGY to 0.83 for PHYSICS indicating that the exogenous latent variables explained a considerable proportion of the endogenous latent variables. The correlational matrix for the independent variables showed that there was a high (as predicted) correlation coefficient between GenSci and GlobalSc and that the relationship was significant as indicated by the *t*-value. Finally, the covariance matrix of latent variables showed strong positive relationships between the latent variables as predicted by the hypothesized model.

These last three results provide strong evidence for the hierarchical nature of the model. The parameter estimates from GenSci to each of the endogenous latent variable were very high. The values range from 0.84 (for BIOLOGY) to 0.91 (for PHYSICS). This was a clear indication that the different components of science were able be represented by a single derived variable in a hierarchical relationship. The path diagram (and the structural equations) indicated the relationship between the independently measured Global Science (GlobalSc) construct and the derived hierarchical factor General Science (GenSci). This was possible because they were both included within the same LISREL analysis. The output shows that the correlation between these two higher order constructs has a value of 0.87. More than three quarters of the variance of these two factors are shared which could possibly indicate that it is difficult to distinguish between them. If this is so then it is difficult to argue that these two factors are indeed different constructs which would lend support to the claim for the hierarchical nature of the academic self-concept of science. This hierarchical aspect will be further investigated below.

The LISREL data explored above confirmed that parameter estimation for Model 4 was successful. There was a good appropriate match between the implied model estimates and the empirical data, the parameter estimates were high and in the correct directional orientation, all values were significant, and the standard errors were appropriate in size. This first check of the viability of the model produced results which were very positive and most welcome outcome.

Similar analyses were carried out on Models 1 to 3 and 5 to 7. No difficulties were encountered with any of the models. In short, the parameter estimates were all feasible; the standard errors were appropriate in size, being neither too large nor approaching zero; all t-values were greater than modulus 1.96 indicating that all parameters were significantly different from zero; the R^2 values were moderate to large indicating well represented latent variables; and correlation coefficients between the various latent variables were moderate to large and statistically significant. Path diagrams for Models 1 to 3 and 5 to 7 will be presented later in this chapter.

(b) Testing of Model Fit

As outlined in the previous chapter, the program LISREL 8 produces a substantial number of fit indices. What follows below is a detailed analysis of the evaluation of fit for Model 4. Following the detailed scrutiny of Model 4, summaries of the fit statistics are provided for the other three models. Table 4.14 contains the exact LISREL statistical fit output for Model 4.

Empirical applications of SEM have typically evaluated model fit by using two methods: (i) the conventional likelihood ratio χ^2 test, which hypothesizes that the specified model holds exactly in the population, i.e. the model is perfect; and (ii) various descriptive measures of fit of the model to the sample data. The former method is particularly difficult to satisfy, and as such, there was no more than the very slightest expectation that the conditions were satisfied and the model accepted using this perfect fit criterion.

Within the data output, the first two results generated by LISREL are fit tests based around the χ^2 statistic. Even given the almost certain negative outcome of this test, it is reported first. The model scrutiny will follow the order of the LISREL output format since this provides a coherent grouping of the different kinds of test statistic. LISREL uses four different types of 'discrepancy' in evaluating the model fit. These discrepancies are between different covariance matrices. The LISREL output groups the tests under the different types and uses of discrepancy, and again this provides a useful structure by which to report and evaluate the model fit outcomes.

Table 4.14 Goodness of Fit Statistics for Model 4

Degrees of Freedom = 183
Minimum Fit Function Chi-Square = 704.315 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 757.865 (P = 0.0)
Satorra-Bentler Scaled Chi-Square = 627.170 (P = 0.0)
Chi-Square Corrected for Non-Normality = 590.706 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 444.170
90 Percent Confidence Interval for NCP = (371.789 ; 524.142)

Minimum Fit Function Value = 0.504
Population Discrepancy Function Value (F0) = 0.318
90 Percent Confidence Interval for F0 = (0.266 ; 0.375)
Root Mean Square Error of Approximation (RMSEA) = 0.0417
90 Percent Confidence Interval for RMSEA = (0.0381 ; 0.0453)
P-Value for Test of Close Fit (RMSEA < 0.05) = 0.884

Expected Cross-Validation Index (ECVI) = 0.611
90 Percent Confidence Interval for ECVI = (0.466 ; 0.575)
ECVI for Saturated Model = 0.331
ECVI for Independence Model = 28.071

Chi-Square for Independence Model with 210 Degrees of Freedom = 39173.131
Independence AIC = 39215.131
Model AIC = 853.865
Saturated AIC = 462.000
Independence CAIC = 39346.230
Model CAIC = 1153.520
Saturated CAIC = 1904.086

Normed Fit Index (NFI) = 0.984
Non-Normed Fit Index (NNFI) = 0.987
Parsimony Normed Fit Index (PNFI) = 0.857
Comparative Fit Index (CFI) = 0.989
Incremental Fit Index (IFI) = 0.989
Relative Fit Index (RFI) = 0.982

Critical N (CN) = 514.259

Root Mean Square Residual (RMR) = 0.0267
Standardized RMR = 0.0332
Goodness of Fit Index (GFI) = 0.951
Adjusted Goodness of Fit Index (AGFI) = 0.938
Parsimony Goodness of Fit Index (PGFI) = 0.753

The results employ one of the following principles. A measure of the discrepancy between:

i) Population covariance matrix, Σ and Model based covariance matrix, $\Sigma(\Theta)$ (Discrepancy of Approximation);

ii) Population covariance matrix, Σ and Implied (sample fitted) covariance matrix, $\hat{\Sigma}$ (Overall Discrepancy);

iii) Sample covariance matrix, S and Implied (sample fitted) covariance matrix, $\hat{\Sigma}$ (Sample Discrepancy);

iv) Model based covariance matrix, $\Sigma(\Theta)$ and Implied (sample) covariance matrix, $\hat{\Sigma}$ (Discrepancy of Estimation).

The first statistic provided by LISREL was the minimum fit Chi-Square statistic (χ^2). Recall that lower values of χ^2 represent better fit, however, since there is no theoretical upper limit to the statistic it is difficult to make an absolute judgment. Model 4 produced a χ^2 value of 704.31 with 183 degrees of freedom and highly significant with $p < 0.01$. The 'significant outcome' indicated that the null hypothesis, i.e. that the data fitted the model perfectly, $\Sigma = \Sigma(\Theta)$, should be rejected meaning that the model failed the fit test. The χ^2 statistic, however, is almost impossible to satisfy, and in practice it is 'unrealistic' (Byrne, 1998, p.110) for models to be expected to meet this stringent requirement. It is important to recall that the χ^2 statistic is sensitive to departures from normality, particularly kurtosis and also extremely sensitive to sample size. As reported by Long (1983), the Chi-Square statistic is likely to reject almost any model with positive degrees of freedom, including those which are only 'minimally false' since the chance of rejection increases with sample size. Since the model was unlikely to fit 'perfectly' and the null hypothesis bound to be rejected it is therefore better to judge whether the χ^2 value per degrees of freedom is lower than other models by which the target model is being compared. With this in mind the χ^2 statistic will be retuned to later in the chapter when other models are available for comparison.

Two variations on the χ^2 statistic were reported next. The first was the Satorra-Bentler Scaled Chi-Square, which is an attempt to allow for problems with kurtosis. The output showed that Model 4 had an $_{(s\text{-}c)}\chi^2$ value of 627.17, with a significant test outcome, which also rejected the null hypothesis and indicated a poor model fit. The second variation was the noncentrality parameter (NCP) and as the name suggests was based on the noncentral χ^2 distribution. This distribution was calculated by (χ^2 – degrees of freedom). Model 4 did not fair much better with the NCP test with a high value of (627.17 – 183) = 444.17. The 90% confidence interval was also reported and was unacceptably large at 371.79 to 524.14. The large confidence interval relative to the χ^2 statistic indicated that the model was a non-perfect fit to the data,

that the null hypothesis has been rejected and that the model judged as insufficient. Again this was not an unexpected outcome given the severity of the test.

As most tested models suffer this rejection fate at the hands of the χ^2 statistic, Jöreskog and Sörbom (1993) and also later MacCallum et al. (1996) suggested that the χ^2 should be viewed on a continuum rather than as a test to pass. They suggested that χ^2 should be judged as large or small relative to the number of degrees of freedom, and that a smaller χ^2 relative to the degrees of freedom would indicate a 'better' fit. For Model 4, the ration χ^2 / degrees of freedom = 3.4. The literature suggests that acceptable ratios are somewhere in the region of 2.0 (Carmines and McIver, 1981) to 5.0 (Wheaton et al., 1977). However, Wheaton (1987) has since cautioned that this ratio practice is no longer representative of best practice in model testing and recommended that its use be discontinued. Even given the nature of Wheaton's warning, the χ^2 /df test for Model 4 provides a more positive test outcome which is a welcome sign at this stage.

The LISREL fit test outputs moved into the various descriptive measures of fit. Recall from the last chapter that the Root Mean Square Error of Approximation (RMSEA), although only first proposed in 1980 by Steiner and Lind (1980), is now 'recognized as one of the most informative criteria in covariance structure modelling' (Byrne, 1988, p. 112). The RMSEA sets out to measure just how well the model would fit the population covariance matrix if optimal parameter values were chosen, although still essentially based on the assumption that $\sum = \sum(\Theta)$. However, it is expressed per degree of freedom, which means that sensitivity to model complexity is naturally accounted for. With respect to Model 4 the LSREL output reported an RMSEA = 0.042 and a 90% confidence interval of 0.038 to 0.045. This is a very positive result for Model 4 and as Browne and Cudeck (1993) and MacCallum et al. (1996) state, that values less than 0.05 represent good model fit, whilst values from 0.05 to 0.08 represent a mediocre model fit and values from 0.08 and 0.10 represent a poor model fit. The fact that the upper range of the 90% confidence interval was below 0.05 was a particularly satisfactory result and showed that over all possible randomly sampled RMSEA values, 90% of them would possess a value smaller than 0.045. Also, the fact that the 90% spread covered such a narrow range (0.038 to 0.045) provided an indication that there was a good degree of precision. The *p*-value for the test of close fit was 0.88; this is a test of the hypothesis that the RMSEA has an associated probability of 0.05. This is one significance test that it is good to fail!!

Jöreskog and Sörbom (1996) have suggested that a good model should display a *p*-value of greater than 0.50. A *p*-value of 0.88 from Model 4 therefore represents a particularly strong RMSEA outcome.

The next reported statistic was the Expected Cross Validation Index, (ECVI). Recall that the ECVI provides a measure of the *overall model discrepancy* across all possible calibration samples. This means it is a measure of both the model error and the sampling error and provides a useful indicator of overall model fit. The ECVI statistic has no absolute range, however, the smaller the ECVI value the better is the fit, since the smaller the overall discrepancy. A good fitting model ought also to have an ECVI value for the hypothesized model smaller than both the ECVI values for the 'independent' model and the 'saturated' model. The output for Model 4 shows an ECVI value of 0.61 (with a 90% confidence interval of 0.46 to 0.58) with an ECVI for the independent model of 28.07 and for the saturated model of 0.31. The ECVI value is low and close to the value for the saturated model, although a better fitting model would have seen its value lower than both the comparative models. The ECVI value can be used as a comparative statistic by which to compare the fit against other models which will be undertaken shortly.

The Akaike's Information Criterion (AIC) is a similar measure to ECVI except that it takes account of model parsimony. The Consistent version (CAIC) is adjusted for sample size effects. For Model 4 hypothesized model AIC is 853.87 compared to Independent AIC of 39215.13 and Saturated AIC of 462.00. Additionally, the hypothesized model CAIC is 1153.52 compared with Independence CAIC of 39346.23 and Saturated CAIC of 1904.09. As is evident from the data, the adjusted sample size CAIC provides a better model fit result, as the hypothesized model has a CAIC value of less than both comparison models, although overall the fit result is inconsistent.

The next collection of output statistics in Table 4.17 are incremental or comparative indices of fit (excluding PNFI). Their values are derived from a comparison between the hypothesized model and the independence model. The Normed Fit Index (NFI) was, for a good while, the practical criterion of choice for many researchers (see Bentler, 1992) although the Comparative Fit Index (CFI) is preferred today as it is more robust across different sample sizes. The results of fit for Model 4 are all extremely positive with values of NFI = 0.99, Non-NFI = 0.99, Parsimony-NFI = 0.86,

CFI = 0.99, Incremental Fit Index = 0.99 and Relative Fit Index = 0.98. According to Bentler (1992) a value > 0.90 indicates an acceptable fit to the data, so these markedly higher figures give confidence that the hypothesized model represents the data better than moderately well.

The Critical N (CN) value for Model 4 is quoted in the LISREL output as CN = 514.26. The CN value is the estimated minimum sample size required by the model in order to produce an acceptable statistic. A value of CN > 200 is quoted by many (see for example Byrne, 1989) as indicative of a model that adequately represents the sample data. The CN for Model 4 is well in excess of that, and is again a positive result.

The Root Mean Square Residual (RMR) represents the average residual value derived from the discrepancies between the sample covariance matrix, (S) and implied (fitted sample) covariance matrix, ($\hat{\Sigma}$) i.e. average Sample Discrepancy. For Model 4 the RMR value = 0.027. This value is difficult to interpret because it is contingent on the original units of the covariances. However, a meaningful result can be obtained by inspection of the standardized RMR. A well fitting model would normally have a value of RMR < 0.05. Model 4 has a standardized RMR = 0.033, which means that the model explains the correlations to within an average error of 0.033.

The final set of outputs was calculated from the Discrepancy of Estimation, i.e. the how well did the model based covariance matrix (in particular the parameter estimates) estimate the sample covariances. The GFI gives the amount of covariance accounted for by the model. For Model 4 the GFI = 0.95. The AGFI is adjusted for the degrees of freedom and for Model 4 AGFI = 0.94 and the parsimony GFI = 0.75. Parsimony GFI indices typically have lower values, and Mulaik et al. (1989) recommended that for an acceptable fit GFI values should be greater than 0.90 and PGFI should be greater than 0.5. Thus the Model 4 performs well against there criteria.

In summary Model 4, as measured by multiple statistical testing techniques, was evaluated to possess the following characteristics of fit to the population and sample data.

i) The Satorra-Bentler Scaled Chi-Square value was overly high and statistically significant at 627.170 (p = 0.00) showing a high error of appropriation and a poor predicted fit between population covariance and model based covariance. *A negative indication of model fit.*

ii) The Root Mean Square of Approximation (RMSEA) value of 0.042 (with 90% confidence interval of 0.038 – 0.045) was well beneath the critically accepted value of 0.05 indicating an estimated low error of fit between population covariance and model based covariance per degree of freedom. *A positive indication of model fit.*

iii) The Expected Cross-Validation Index (ECVI) of 0.61 (90% confidence interval of 0.47 – 0.58) although a low value, and certainly lower than the Independence Model of 28.07 was larger than the value of the Saturated Model of 0.33. The CAIC model value (1153.52) by contrast was lower than both its reference values (39346.23 and 1904.09). *A mixed indication of model fit.*

iv) The Comparative Fit Index (CFI) was an extremely high 0.99 and comfortably in excess of the 0.90 benchmark. *A positive indication of model of model fit.*

v) The Critical N value was a high 514.26, well in excess of the 200 threshold. *A positive indication of model fit.*

vi) The Standardized Root Mean Residual (RMR) was 0.033, well beneath the guide level of 0.05. *A positive indication of model fit.*

vii) The Goodness of Fit index was 0.95 and well above the target level of 0.90. *A positive indication of model fit.*

The indices above, chosen as representative of the various measures, show that the model performed well on Sample Discrepancy and Discrepancy of Approximation. It did not perform well on the Discrepancy of Approximation or Overall Discrepancy. A key component of performance seems to be the involvement of the hypothesized

population covariance matrix, Σ. All indices that contained Σ in the discrepancy calculation, with the exception of RMSEA which was expressed per degree of freedom, gave poor values of fit. Indices which contained only the model based covariance matrix, $\Sigma(\Theta)$, or the sample covariance matrix, S, or the implied covariance matrix, $\hat{\Sigma}$, performed considerably better.

Having examined the degree to which Model 4 fitted the data the next procedure was to examine the residuals in order to identify any areas of misfit. The residuals represent discrepancies between the sample covariance matrix, S and the implied covariance matrix. LISREL produced two residual outputs, *fitted residuals* and *standardized residuals*. The standardized residuals represent the number of standard deviations the observed residuals are away from a perfect model fit, i.e. how far away from the residuals = zero. Two hundred residuals were calculated for Model 4 and the values of all 200 are represented in a stem-leaf diagram in Table 4.15a. Good models would have a symmetric stem-leaf diagram with most values being clustered around the zero point and few points at the end of the tails. Residuals with values > 2.58 (see Byrne, 1998) are considered large, reflecting badly on the model.

For Model 4 the stem-leaf diagram had a good central cluster which was roughly symmetric with a slight tendency to skew to the positive-signed residuals. This skew indicated that the model had slightly underestimated the covariance between the variables which had resulted in underfitting. This pointed to the fact that the model might benefit from being modified with the inclusion of additional pathways (i.e. freeing some parameters). Of the 200 residuals 20 had residual values smaller than -2.58 and 18 had residual values greater than +2.58. The fitted residual stem-leaf plot can be found in Table 4.15a and the standardized stem-leaf plot, along with the residuals exceeding modulus 2.58, found in Table 4.15b.

Table 4.15a Summary Statistics for Fitted Residuals

```
Smallest Fitted Residual =   -0.089
  Median Fitted Residual =    0.000
 Largest Fitted Residual =    0.097


Stemleaf Plot

- 8|95
- 7|1
- 6|5410
- 5|97
- 4|9975541100
- 3|6000
- 2|9988888876666443321
- 1|999987776655544443333222221111110000
- 0|99999988776444333222111000000000000000000000000
  0|111122223344555666777799
  1|000001111222333334445566677777888999
  2|0012223444445568899
  3|00001134467899
  4|001155
  5|02345
  6|16
  7|
  8|
  9|7
```

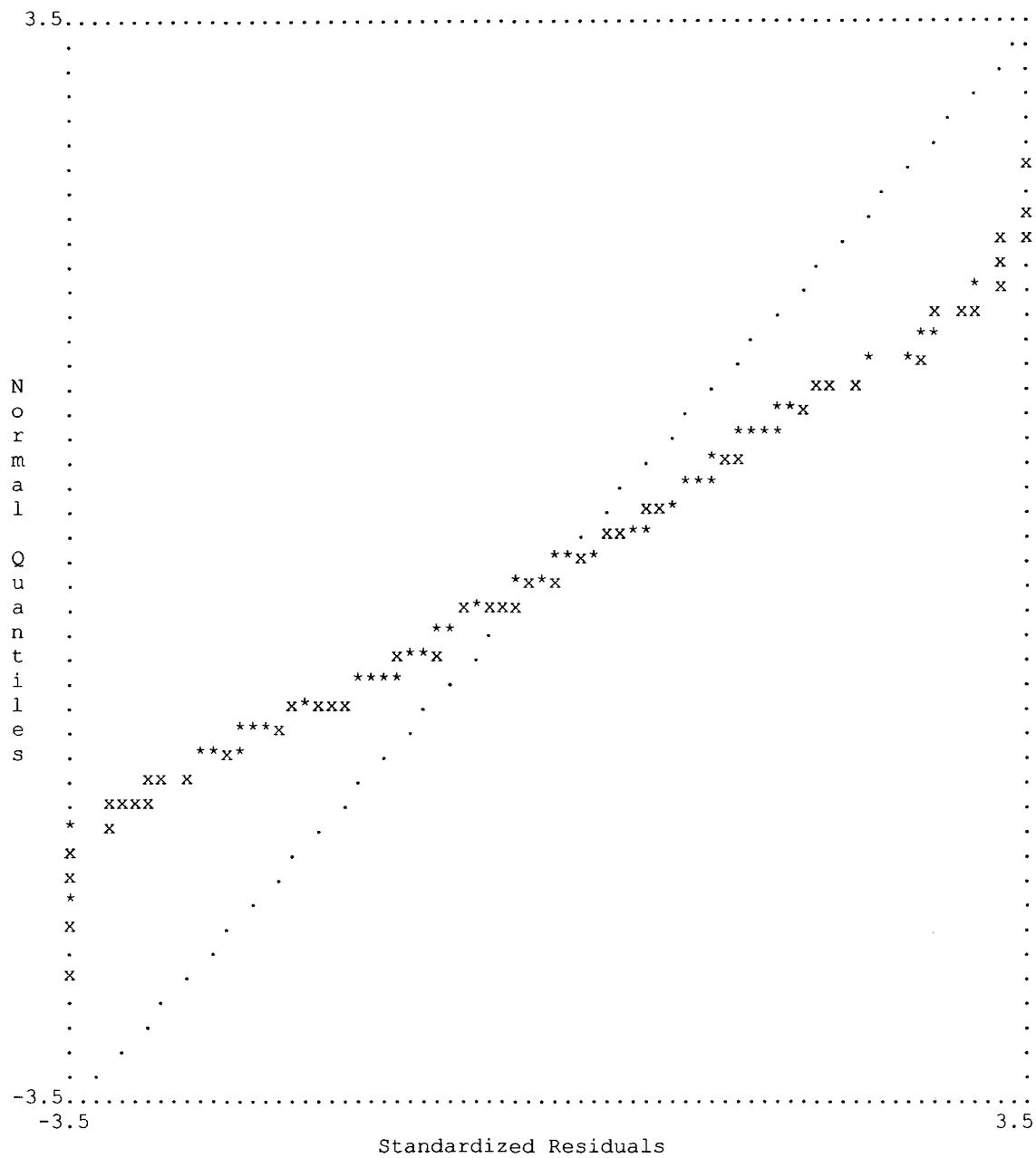
Table 4.15b Summary Statistics for Standardized Residuals

```
Smallest Standardized Residual =   -5.011
  Median Standardized Residual =    0.000
 Largest Standardized Residual =    5.509


Stemleaf Plot

- 5|0
- 4|95510
- 3|87752220
- 2|998666655533332210000
- 1|9988877666544332211111000
- 0|99999999888777777777666555444433222111000000000000000000000000000000
  0|11112222223344444455566777778888899999
  1|000011111122222233333444555666677778889999
  2|0011133477788888
  3|11113339
  4|4
  5|5


Largest Negative Standardized Residuals

Residual for     ENQ6 and     ENQ5  -2.997
Residual for    PHYS1 and    BIOL3  -4.906
Residual for    PHYS2 and    BIOL1  -3.799
Residual for    PHYS3 and    BIOL1  -2.606
Residual for    PHYS4 and    BIOL1  -3.151
```

```
Residual for    CHEM1 and    BIOL3  -3.680
Residual for    CHEM2 and    CHEM1  -3.681
Residual for    CHEM3 and    BIOL1  -3.544
Residual for    CHEM4 and     ENQ6  -2.820
Residual for    CHEM4 and    BIOL2  -4.543
Residual for    CHEM4 and    BIOL3  -4.031
Residual for     NOS2 and     ENQ5  -3.183
Residual for     NOS2 and    BIOL3  -2.944
Residual for     NOS2 and    PHYS2  -4.140
Residual for     NOS3 and     NOS1  -2.622
Residual for     SCI2 and     ENQ4  -2.644
Residual for     SCI2 and    BIOL3  -5.011
Residual for     SCI2 and    BIOL4  -2.882
Residual for     SCI2 and    PHYS2  -3.189
Residual for     SCI2 and    PHYS3  -4.544


Largest Positive Standardized Residuals
Residual for    BIOL2 and     ENQ4   2.774
Residual for    PHYS3 and     ENQ5   3.266
Residual for    PHYS3 and    PHYS2   3.084
Residual for    PHYS4 and     ENQ5   2.801
Residual for    PHYS4 and    BIOL3   3.087
Residual for    CHEM2 and    BIOL1   2.789
Residual for    CHEM2 and    BIOL2   3.307
Residual for    CHEM2 and    BIOL4   5.509
Residual for    CHEM3 and    PHYS4   4.423
Residual for    CHEM4 and    CHEM1   3.910
Residual for     NOS1 and     ENQ6   2.810
Residual for     NOS1 and    PHYS3   3.057
Residual for     NOS1 and    PHYS4   2.651
Residual for     NOS2 and    BIOL4   2.677
Residual for     NOS2 and    CHEM2   2.777
Residual for     SCI2 and    BIOL1   2.726
Residual for     SCI2 and    PHYS1   3.319
Residual for     SCI3 and    BIOL1   3.149
```

A normal probability (or Q-plot) of residuals was generated next. This provided a graphical display of the standardized residuals (on the abscissa) against the quartiles of the normal distribution (on the ordinate). Each x in the plot signifies a single plot and each * signifies multiple plots. A vertical distribution of points represents the best possible fit and a horizontal distribution of points represents the worst possible fit. A distribution of points along the 45° line (or steeper) indicates an acceptable fit. A non-linear Q-plot indicates a departure from normality. As can be seen from Chart 4.3, the plot is linear, confirming there were no departures from normality, although the Q-plot is less steep than the acceptable 45° indicating a less than acceptable fit and the possibility that certain parameters were mis-specified. There appear not to be any outliers in the Q-plot which would have been a further indication of possible specification errors in the model.

Chart 4.3 Q-plot of Standardized Residuals



The linear plot indicates that there is no deviation from normality. The presence of outliers would have indicated possible specification error. A gradient shallower than 45° indicates a less than satisfactory fit.

Validity and Reliability of the Model

An evaluation of the validity of the model might ask the question 'was the model measuring what it was intended to measure?' By contrast a reliability measure might ask the question 'to what extend were the indicator variables free from random error?' From these perspectives the model's validity and reliability can be determined by analyzing the model averaged values of the parameter loadings and their related measurements errors.

Recall that for each model parameter the equation linking the indicator variable and the latent variable is expressed generically as:

$$X = \lambda\xi + \delta \qquad \text{(equation 4.1)}$$

Where, X = a measure of the indicator variable
ξ = latent variable
λ = parameter loading
δ = measurement error

Earlier in the chapter each of the parameter estimates for each of the equations were examined individually to confirm that they were both (i) substantial and (ii) significant. These results were displayed in the measurement equations in Table 4.13a. The outcome of that evaluation was very positive with all the factor loadings reported as being appropriately large in size and all were significant at the level of $p = 0.05$ or better. This gave support to the fact that the indicator variables were validly measuring the underlying construct as represented by the latent variable. The parameter loadings reported in Table 4.13a were in their unstandardized form. This is quite appropriate for the role which it was undertaking at the time, although the unstandardized format has the limitation that direct comparisons with other measures of the same construct may not be easily made if the measures have been taken on different scales. To resolve this issue and to provide other benefits (see below) it is necessary to consider the parameter loadings in their standardized form. These were presented as LAMBDA-X and LAMBA-Y in the LISREL output and are reported in Table 4.16. As can be seen from the table of values the parameter loadings occupy values that are reassuringly high from a validity perspective, (all but one value in excess of 0.6). Table 4.16 reveals that Item pair SCI1 held the highest value

and therefore indicates strongest claims to validity with a standardized parameter loading of 0.870 whilst CHEM2 makes the least claim to validity with a parameter loading of 0.570. None of the parameter loadings call into question any issues concerned with poor validity.

Whilst the validity seems secure at the level of the individual item pairs, it is also possible to test the validity of each constructed latent variable and hence, through this, the validity of the whole model. Equation 4.2 provides a means of calculating the validity of the latent constructs where, ρv, is 'the amount of variance that is captured by the construct in relation to the amount of variance due to the measurement error.' (Fornell and Larcker, 1981, p.45).

$$\rho_v = (\Sigma\lambda^2) / [\Sigma\lambda^2 + \Sigma(\theta)] \qquad \text{(equation 4.2)}$$

where ρ_v = convergent validity

λ = indicator loadings

θ = indicator error variances (variances of the δ's and ε's)

Σ = summation over the indicators of the latent variables

A positive result would be for ρ_v to be greater than 0.5, as this would indicate that the underlying latent variables account for a greater proportion of the variance than does the measurement error. If more than 50% of the variance is accounted for by measurement error then this might call into question the validity of the model. There is no automated procedure within either LISREL or SPSS to calculate these values so they had to be calculated by long hand. The standardized parameter loadings and the error variances (labeled THETA) used in the calculation are found in Table 4.16 and the final calculations can be found in Table 4.17.

Table 4.16 Completely Standardized Solution of Parameter Loadings and Error Variances for Model 4

	LAMBDA - Y					THETA
	Enquiry	Biology	Physics	Chemistry	NoS	EPS
ENQ4	0.638					0.593
ENQ5	0.601					0.639
ENQ6	0.707					0.500
BIOL1		0.642				0.588
BIOL2		0.752				0.435
BIOL3		0.605				0.634
BIOL4		0.636				0.595
PHYS1			0.667			0.555
PHYS2			0.599			0.641
PHYS3			0.709			0.497
PHYS4			0.631			0.602
CHEM1				0.724		0.476
CHEM2				0.570		0.675
CHEM3				0.608		0.630
CHEM4				0.670		0.551
NOS1					0.682	0.535
NOS2					0.685	0.531
NOS3					0.766	0.414

	LAMBDA - X		THETA
	GenSci	GlobalSc	DELTA
SCI1		0.870	0.243
SCI2		0.659	0.566
SCI3		0.623	0.612

In a consideration of the issue of model reliability, the squared multiple correlations (R^2) of the indicator variables can be examined and evaluated. Again, this was first discussed in the consideration of *parameter estimation* and the R^2 values were presented in Table 4.16. These values show the proportion of the variance explained by its underlying latent variable with the rest of the variance attributable to measurement error. Recall that the R^2 measures were appropriately high signaling the high reliabilities of the individual measurement items. SCI1, in keeping with the validity measures above, had the highest measures reliability with a value of R^2 = 0.76, whilst CHEM2 had the lowest reliability score with a measurement of R^2 = 0.33. None of these values brought into question any issues concerned with reliability.

It is possible to go one step further than the evaluation of the individual parameter loadings as an assessment of reliability. It is also possible to calculate *composite reliability*.

The composite reliabilities, which again are not calculated automatically by LISREL can be worked out using the equation:

$$\rho_c = (\Sigma\lambda)^2 / [(\Sigma\lambda)^2 + \Sigma(\theta)] \qquad \text{(equation 4.3)}$$

where ρ_c = composite reliability

λ = indicator loadings

θ = indicator error variances (variances of the δ's and ε's)

Σ = summation over the indicators of the latent variables

A value for $\rho_c > 0.6$ is seen to be a desirable outcome (Bagozzi and Yi, 1988).

The results for both the composite reliabilities and the mean variance extracted are presented in Table 4.17

Table 4.17 Composite Reliabilities and Mean Variance Extracted

	Composite Reliabilities	Mean variance Extracted
ENQUIRY	0.686	0.423
BIOLOGY	0.755	0.437
PHYSICS	0.747	0.426
CHEMISTRY	0.739	0.417
NoS	0.755	0.507
GLOBALSC	0.765	0.526

As can be seen from Table 4.17, the composite reliabilities are all greater than the threshold value of 0.60, thus giving a strong indication of a reliable

measurement. The mean extracted variance was more borderline with four of the values dipping slightly under the benchmark figure of 0.50. With this in mind it was reported by Fornell and Larker (1981) that:

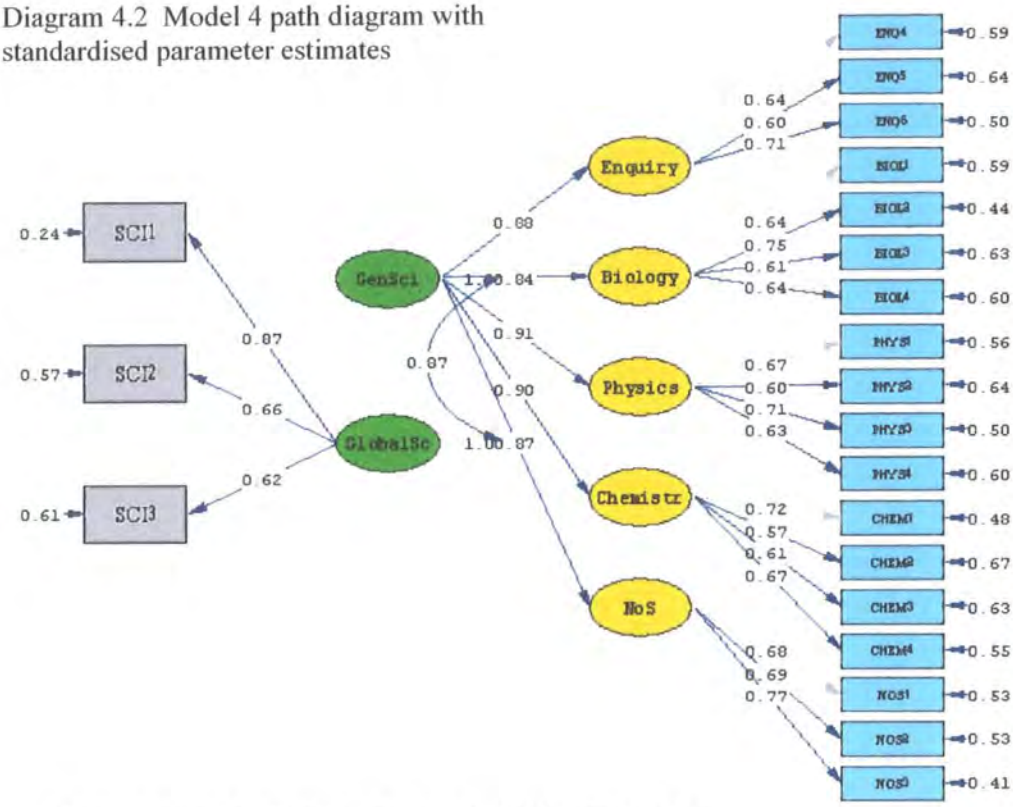
> (ρ_v) is a more conservative measure than (ρ_c). On the basis of (ρ_c) alone, the researcher may conclude that the convergent validity of the construct is adequate, even though more than 50% of the variance is due to error (p. 46).

On balance there seems to be sufficient evidence to suggest that the reliability and validity of the model is not called into question.

Table 4.18 Model fit indices for Models 1, 2, 3 and 4

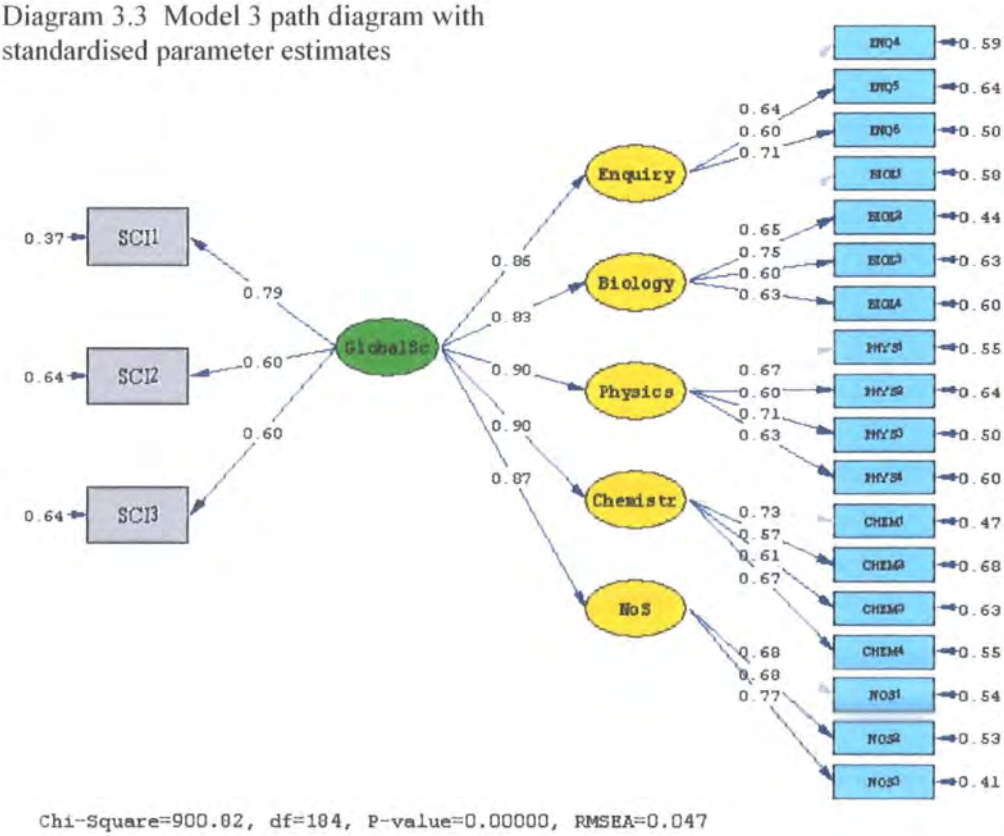
Fit Index Information	Model 4	Model 3	Model 2	Model 1
Degrees of Freedom	183	184	130	125
Minimum Fit Function Chi-Square	704.32 p=0.00	832.46 p=0.00	553.96 p=0.00	538.40 p=0.00
Satorra-Bentler Scaled Chi-Square	627.17 p=0.00	743.79 p=0.00	481.01 p=0.00	465.88 p=0.00
Chi-Square Corrected for Non-Normality	590.71 p=0.00	674.26 p=0.00	450.91 p=0.00	435.59 p=0.00
Estimated Non-centrality Parameter (NCP)	444.17	559.97	351.01	340.88
90 Percent Confidence Interval for NCP	371.79 524.14	479.71 647.79	287.66 421.94	278.55 410.78
Root Mean Square Error of Approx (RMSEA)	0.0417	0.0467	0.0440	0.0442
90 Percent Confidence Interval for RMSEA	0.0381 0.0453	0.0432 0.0502	0.0398 0.0479	0.0339 0.0485
P-Value for Test of Close Fit (RMSEA < 0.05)	0.884	0.878	0.454	0.440
Expected Cross-Validation Index (ECVI)	0.611	0.712	0.479	0.472
90 Percent Confidence Interval for ECVI	0.466 0.575	0.542 0.663	0.358 0.454	0.355 0.449
ECVI for Saturated Model	0.331	0.331	0.245	0.245
ECVI for Independence Model	28.071	28.071	19.839	19.839
Normed Fit Index (NFI)	0.984	0.981	0.983	0.983
Non-Normed Fit Index (NNFI) (Tucker Lewis Index, TLI)	0.987	0.984	0.985	0.985
Parsimony Normed Fit Index (PNFI)	0.857	0.860	0.835	0.803
Comparative Fit Index (CFI)	0.989	0.986	0.987	0.988
Incremental Fit Index (IFI)	0.989	0.986	0.987	0.988
Relative Fit Index (RFI)	0.982	0.978	0.980	0.979
Critical N (CN)	514.26	435.78	495.96	494.86
Root Mean Square Residual (RMR)	0.0267	0.0288	0.0265	0.0260
Standardized RMR	0.0332	0.0358	0.0339	0.0332
Goodness of Fit Index (GFI)	0.951	0.942	0.955	0.957
Adjusted Goodness of Fit Index (AGFI)	0.938	0.927	0.941	0.941
Parsimony Goodness of Fit Index (PGFI)	0.753	0.750	0.726	0.699

Diagram 4.2 Model 4 path diagram with standardised parameter estimates



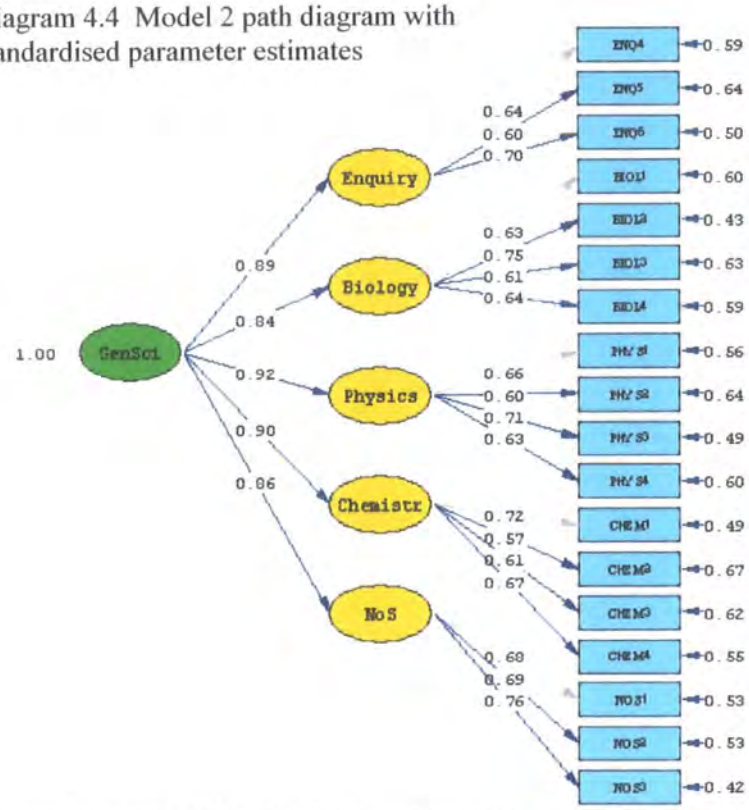
Chi-Square=757.87, df=183, P-value=0.00000, RMSEA=0.042

Diagram 3.3 Model 3 path diagram with standardised parameter estimates



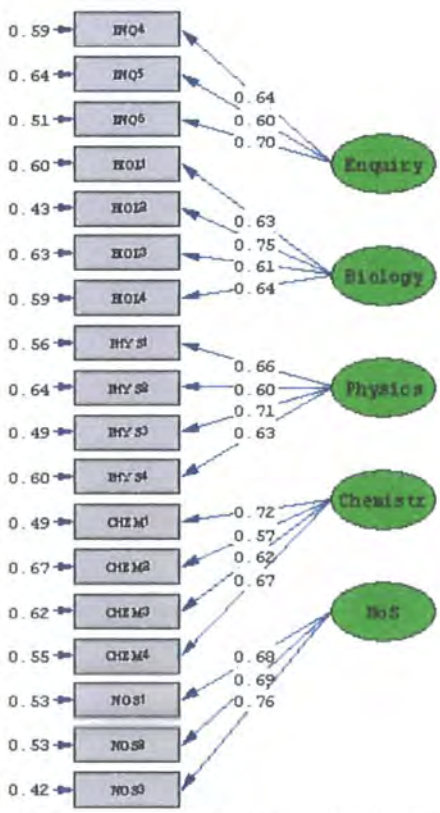
Chi-Square=900.02, df=184, P-value=0.00000, RMSEA=0.047

Diagram 4.4 Model 2 path diagram with standardised parameter estimates



Chi-Square=587.54, df=130, P-value=0.00000, RMSEA=0.044

Diagram 4.5 Model 1 path diagram with standardised parameter estimates



Chi-Square=567.35, df=125, P-value=0.00000, RMSEA=0.044

Comparison of Model Fits

Table 4.18 and Diagrams 4.2 to 4.5 show information about all four models under investigation. A comparison between the Model 1 and Model 2 fit statistics reveals very little difference between the two models in any of the presented fit parameters. Any measured difference only appeared in the third significant figure, and as such probably indicates negligible fit differences. Both Model 1 and Model 2 (as did Models 3 and 4) demonstrate the possible multidimensionality of the science self-concept with the formation of five distinct factors all with substantial factor loadings. In keeping with the exploratory factor analysis, the confirmatory factor analysis showed that the factor loadings were both large and significant and ranging from 0.57 to 0.77. The correlational data between factors was also high (between 0.73 to 0.82, see Table 4.13d) indicating that although the pupils could discriminate between the factors there was a strong possibility of a derived general hierarchical term. In the absence of other indicators Model 2 was likely to be a stronger candidate for a solution than Model 1 because it did include an inferred higher order term, thus providing fuller information about the model whilst its fit statistics were in every way equal to the alternative model.

An inspection of the data in Table 4.18 indicates that Model 4 has a better fit to the data than Model 3 on every measure. Model 4 has a lower Chi-square statistic ($_{SB}\chi^2$ = 627 and 743) and a lower Non-centrality Parameter (NCP = 444 and 480). It has a lower RMSEA (0.0417 to 0.0467) and the 90% confidence interval for Model 4 remained beneath the important 0.05 threshold for the entire range (0.0381 - 0.0453). The Expected Cross- Parameter was lower (0.611 and 0.712), although neither was beneath the Saturated Model value of 0.331. The fit indices of both models were extremely high, although Model 4 consistently out performed Model 3 on all the fit indices including the Tucker Lewis Index (TLI = 0.987 and 0.984), the Standardized Root Mean square Residual (SRMR = 0.0332 and 0.0358), and the Goodness of Fit Index (0.951 and 0.942). Model 4 also achieved a higher Critical N with a value of 514 compared with 435. From these data it appears safe to suggest that Model 4 has a better fit to the data than Model 3 and as such Model 3 will be rejected in preference to Model 4. The increased level of fit was consistent enough across all indicators to outweigh the increase in parsimony of Model 4.

Model 4 and Model 2 presented two equally feasible solutions and these two models will be considered first. They both had good fit to the data and both had strong evidence from the Tucker Lewis Index, the Standardized Root Mean square Residual, the Goodness of Fit Index and the Critical N value. The benefit of Model 4 was that it showed the high correlation between the derived general science factor (GenSci) and the independently measured global science factor (GlobalSc). This high correlation was probably indicating that the two latent variables were measuring the same construct. If they were indeed the same construct then Model 2 would be preferred over Model 4 on the basis of parsimony, since parsimony requires the less complex model to be chosen if all other things are equal.

In order to help reach a decision on the preferred model, additional insight into the performance of the two models was sought. Up to this stage, the model testing and analysis procedures had treated the data sample as one large homogeneous group of 1398 individuals. In the process of model selection the preferred model had to be recognized as being able to be robustly and consistently employed across a wide range of differently constituted data sets. The success of this procedure was evaluated in a number of ways with each procedure having its own properties and purposes. The first technique was to check for model fit with key independent subgroups, i.e. to perform a loose replication strategy test. This constituted each of the models being fitted to the different subgroup samples (and in this case gender subgroups) and the parameter estimates and fit statistics for the two models being evaluated and compared. The technique would also serve as a cross-validation procedure, however the loose replication strategy can not be regarded as 'true cross-validation since the in the analysis of the validation sample in no way depends on results from analysis of the calibration sample' (MacCallum et al., 1994, p.13). This procedure served two purposes; firstly, it allowed the models to demonstrate the extent of their validity and utility when being applied specifically to important sub-group data; additionally, it provided multiple ranges of fit indices (rather than just one set) with which to evaluate each of the models.

Tables 4.19a and 4.19b show the fit statistics for Models 4 and 2 from various combinations of sub-group Key Stage and Gender. It is important to recognize the significance of these data sets at this stage. Table 4.19a contains the simplest and largest subgroups of the data. The data were roughly divided into two halves using one of two criteria, firstly, selecting by Key Stage and then secondly, selecting by

gender. Later in the research those groupings were subdivided again into boys and girls within KS3 and boys and girls within KS4 and these latter results are presented in Table 3.19b. The first point to notice is that the model-to-data fit is, in most cases, slightly better for the sub-groups than for the entire sample. In fact, the $_{SB}$Chi-square for Key Stage 4 Girls is almost non-significant, which is a very positive result indeed, as this is the most stringent test of them all. This is a very reassuring outcome as it shows that the instrument is applicable to age differentiated and gender differentiated subgroups, and the modelling process holds firm with the differently constituted data sets.

The fact the fit statistics are as good as, and sometimes better for, the subgroups than for the whole group might be explained in two ways. First, that the sample size, by definition, is reduced when selecting a sub-group, and as has been well rehearsed in the sections above, the Chi-square statistic along with many of the other fit statistics is sensitive to sample size. Second, the individuals who constitute a subgroup are more likely to be more homogeneous in the responses than a more widely constituted group, which could result in smaller variances across the measured data. This means that the results indicating better fit need to be tempered with this realism. However, and this a very significant 'however', the model fit to the data when sampled by subgroup is certainly no worse fitting than when the data was applied in its entirety. This is consistently so across most of the fit indices for most of the permutations of sub-group choice. This is a key outcome and shows that both Model 4 and Model 2 behave consistently and robustly across different segments of the data. Two inferences can be drawn from this; that the models have satisfied what could be termed 'a loose replication strategy' test (as the models are being fitted to two independent data sets) indicating that the models may be generalisable to a wider population. Also, that the instruments are appropriate to use with all age groups within secondary education and both sexes. This second point will be explored at a later stage below.

Table 4.19a Fit Statistics for Models 4 and 2 with Subgroups Key Stage and Gender

Fit Index Information	Boys n = 711		Girls n = 687		Key Stage 4 n = 507		Key Stage 3 n = 891	
	Model 4	Model 2	Model 4	Model 2	Model 4	Model 2	Model 4	Model 2
Degrees of Freedom	183	130	183	130	183	130	183	130
Minimum Fit Function Chi-Square	419.53 p=0.00	307.77 p=0.00	538.44 p=0.00	411.94 P=0.00	531.50 p = 0.00	401.42 p = 0.00	552.31 p=0.00	443.17 p=0.00
Satorra-Bentler Scaled Chi-Square	361.24 p=0.00	259.30 p=0.00	474.08 p=0.00	360.04 P=0.00	480.94 p = 0.00	349.43 p = 0.00	494.56 p=0.00	395.30 p=0.00
Root Mean Square Error of Approx (RMSEA)	0.0370	0.0374	0.0482	0.0508	0.0567	0.0578	0.0437	0.0479
90 Percent Confidence Interval for RMSEA	0.0314 0.0426	0.0308 0.0441	0.0429 0.0535	0.0446 0.0571	0.0506 0.0629	0.0505 0.0651	0.0391 0.0484	0.0425 0.0533
P-Value for Test of Close Fit (RMSEA < 0.05)	0.944	0.880	0.0685	0.0160	0.000	0.000	0.354	0.037
Expected Cross-Validation Index (ECVI)	0.761	0.567	0.953	0.745	1.296	0.974	0.789	0.641
90 Percent Confidence Interval for ECVI	0.573 0.726	0.422 0.551	0.743 0.930	0.568 0.732	1.019 1.276	0.750 0.970	0.594 0.742	0.473 0.608
ECVI for Saturated Model	0.651	0.482	0.673	0.499	0.913	0.676	0.519	0.384
ECVI for Independence Model	27.770	19.552	27.460	19.723	26.565	18.115	29.057	20.998
Non-Normed Fit Index (NNFI) (Tucker Lewis Index, TLI)	0.989	0.989	0.982	0.980	0.974	0.971	0.986	0.983
Parsimony Normed Fit Index (PNFI)	0.855	0.834	0.849	0.827	0.840	0.817	0.855	0.832
Comparative Fit Index (CFI)	0.991	0.991	0.984	0.983	0.977	0.976	0.988	0.986
Critical N (CN)	453.89	467.64	334.42	325.71	243.43	247.78	415.66	384.70
Root Mean Square Residual (RMR)	0.0287	0.0273	0.0307	0.0308	0.0451	0.0463	0.0277	0.0282
Standardized RMR	0.0350	0.0345	0.0402	0.0413	0.0588	0.0619	0.0347	0.0359
Goodness of Fit Index (GFI)	0.944	0.952	0.928	0.935	0.905	0.917	0.939	0.943
Parsimony Goodness of Fit Index (PGFI)	0.748	0.724	0.735	0.711	0.717	0.697	0.744	0.717

Table 4.19a Fit statistics Models 2 and 4 with the data divided into two sub-samples. Comparisons can be made between the two models when they are fitted with data from one Key stage or one sex only.

Table 4.19b Fit Statistics for Models 4 and 2 with mini-subgroups Key Stage and Gender

Fit Index Information	KS4 Girls n = 235		KS4 Boys n = 272		KS3 Girls n = 452		KS3 Boys n = 439	
	Model 4	Model 2	Model 4	Model 2	Model 4	Model 2	Model 4	Model 2
Degrees of Freedom	183	130	183	130	183	130	183	130
Minimum Fit Function Chi-Square	245.794 p=0.001	185.636 p=0.001	322.303 p=0.00	218.845 p=0.000	425.327 p=0.00	325.066 p=0.00	392.22 p=0.00	296.85 p=0.00
Satorra-Bentler Scaled Chi-Square	236.709 p=0.005	173.254 p=0.007	275.518 p=0.00	183.997 p=0.001	375.932 p=0.00	290.286 p=0.00	328.25 p=0.00	246.85 p=0.00
Root Mean Square Error of Approx (RMSEA)	0.0354	0.0377	0.0432	0.0391	0.0483	0.0523	0.0426	0.0453
90 Percent Confidence Interval for RMSEA	0.0206 0.0477	0.0207 0.0518	0.0323 0.0534	0.0250 0.0516	0.0414 0.0553	0.0442 0.0604	0.0351 0.0499	0.0366 0.0539
P-Value for Test of Close Fit (RMSEA < 0.05)	0.868	0.745	0.287	0.499	0.054	0.009	0.225	0.115
Expected Cross-Validation Index (ECVI)	1.516	1.162	1.546	1.104	1.202	0.957	1.153	0.889
90 Percent Confidence Interval for ECVI	1.270 1.609	0.962 1.255	1.221 1.550	0.864 1.129	0.932 1.178	0.724 0.944	0.862 1.093	0.658 0.861
ECVI for Saturated Model	1.974	1.462	1.705	1.262	1.024	0.758	1.055	0.781
ECVI for Independence Model	26.661	18.197	26.312	18.379	28.821	21.341	29.493	20.947
Non-Normed Fit Index (NNFI) (Tucker Lewis Index, TLI)	0.990	0.987	0.985	0.987	0.983	0.980	0.987	0.985
Parsimony Normed Fit Index (PNFI)	0.838	0.815	0.838	0.818	0.846	0.824	0.849	0.827
Comparative Fit Index (CFI)	0.991	0.989	0.987	0.989	0.985	0.983	0.989	0.987
Critical N (CN)	228.79	231.18	227.65	252.01	277.44	265.78	308.46	303.39
Root Mean Square Residual (RMR)	0.0451	0.0463	0.0439	0.0431	0.0315	0.0317	0.0337	0.0331
Standardized RMR	0.0588	0.0619	0.0572	0.0580	0.0420	0.0431	0.0406	0.0408
Goodness of Fit Index (GFI)	0.905	0.917	0.898	0.918	0.914	0.921	0.918	0.928
Parsimony Goodness of Fit Index (PGFI)	0.717	0.697	0.711	0.698	0.724	0.700	0.728	0.705

Table 4.19b Fit statistics Models 2 and 4 with the sub-samples again sub-divided. Comparisons can be made between the two models when they are fitted with data from one Key Stage and one sex only.

Cross Validation

One procedure which helped to inform the decision about model selection was to obtain the cross-validation index (CVI) for each model. In fact, a double cross-validation was undertaken by reversing the roles of the first and second sub-sample. To carry out the procedure the data sample was divided into two using sex as the dividing criterion. Initially the boys' sample was treated as the calibration sample and the girls sample was treated as the validation sample. Each of the models (Model 4 and Model 2) was fitted in turn to the boys' sample and a fitted covariance matrix for each model was obtained. Fitting functions were then formed which showed the discrepancy between the calibration sample covariance matrix and the validation sample covariance matrix. The order of the samples was then reversed such that the girls' sample was treated as the calibration sample and the boys sample as the validation sample and procedure repeated. What emerged were two CVI values for each model. According to Bagozzi and Yi (1989) the model which produces the smallest CVI in each of the rounds is the one that can be considered as having the greatest predictive validity. The results for the two models are presented below.

Table 4.20 Double Cross-validation for Models 4 and 2

Sample combination	CVI for Model 4	CVI for Model 2
Calibration boys, validation girls	3.68	2.90
Calibration girls, validation boys	3.70	3.06

Table 4.20 summarizes the results of the double cross validation procedure and each of the CVI values were obtained from a separate run from LISREL. It can be seen that the CVI values for Model 2 are clearly lower than for Model 4 in both rounds of the validation procedure. This provides clear evidence to indicate that Model 2 is most likely to have the greatest predictive validity of the two models and should therefore be chosen in preference over Model 4.

In order to confirm the decision to choose Model 2, over Model 4, the more stringent cross moderation test instigating both the moderate and tight replication strategies was carried out. This involved each model being fitted to two samples simultaneously thus enabling the testing of invariance constraints. The difference between the moderate and the tight strategies was that in carrying out the moderate replication strategy certain parameters were allowed to be freely estimated by removing a number of invariance constraints from the endogenous latent variables. This was a much less onerous condition than the tight replication strategy and placed less stress on fitting the data to the model. The results are displayed in Table 4.21.

There are three ways in which the data in Table 4.21 were evaluated. First, by assessing the fit index information in its own right; second, by assessing the fit index information against the original data for Models 2 and 4 as presented in Table 4.18; third, by comparing the fit data between the tight and the moderate replication strategies to compare the models performance under these circumstances.

(i) When considering Model 2 in absolute terms, the fit data for both tight and moderate strategies are good. All the Global goodness of fit data are more than satisfactory with RMSEA values beneath 0.05, the p-values for RMSEA above 0.05, fit indices all well above 0.90, and the Critical Ns safely above 200. For the Group goodness fit data, the RMRs are below 0.05 and the Goodness of Fit Index around 0.95. These values lend support to the notion that Model 2 performs well in the cross-validation process. When considering Model 4 in absolute terms, the fit data for both tight and moderate strategies are slightly poorer, although there is not a great deal between them.

(ii) When the original fit statistics (shown in Table 4.18 and using one sample) are compared with replication strategy fit statistics (shown in Table 4.21 and using two samples simultaneously), Models 2 and 4 also behave quite similarly. They both have values which are similar in scale order between the two types of fit

Table 4.21 Tight, moderate & extra Moderate replication strategies for Models 2 and 4

Fit Index Information	Model 2 Tight	Model 2 Moderate	Model 2 Xmod	Model 4 Tight	Model 4 Moderate	Model 4 Xmod
Global Goodness of Fit Statistics						
Degrees of Freedom	301	296	279	415	410	392
Normal theory weighted Chi-Square	788.57	811.99	788.57	1329.24	1315.50	1293.90
Satorra-Bentler Scaled Chi-Square	681.82	672.76	652.26	1133.22	1122.05	1103.87
Root Mean Square Error of Approx (RMSEA)	0.043	0.042	0.044	0.051	0.050	0.051
P-Value for Test of Close Fit (RMSEA < 0.05)	0.516	0.497	0.320	0.001	0.001	0.000
Non-Normed Fit Index (NNFI or TLI)	0.986	0.986	0.985	0.981	0.981	0.981
Parsimony Normed Fit Index (PNFI)	0.959	0.944	0.890	0.959	0.948	0.907
Comparative Fit Index (CFI)	0.986	0.986	0.986	0.981	0.981	0.981
Critical N (CN)	740	738	722	598	597	582
Group Goodness of Fit Statistics						
Group 1 Girls						
Contribution to Chi-Square	458.96	458.62	438.79	819.65	819.62	799.52
Percentage Contribution to Chi-Square	56.81	57.49	57.45	54.65	55.33	55.19
Root Mean Square Residual (RMR)	0.0401	0.039	0.0397	0.264	0.253	0.255
Standardized RMR	0.0518	0.0512	0.0516	0.245	0.238	0.241
Goodness of Fit Index (GFI)	0.929	0.929	0.932	0.904	0.904	0.907
Group 2 Boys						
Contribution to Chi-Square	348.89	339.186	324.93	680.26	661.72	649.07
Percentage Contribution to Chi-Square	43.19	42.52	42.55	45.35	44.67	44.81
Root Mean Square Residual (RMR)	0.0369	0.0354	0.0363	0.249	0.233	0.238
Standardized RMR	0.0476	0.0453	0.0461	0.230	0.218	0.220
Goodness of Fit Index (GFI)	0.947	0.949	0.950	0.922	0.924	0.925

Table 4.21 shows the results of the tight, moderate and loose replication strategies for Model 2 and Model 4. The replication was carried out by first fitting the sample from group one to the model and then fitting the model to group two. The fitting function was calculated by evaluating the discrepancy between the two covariance matrices. In the tight replication all parameters were forced to be equal across groups. In the moderate replication strategy the endogenous latent variables were allowed to be freely estimated.

(iii) In the final test a comparison was made between the performance of each model in the tight, moderate and extremely moderate replication mode to check whether there was a statistically significant difference between operating in the three modes. Due to the fact that the models were nested, that is, 'the tight replication strategy can be derived from the moderate replication strategy by introducing additional equality

constraints' (Diamantopoulos, 2000), the difference in Chi-square values, D^2, could be compared with published tables to check on statistical significance. It was therefore possible to directly test whether a more constrained model had a *significantly* poorer fit than a less constrained model; this feature of CFA is one of its major advantages over EFA. However, a decision was made early in the study to correct for possible skewed and kurtotic data through the use of the Satorra-Bentler Chi-square statistic. As a consequence of this, the analysis was slightly more complicated because the Satorra-Bentler chi-square statistic (S-Bχ^2) for nested models is typically not distributed as Chi-square (Satorra, 2000). However, Satorra and Bentler (2001) have developed a scaled difference Chi-square test statistic that can be used to compare S-Bχ^2 from nested models. This statistic was used in the data analysis. A piece of software which ran the test was downloaded from http://www.abdn.ac.uk/~psy086/dept/psychom.htm and executed on the data. The following output was achieved.

Model 2

```
SBDIFF.EXE: Computes significance test on the difference between Satorra-
Bentler scaled chi square statistics

User's Notes: Model 2 Comparison between Tight and Moderate restrictions

 INPUTS:
 Satorra-Bentler chi square for the MORE constrained model= 681
 Normal chi square for the MORE constrained model= 823
 Degrees of freedom for the MORE constrained model= 301
 Satorra-Bentler chi square for the LESS constrained model= 678
 Normal chi square for the LESS constrained model= 817
 Degrees of freedom for the LESS constrained model= 296


 OUTPUTS:
 Satorra-Bentler Scaled Difference = 4.2378  df =  5
 Chi Square probability = 0.515717

-----------------------------------

User's Notes: Model 2 Comparison between Tight and Xmod restrictions

 INPUTS:
 Satorra-Bentler chi square for the MORE constrained model= 681.82
 Normal chi square for the MORE constrained model= 823.18
 Degrees of freedom for the MORE constrained model= 301
 Satorra-Bentler chi square for the LESS constrained model= 652.26
 Normal chi square for the LESS constrained model= 788.57
 Degrees of freedom for the LESS constrained model= 279


 OUTPUTS:
 Satorra-Bentler Scaled Difference = 29.1734  df = 22
 Chi Square probability = 0.139978

-------------------------------------------------------------------------------
```

Model 4

```
User's Notes: Model 4 Comparison between Tight and Xmod restrictions

INPUTS:
Satorra-Bentler chi square for the MORE constrained model= 1133.2
Normal chi square for the MORE constrained model= 1329.2
Degrees of freedom for the MORE constrained model= 415
Satorra-Bentler chi square for the LESS constrained model= 1122
Normal chi square for the LESS constrained model= 1315.5
Degrees of freedom for the LESS constrained model= 410


OUTPUTS:
Satorra-Bentler Scaled Difference = 11.2841  df =  5
Chi Square probability = 0.046029

-------------------------------------

User's Notes: Model 4 Comparison between Tight and Moderate restrictions

INPUTS:
Satorra-Bentler chi square for the MORE constrained model= 1133.2
Normal chi square for the MORE constrained model= 1329.2
Degrees of freedom for the MORE constrained model= 415
Satorra-Bentler chi square for the LESS constrained model= 1103.9
Normal chi square for the LESS constrained model= 1293.9
Degrees of freedom for the LESS constrained model= 392


OUTPUTS:
Satorra-Bentler Scaled Difference = 29.7300  df = 23
Chi Square probability = 0.157312
```

The output for Model 2 shows that the Chi-Square probability for the models performing differently for the two gender groups was equal to 0.516 for Tight - Moderate; and equal to 0.140 for Tight - Xmod. In other words the models were not statistically different at the 5% level, indicating that they performed equally well on both groups even when the parameters were severely constrained to be equal. For Model 4, a probability for the models being different was 0.046 for Tight - Moderate constraint and 0.157 for Tight – Xmod constraint. This shows that there was a statistical difference between the performances of the model in the tight compared with the moderately restricted modes. The outcome of this shows that Model 4 did not perform as well as Model 2 when considering parameter invariance across gender subgroups. For Model 2 the results of the tight replication strategy were not statistically significantly different than for the moderate strategy indicating that the model cross-validated under the strictest conditions which was a very strong sign for its generalisability to different samples.

This was the second test that showed Model 2 to be superior to Model 4. Before Model 2 could be definitively chosen, the other models (Models 5 to 7) that might have provided an alternative and a better fit to the data were considered. The a priori position was to consider science as having an identifiable separation between physics, chemistry and biology, with the model also including a factor representing the methods of science and further factor about the institution of science and scientists. However, for completeness an alternative to this notion was tested. The three alternatively conceived models were, (i) Model 5: The PCB model, (ii) Model 6: The Physical-biological Model, and (iii) Model 7: The Knowledge-process model.

The fit statistics for Models 5 to 7 in comparison to the fit statistics of Model 2 are shown in Table 4.26. The path diagrams for Models 5 to 7 are presented in Diagrams 4.6 to 4.9. As can be seen, Model 5 (Diagram 4.6) had negative parameter values indicating correlational relationships in a direction opposite to that indicated in the model specification. This was evidence of a problematic fit between model and data. The fit statistics other than this were adequate although they were less strong than the original Model 2, and certainly did nothing to suggest that this should be the preferred model.

Model 5 had a higher RMSEA, a lower p-value for RMSEA, lower TLI, lower Critical N, and higher GFI, all suggesting a slightly less well fitting model than Model 2. The Chi-square value of Model 5 was lower than Model 2 although there were far fewer items and it would be expected that the Chi-value would drop under these circumstances.

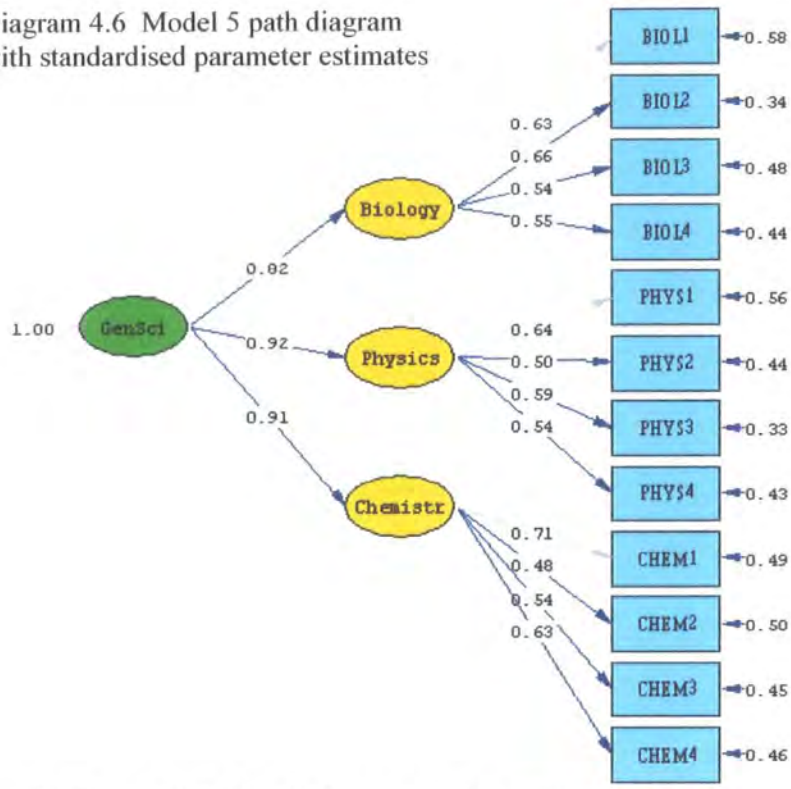
A comparison between Model 2 and Models 6 and 7 yielded a similar outcome. For Models 6 and 7 the Chi-square value increased for the same number of items, which was a poor sign; the ECVI increased, even though the models had similar saturated ECVI values, which was also a poor sign. These ECVI results indicated that Models 6 and 7 have a lower predicted fit to other samples, i.e. the expected cross-validation was poorer. All the other fit indices moved slightly in the wrong direction, which was also a poor sign. The outcome of this was to conclude that none of the alternatively proposed models provided a superior fit to the data than the original a priori hypothesis represented by Model 2.

Table 4.22 Fit Indices for Models 2 and 5 – 7

Fit Index Information	Model 2	Model 5	Model 6	Model 7
Degrees of Freedom	130	186	184	185
Minimum Fit Function Chi-Square	553.96 p=0.00	436.52 p=0.00	795.84 p=0.00	1038.3 p=0.00
Satorra-Bentler Scaled Chi-Square	481.01 p=0.00	388.83 p=0.00	714.34 p=0.00	953.86 p=0.00
Root Mean Square Error of Approx (RMSEA)	0.0440	0.0502	0.0454	0.0545
90 Percent Confidence Interval for RMSEA	0.0398 0.0479	0.0452 0.0553	0.0419 0.0490	0.0511 0.0580
P-Value for Test of Close Fit (RMSEA < 0.05)	0.454	0.016	0.224	0.000
Expected Cross-Validation Index (ECVI)	0.479	0.385	0.688	0.902
90 Percent Confidence Interval for ECVI	0.358 0.454	0.286 0.374	0.523 0.640	0.682 0.820
ECVI for Saturated Model	0.245	0.172	0.331	0.331
ECVI for Independence Model	19.839	14.395	28.071	28.071
Non-Normed Fit Index (NNFI) (Tucker Lewis Index, TLI)	0.985	0.981	0.984	0.978
Parsimony Normed Fit Index (PNFI)	0.835	0.803	0.860	0.860
Comparative Fit Index (CFI)	0.987	0.985	0.986	0.980
Critical N (CN)	495.96	430.04	453.82	341.76
Root Mean Square Residual (RMR)	0.0265	0.0303	0.0275	0.0317
Standardized RMR	0.0339	0.0364	0.0342	0.0391
Goodness of Fit Index (GFI)	0.955	0.957	0.944	0.926
Parsimony Goodness of Fit Index (PGFI)	0.750	0.686	0.752	0.742

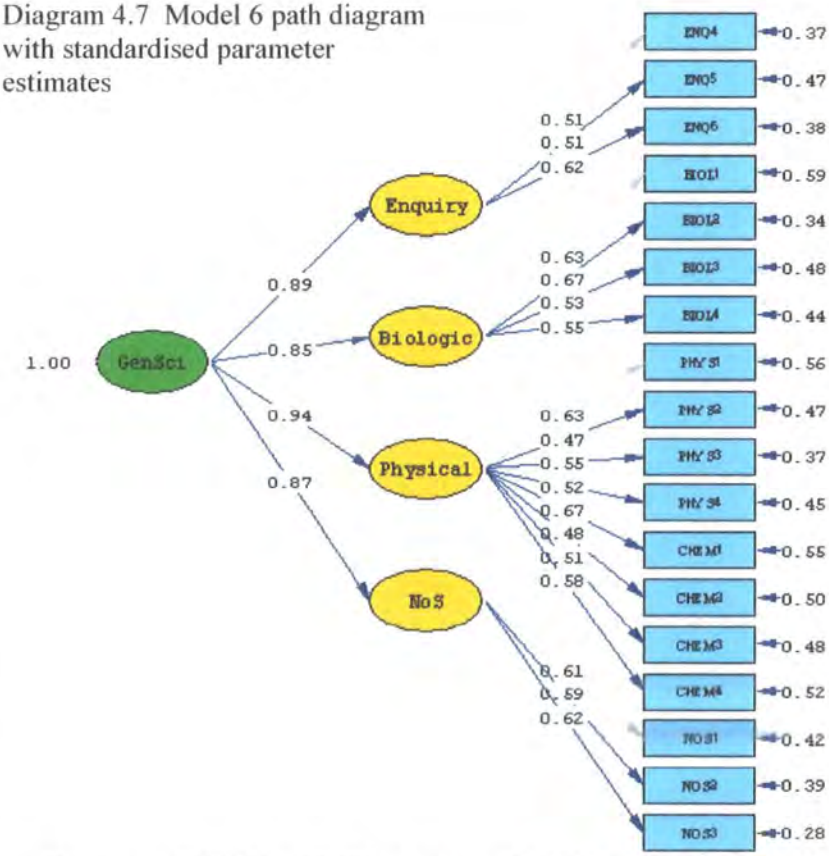
Table 4.22 shows alternative models for the conception of science. Model 5 is the Physics, Chemistry, Biology model. Model 6 is the Physical-Biological science model. Model 7 is the Knowledge-Procedures-Institution model.

Diagram 4.6 Model 5 path diagram with standardised parameter estimates

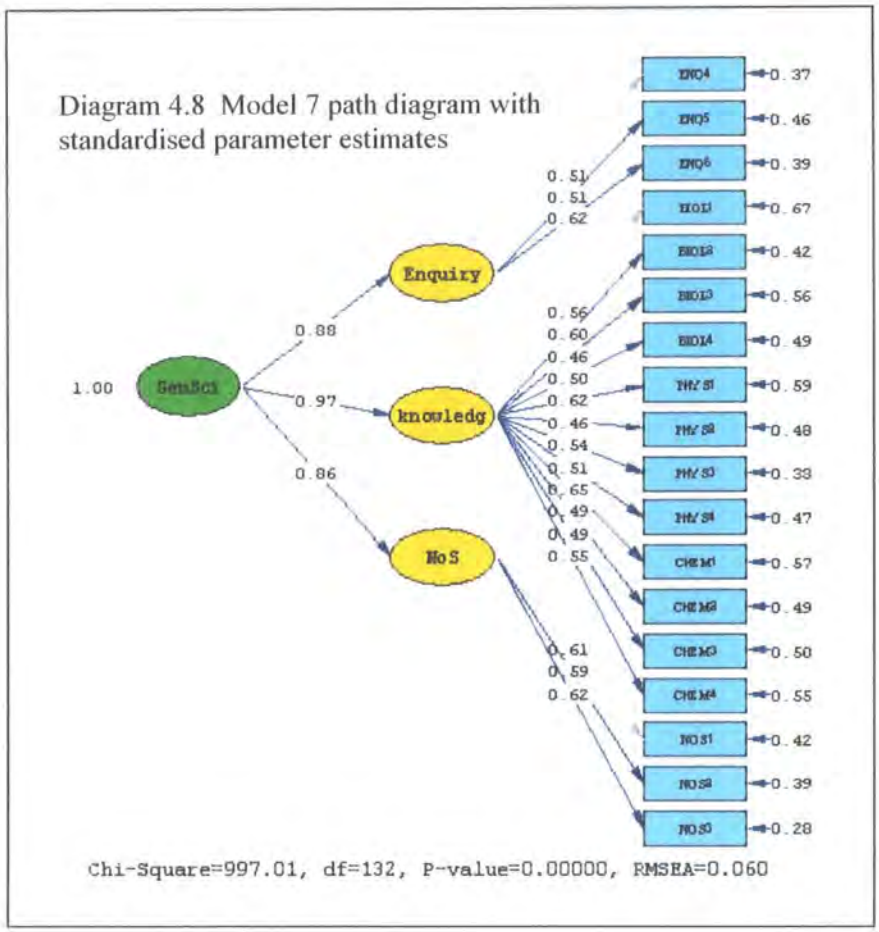


Chi-Square=333.74, df=51, P-value=0.00000, RMSEA=0.056

Diagram 4.7 Model 6 path diagram with standardised parameter estimates



Chi-Square=695.94, df=131, P-value=0.00000, RMSEA=0.049



In summary, seven a priori models were proposed to represent a multidimensional and hierarchical academic self-concept in science. The most complex of these models was Model 4 which included two higher order exogenous science latent variables. The first of these was GenSci which was a derived (inferred) variable from the five endogenous latent variables of Physics, Chemistry, Biology, Enquiry and Nature of Science. The second was the variable GlobalSc which was a constructed latent variable from three 'general science' manifest variables. Models 3, 2 and 1 were simplified and more parsimonious versions of Model 4 whilst retaining the original number of endogenous variables. Models 5 to 7 conceived science in a slightly less extended manner.

Summary

The Data

The sample data was appropriate for the use in model testing. It was wide ranging, from a good collection of schools of different characteristics and didn't suffer from issues of missing data, skewness or kurtosis. The implementation of data onto the theoretical scale structure provided no apparent problems. The reliability analysis of the scales was good with high Cronbach alpha scores across all scales.

Exploratory Factor Analysis

Although several models were proposed in an a priori manner, EFA was employed as the first step in the model testing analysis. EFA with six resolved factors produced the most cognate and meaningful distribution of items across the factors which was consistent with expectation. The EFA provided support for the a priori assumptions.

Parameter Estimates

Confirmatory factor analysis and structural equation modelling was employed for the main body of analysis. No difficulties were encountered with any of the models (except Model 5) and parameter estimates were all feasible. The standard errors were appropriate in size, being neither too large nor approaching zero; all t-values were greater than modulus 1.96 indicating that all parameters were significantly different from zero; the R^2 values were moderate to large indicating well represented latent variables; and correlation coefficients between the various latent variables were moderate to large and statistically significant. These were very positive indications for all models.

Fit Indices

All models performed well on most of the indices of fit. The only drawback was that the Chi-square statistic was always significant indicating that none of the models was a perfect fit to the data. This is a stern criterion to achieve and it is recognized that in many instances of model testing it is too arduous to achieve (see Byrne, 1998). On

most other fit criteria the models did well on TLI, CFI RMR and GFI, and Models 2 and 4 did particularly well on these tests.

Reliability and Validity

Reliability and validity measures at the level of items, factors and model were undertaken, and the resulting data showed the models in a very positive light. This was particularly true with the convergent validity and composite reliability tests.

Cross Validation and Replication

Cross validation was carried out on sub-group samples for Models 2 and 4. This had the advantage of examining the sub-group characteristics and fit statistics at the same time as conducting the cross-validation procedures. Both Model 2 and Model 4 performed well in sub-group analysis with the models behaving consistently and reliably under these conditions. A double cross-validation procedure was undertaken and the Cross validation Indices (CVI) for each model was calculated. The outcome conclusively indicated that Model 2 possessed the greatest predictive validity of the two models. The superiority of Model 2 was confirmed from the results of the tight and moderate replication strategies in which Model 2 performed extremely well and out performed Model 2. At this stage Model 2 was confirmed as the best of the a priori models.

Choice over other Possible Models

Science could be conceived in a manner which was alternative in nature to the structure proposed through the rationale of this study. None-the-less, in order to satisfy the notion of completeness three alternative models for the structure of science self-concept were included for comparison against the final chosen model. Each of these three models (the physics-chemistry-biology model; the physical-biological science model; and the knowledge-procedure-institution model) were compared for goodness of fit against Model 2 and each model was rejected as possessing inferior fit statistics. The final outcome was to choose Model 2 as the best representation for a model of academic self-concept in science.

DISCUSSION AND CONCLUSIONS

The key purposes of this study were to:

- Justify a view for a multidimensional and hierarchical nature of academic self-concept in science;
- provide a model for the way in which academic self-concept in science may be conceived within the context of an extension of the Marsh/Shavelson Model (Marsh et al., 1985);
- use the model as a basis for forming a comprehensive instrument for the measurement of academic self-concept in science;
- clarify issues of gender and age group in relation to academic self-concept in science, particularly concerning issues of measurement validity;
- explore the potential utility of the self-concept measurement instrument for the teachers and learners of science.

This final chapter reviews the evidence built from the collected data. It reviews the outcome from the exploratory factor analysis, including item scores, factor loadings, and differences between gender groups and Key Stages; it reviews the evidence from the confirmatory factor analysis and structural equation modelling, including the quality of the parameter estimates and the model fit statistics; it reviews the issues of structural and external validity, and in particular the evidence that the models are legitimate tools which can be validly used to measure academic self-concept in science across different groups of pupils; it introduces some implications and recommendations for teachers of having access to a science self-concept profile and its application to the context of learning and teaching; it considers limitations of this research; finally, it proposes further research possibilities.

At the heart of this study has been the attempt to demonstrate that that academic self-concept in science can be conceived as being multidimensional and hierarchical in nature. The data emerging from this research does appear to confirm that this is

indeed the case. Evidence for this assertion will be presented in the next few sections, with the remainder of the chapter will be presented thus:

i) Science is multidimensional
ii) Exploratory factor analysis and data processing outcomes
iii) Structural equation modelling, including parameter testing and goodness of fit
iv) Structural validity, external validity and use across groups
v) Implications and recommendations for teachers
vi) Strengths and limitations of this research
vii) Proposals for further research possibilities
viii) Concluding remarks

Science as Multidimensional

At the beginning of this thesis an hypothesis was presented suggesting that for many secondary aged school pupils, rather than science being conceived as a smooth, homogeneous, single discipline, it is conceived as a multidimensional and heterogeneous subject. The justification for this hypothesis is tripolar and based upon (i) curriculum organization and teaching structure, (ii) epistemological differences between physics, chemistry and biology, (iii) pupil interest, choice and motivation.

First, for a large majority of pupils their curriculum experience of science is one of a study of different scientific modules each fitting, more or less neatly, into the physics, chemistry and biology family structure. In fact, even the much heralded introduction of the 2006 National Curriculum for Science (DfES, 2006) with its forward thinking and innovative stance seems to have little potential to alter this situation. The examination boards, in response to the new National Curriculum, have created course specifications which still very much adhere to the traditional tripartite structures. This is true even of the more pioneering specifications like Twenty First Century Science (OCR, 2006). This is not to say that collecting topics together and teaching like-content within an ordered framework is a bad thing; in fact quite the contrary, there is a strong argument for its effectiveness and continued retention.

The issue is not that the situation needs to change, only that this situation needs to be recognized as it will have a cognitive and affective influence on the pupils in our secondary schools.

Second, it is hypothesized that physics, chemistry and biology draw upon, value and build different types of knowledge, employ different sorts of skills, and operate different kinds of success criteria in their pedagogies and assessment frameworks. Two examples will serve to illustrate this. One is the role and application of mathematics in both defining meaning and in establishing outcome in physics vis-à-vis biology (Murphy and Whitelegg, 2006); the second is the differential way in which reductionist thinking is employed and valued in knowledge formation in different ways in the different sciences, e.g. variation in the conception and utility of 'energy' in physics, chemistry and biology.

Third, there is a clear indicator that the sciences have a clearly distinctive nature in the minds of young learners by the sad and frustrating way in which different science subjects differentially attract pupils into further optional study. As a subject at university level, physics has suffered almost to the point of collapse and this has been compounded by the added impediment that physics has historically been much less popular with girls and women when compared with their male counterparts (Smithers and Robinson, 2006).

These three factors, together with others have led to a position, or at least indicate the situation, that secondary aged pupils do not perceive science as unidimensional, but instead, see science as three distinct but allied subjects. These subjects are probably taught by teachers who themselves have these same perceptions, or at the very least possess a single specialism, and quite often an enthusiasm and security of knowledge to match. If this assertion is valid then it is not wholly plausible to suggest that academic self-concept in science can be conceived as being unidimensional. It was this premise, that pupils' academic self-concept in science is multidimensional, which was at the centre of this thesis.

Exploratory Factor Analysis and Data Processing Outcomes

Recall that in constructing the self-concept questionnaires, the National Curriculum for Science was divided up and represented by 32 statements at two levels of specificity (excluding from this items relating to the generic term 'science'). This allowed for an exploration of pupils' academic self-concept across the whole breadth of science and emerging from this were clearly identifiable and distinct science factors occupying at least two levels of hierarchy.

The most striking outcome from the exploratory factor analysis was the cohesive manner in which the items-pairs pulled together to form the various factors. This was striking in two ways. Firstly, the three item pairs occupying the highest level of specificity within each subject domain loaded almost without falter onto the same factor, as anticipated. For the KS4 pupils, there were only two exceptions to an item pair not loading highest on its target factor. The first of these was for the item-pair CHEM2, which didn't load onto the factor with the other chemistry items as predicted. CHEM2 was made up of the two statements 'I have difficulty learning new things about products from oil and rocks', and 'I am good at understanding changes in the Earth and its atmosphere'. On further reflection, the specific content of these two items does suggest that this outcome is not without reason. This is because an item targeted on understanding about 'earth and atmosphere' might well be identified as a topic which finds itself either within physics or biology, depending on the exact nature of the content area. Similarly, 'oil and rocks' does not necessarily and explicitly define itself as belonging to chemistry. These ambiguous content items, although breaking up the predicted pattern, do not diminish from the validity of the instrument, in fact quite the opposite; they highlight the notion that items which contain non-subject specific subject content are likely to receive a non-specific subject response from the pupils. This says more about the sometimes arbitrary nature of our curriculum structures which organizes particular topics to be within certain subject frameworks for pragmatic or historical reasons. In all other cases in KS4 each item loaded most highly on its target factor, and indeed, items often had negligible loadings on all other factors.

Within KS4, and excluding CHEM2, from a total of 126 factor loadings there were only 8 occasions when items loaded on a factor other than their target factor with a value greater that 0.15. By way of comparison, the mean value for items which did

load on their target factor was 0.64. This gives an indication of the distinctness of the exploratory factor analysis outcome. A similar pattern emerged from an examination of the KS3 data which had a consistency of factor loadings which was extremely close to the outcome of KS4. Within KS3 one item clearly did not fit in with the emerging pattern. PHYS1 contained the statement pair "I am good at understanding about physics" and "I am pretty bad when it comes to physics". This pair loaded at near zero on the physics factor but loaded moderately highly on all other factors except NoS. As was reported in an earlier chapter, a number of KS3 pupils when the questionnaire was being administered asked the supervisor for clarification as to the meaning of the word 'physics'. The result may have been as a consequence of the younger pupils not having a clear conception of the nature and content of 'physics' and hence it being associated with a whole array of different factors. This misunderstanding of the term physics may be more significant than the misunderstanding the terms chemistry or biology for younger pupils. Biology, for instance, may have a connection with 'living things' and chemistry with 'chemicals'. This, for the moment, remains unresolved.

In close association with the situation outlined above, a number of other very interesting patterns emerged around the understanding of the term 'physics'. Reference to Table 4.8 will help to clarify these features. Table 4.8 shows two sets of scores, the upper half presents values for the item pairs PHYS1, CHEM1 and BIOL1. These were the generic physics, chemistry and biology statements. The lower half of the table presents the average values for the second, third and fourth item pairs for physics, chemistry and biology which contained the high specificity statements.

The first pattern of note is related to a possible contradiction in the pupils' responses about physics. This can best be seen by comparing the values on the top section of the table with the values on the bottom section of the table. PHYS1 was scored more negatively than either biology or chemistry with a statistically significant difference at a level of $p < 0.0001$. This is not unexpected per se, as physics is often viewed as a learners' *worst subject* (Murphy et al., 2006). However, paradoxically, on considering the high specificity content items, the same individuals scored the average of the PHYS2-4 items significantly lower at the 0.001 level (i.e. more positive) than either the biology scores or the chemistry scores. The effect sizes, (for all pupils), between the physics and the biology scores were 0.14 and the effect sizes

between physics and chemistry scores were 0.24. This clearly makes physics the pupils' most positive self-concept area when measured at the highest level of specificity. This contradiction occurs across both age and gender sub-groups, although it is particularly strong with the boys. This seems paradoxical, with the 'whole' being scored differently to the 'sum of the parts'. These data appear to be indicating that many pupils had either a poor conception of the term 'physics' or a strong negative association with the term 'physics', whilst at the same time rating their self-concept highly against the content of physics. It is difficult to perceive that the label 'physics' would create such a negative connotation for all pupils, but this does in fact seen to be the case. Maybe it shows that these individuals do not dislike physics quite as much as they think they do. This is doubly intriguing as the result is consistent across all the subgroups, is based on a large sample of 1398 pupils, is statistically significant at a level better than 0.0001, and generates meaningfully large effect sizes. This may or may not be of *educational significance* but it certainly warrants further investigation. Interestingly, the scores for SCIENCE are the most negative of almost all the individual scores. Maybe a clue lies in here to explain the physics paradox. Maybe there is a tendency for pupils to score items more negatively as the items become more general. This does not seem to have been reported in the literature and so maybe it is an idiosyncratic effect of this particular data sample. This effect will be monitored in any follow up research.

This issue of PHYS1 compared with the PHYS2-4 goes a little further. For the Biology factor, for instance, there was no statistical difference (at the generous 0.05 level) between the average score of BIOL2, BIOL3 and BIOL4 (2.526) and the score for BIOL1 (2.532); the same was also true for the Chemistry factor (2.629 – 2.665). However, with the Physics factor the difference between the score for PHYS1 (2.750) and the average score for PHYS2-4 (2.378) was statistically significant with $p <$ 0.0001, and with a huge size effect of 0.48. There is certainly something happening with the term 'physics', with a huge number of pupils scoring it disproportionately low compared to other measures of their self-concept.

The collected scores for the KS3 and KS4 pupils showed very strong age consistency. The factor loadings for KS3 were only just slightly less strong than for KS4. This was an unexpected outcome as the literature has suggested that as the pupils' age increases so their self-concept becomes progressively more distinct, (see Marsh and Ayotte, 2003). This does not seem to be the case here. Additionally the

factor scores of the KS3 and KS4 pupils follow a similar profile pattern of raising and lessening negativity across the different items. The KS3 scores always remained slightly more positive than the KS4 scores across all areas with an average effect size of around 0.18. Full data comparisons with significance testing and size effect calculations can be found in Table 4.9. These results are consistent with other outcomes presented in the literature with Marsh (1993b) for instance, reporting that self-concepts decline from early preadolescence to middle adolescence before rising again as the individuals move to adulthood

The gender differences were more marked than the age related differences, which again is not unexpected. Except for the biology scores, where there was no significant different between the sexes; the boys were always more positive than the girls with an average size effect of 0.25. This was not unexpected either, with biology being seen as more gender neutral than physics or chemistry (Murphy et al. 2006). The full data can be found in Table 4.14. These data are consistent with similarly reported results, for instance, Hattie (1991) has reported self-concept differences favouring males in general, physical and mathematics self-concept, and although there were no science data in that study, mathematics and science are often co-related.

Structural Equation Modelling, including Parameter Testing and Goodness of Fit

Following the loose exploration of factor structure from the exploratory factor analysis, four a priori models were proposed to represent possibilities of multidimensional and hierarchical academic self-concept models in science. All four models were based around the notion of science being defined through the latent variables of 'physics', 'chemistry' and 'biology' to represent the conceptual knowledge and understanding aspects of the discipline; 'enquiry' to represent the methodology of science exploration; and 'nature of science' to represent the creative human involvement and the institution of science. This model was consistent with the DES (1988) definition which originally shaped the science of the English National Curriculum. Of these four models, the most complex included two higher order exogenous science latent variables. The first of these was GenSci which was a derived (inferred) variable from

the five endogenous latent variables previously mentioned. The second exogenous variable was GlobalSc which was a constructed latent variable from three 'general science' manifest variables. Models 3, 2 and 1 were adaptations of Model 4, as well as being further simplified and more parsimonious versions of the model.

A full explicit analysis was carried out on Model 4 as this contained most of the structural components of all the other models. The data showed itself to be appropriate for the use of model testing. It was wide ranging, as it had been collected from schools with a good cross-section of characteristics, and the data didn't suffer unduly from issues of missing values, skewness or kurtosis. The implementation of data onto the theoretical scale structure provided no problems. The reliability analysis of the scales was good with high Cronbach alpha scores across all the scales.

Confirmatory factor analysis and structural equation modelling was employed for the main body of analysis. No difficulties were encountered with any of the models with all parameter estimates being feasible with positive outcomes. The standard errors were appropriate in size, being neither too large nor approaching zero; all t-values were greater than modulus 1.96 indicating that all parameters were significantly different from zero; the R^2 values were moderate to large indicating well represented latent variables; and correlation coefficients between the various latent variables were moderate to large and statistically significant. These were very positive indications for all four models showing that the models had passed the first important stage of parameter estimation and that they were structurally secure in the first instance.

The model fit was evaluated from the LISREL output which offers a real 'belt and braces' approach to model fit analysis. However, given all the hard data, the final choices are still something of a mixture of science and art. Marsh and Hau (2002) explain that the way they arrive at a final choice of model is to:

> emphasize the Tucker-Lewis index (TLI), the relative noncentrality index (RNI), and root mean square error of approximation (RMSEA) to evaluate goodness of fit, but also present the χ^2 test statistic and an evaluation of parameter estimates. Whereas tests of statistical significance and indices of fit aid in the evaluation of the fit of a model,

there is ultimately a degree of subjectivity and professional judgment in the selection of a "best" model. (p3).

All four models performed well on the indices of fit with the Tucker Lewis Indices ranging from 0.984 to 0.987, and the RMSEA ranging from 0.0339 to 0.0381. These were all excellent values and above many of the fit statistics for models in the literature, e.g. Marsh and Hau (2002) quoted TLI of 0.97 and RMSEAs all larger than 0.05 for their accepted models. The major drawback with the fit indicators was the Chi-square statistic, which was always significant and indicating a non-perfect fit. However, this Chi-square test is a stern criterion to achieve and it is recognized that in many instances it is too arduous to attain (see Byrne, 1998). Following the analysis of these data, Models 1 and 3 were rejected as being inferior to Models 2 and 4.

Structural Validity, External Validity and Use across Groups

Reliability and Validity

At the level of items and factors the data indicated very strong reliability and validity measures for the remaining models. This was particularly true for the convergent validity and composite reliability tests. In considering Model 4, for instance, the composite reliabilities for the latent variables were all safely in excess of the threshold of 0.6 suggested by Bagozzi and Yi (1988) with latent factor values ranging from 0.686 for Enquiry to 0.765 for Global Science (see Table 4.24).

The final test of structural validity related to comparisons between gender and age sub-groups since it was important to be secure in knowing that the instruments measured the same component of self-concept with equal validity for both boys and girls of different ages. That is, key checks needed to be made to ensure that the instruments were measuring the same psychological construct in exactly the same way across the different groups. To this end the instruments were tested such that they could clearly demonstrate that there was both measurement (i.e. factor loadings) and structural (i.e. factor covariances) equivalence across the groups. This was implicitly assumed when making comparisons of observed means across gender groups or across ages, and if this wasn't true then these comparisons were

unjustified (see for example Byrne and Shavelson, 1987; Marsh, 1993b). The construct validity of self-concept is contingent on this, and Byrne and Shavelson (1987) have argued that the inconsistency manifest in earlier reported research findings was undoubtedly attributable to the violation of the assumption of equivalent self-concept measure and structure.

It is possible to carry out multiple group comparisons of factor means analysis using exploratory factor analysis, and this would provide factor loading comparisons and relations among self-concept facets across gender groups. Indeed this has already been undertaken and reported earlier. However, there are a severe limitations with this technique in that it doesn't have the capability to address the invariance of the factor loadings, or the covariance across groups, or the power to test specific hypotheses (see Marsh, 1993b). This limitation was addressed by undertaking the evaluation by using confirmatory factor analysis, which by contrast, does not suffer from these limitations and does allow for multi-group equivalence testing through the determination of factorial invariance. It is also able to take account for error variance by simultaneously analyzing data across different groups.

The factorial invariance was tested by systematically fitting the two data sets to ever more restricted models by constraining some parameter estimates (i.e. factor loadings, variance and covariance parameters) to be equal across groups. The imposition of these constraints did not result in a noticeably poorer fit for either model and the results provide strong evidence for the invariance of factor loadings, factor correlations and factor invariance. Table 4.25 shows that the TLI and the RMSEA barely changed as more restrictions were imposed which was excellent support for the invariance across gender.

The performance of Model 2 was slightly superior to that of Model 4. The TLI, the GFI and the RMSEA were all slightly stronger, but more importantly for Model 2 there was no statistical difference in the change of Chi-square statistic as more restrictions were imposed on the model. This is an extremely positive outcome and shows that the model replicates well even under the strictest of conditions.

Cross validation was carried out on sub-group samples. This had the advantage of examining the sub-group characteristics and fit statistics at the same time as conducting the cross-validation procedures. Both Model 2 and Model 4 performed

well in sub-group analysis with the models behaving consistently and reliably under these conditions. A double cross-validation procedure was undertaken and the Cross validation Indices (CVI) for each model was calculated. The outcome conclusively indicated that Model 2 possessed the greatest predictive validity of the two models. The superiority of Model 2 was confirmed from the results of the tight and moderate replication strategies in which Model 2 performed extremely well and Model 4 not unwell, but not as strong as Model 2. At this stage Model 2 was confirmed as the best of the a priori models.

This is a significant stage of the research as the position of deciding on a single clearly best model has been reached. The model demonstrates multidimensional and hierarchical features and behaves robustly under many of the test conditions. A threshold has been reached whereby the utility of the model needs to be investigated. Some of this potential will be explored, in outline, in the sections below.

Implications and recommendations for teachers

Self-concept as an evaluative and measurable characteristic of young learners is sadly neglected in most schools today, and this despite its potential to influence a number of different and important educational processes and outcomes. Skaalvik and Rinkin, (1995) for instance, have shown that self-concept can directly influence motivation and effort. Chapman, (1988), along with Marsh, Byrne and Yeung (1999), and Valentine, DuBois and Cooper (2004) have demonstrated that a positive self concept has a direct correlation with raised academic performance; and Marsh, Craven and McInerney (2005) have gone so far to suggest that their results demonstrate a *causal influence* between self-concept and educational outcomes like attendance, participation, achievement, coursework selection and school enjoyment.

In the face of a growing body of respected research outcomes, self-concept has yet to attain any sort of profile in the minds of teachers, nor indeed have many other related psychological measures including self-efficacy. On those occasions when teachers have been asked explicitly to consider a construct like self-concept (usually within research projects) the outcomes have been rather mixed. Firstly, teachers have shown themselves to be particularly poor as judging the level of a learner's self-concept. Carr and Kurtz-Costes (1994) showed that when teachers were asked to

estimate their pupils' self-concept they used 'achievement' as a reference marker which resulted in them attributing high self-concepts to highly achieving pupils, and vice versa. Psychological measuring instruments later found the pupils' self-concepts to be not nearly so correlated to achievement as the teachers thought, resulting in the teachers' self-concept estimates being disproportionately high for their 'smarter' pupils. The difficulty for teachers is that self-concept is not a manifestly visible attribute, and therefore it is not unexpected that they use other educational indicators (like achievement) to assist the formation of their judgment, even when this turns out to be unhelpful.

Without a self-concept instrument to provide a valid and reliable measure in science, the chance of self-concept theory making any noticeable impact on science educational practice is negligible. There are existing instruments which can be used 'off-the-shelf', and the ASDQ (Marsh, 1999) is an excellent example. However, self-concept research has come more and more to the conclusion that cross-network associations are more secure (and valid) when the self-concept measurements are collected at a higher level of specificity. This has been shown to be particularly important when considering academic self-concept and achievement measures (see Marsh, 1992; Marsh et al., 2005). If the self-concept is measured at a lower level of specificity, and teachers or researchers make unwarranted claims (about associations) then there is the danger that the critically negative scenarios suggested by Baumeister et al. (2003) which were concerned with unproven linkages, or worse still, being trapped by the *broad seductive appeal* of self-concept, could become legitimated. A high specificity self-concept instrument is needed in science. In other curriculum areas, the predictive utility of a self-concept measure has been shown to bring about desirable educational results, see for example Chen's (2006) work in mathematics, and this could be an extremely useful additional tool in the armory of teachers of science. Mathematics is far more advanced, in a research perspective, (e.g. Marsh et al., 2004; Pietsch, 2003) and science needs to engage in some catch-up.

The pace of development however, needs to be slow and considered. Despite the research evidence being strong, what is not needed in our schools is an avalanche of scales or inappropriately applied theory. There are real and active dangers when moving from a research context to wholesale pedagogical implementation. This has happened before on more than one occasion and there is the potential to cause more

harm than good. It would be unfortunate if self-concept research (on an admittedly less grand scale) suffered the same fate as the learning styles research. Significant amounts of attention has being paid to implement the learning styles agenda and schools, either through push or jump, have pulled their bandwagons into a protective circle to guard against the OfSTED intruders by providing kinesthetic and aural learning experiences to cater for their pupils 'individual' needs. This is not to suggest at all that notions of learning styles are invalid per se, for there are many important principles that ought to find their way into the policies and practices of school. What has much less validity though, has been the way that the learning-styles ideology has been operationalised into our schools, by making *simplistic judgments* or using instruments with 'low reliability, low validity and negligible impact on teaching and learning' (Coffield, Moseley, Hall and Ecclestone, 2004, p.56). Even when good quality learning styles instruments were available they were often inappropriately used resulting in less meaningful outcomes (Veenman, Prins and Verheij, 2003). Education, including self-concept researchers need to learn from these lessons.

There is a real and useful role for an academic self-concept instrument in science. There is a strong evidence to link high self-concept in a subject with high achievement in that subject and vice versa. Some pupils may have a lower self-concept in science (for them) than their understanding and performance warrants. These pupils *may* be underachieving either because of this low self-concept or because of reasons closely associated with this low academic self-concept. Targeted and *contingent* self-concept (Crocker and Knight, 2005) enhancement may be of significant benefit to these individuals. This is not to say that wholesale self-concept enhancement is being advocated, in fact quite the opposite. Excessive or unjustified teacher praise may well be to the learners' detrimental. Elliott et al. (2001) have pointed out that learners in schools both in the UK and the US have been conditioned by overly positive teacher evaluations resulting in negative educational outcomes. Other researchers (e.g. Damon, 1995; Stevenson and Lee, 1990; Stevenson and Stigler, 1992) have found evidence of a relationship between high levels of self-satisfaction and lower expectations and work rates. They have consequently called for restraint on any hyping up, by advocating that appraisals be more realistic.

Given that raising levels of self-concept when it is clearly not warranted serves little educational purpose and may indeed cause educational harm, we are led towards a

dilemma. Excessively low levels can be debilitating whilst excessively high levels creates complacency. There is a third element to add to this mix. Science, and particularly physics, is perceived by many young learners (and especially girls) to be excessively hard (Murphy et al., 2006). This has the consequence for many of opting out or giving up.

Without instruments which provide a reliable and valid measure of science self-concept it would not be possible to accurately discriminate between where the lifting of self-concept is warranted and where it is not. It certainly cannot be left for the judgment of teachers (see Carr et al., 1994). If learning mentors and other education professionals are to have a powerful impact in our schools, then what better role could they take than in helping to raise, where necessary, the self-concept of our learners in a specific aspect of science (for instance) to a level which is consistent with a more appropriately considered target level. Pupils need to know that their effort will pay off and that a personal investment of energy and emotion into their learning will have positive outcomes. As Hufton, Elliott and Illushin (2002) have argued, one of the answers to better motivation and engagement lies, not with reforms or systems, but with convincing learners that 'working harder will produce gains that have both meaning and value' (p.284).

Achievement in science, although important is not the only issue where an accurate measure of self-concept is needed. Science, and especially physics, has a problem with attracting and retaining students to its courses (Smithers et al., 2005). If this problem is not resolved in the near future there are possible catastrophic effects ahead as the decline in students studying science has been increasing year on year (Institute of Physics, 2006). This is particularly relevant in physics and most acute in relation to the numbers of girls studying physics. The reasons for this are too complex to rehearse here, however, Murphy and Whitehead (2006) have hypothesized that one of the factors influencing the decline in physics numbers relates to the effect of differences in self-concepts of boys and girls. They call for more research in this area to 'understand how boys and girls feel about themselves in relation to science and physics (together with) evidence and tools for teachers to use in dealing with gender differences in physics' (p.11). Reliably and validly measuring a learners' science self-concept profile is the first step in addressing a quite complex issue. If this instrument is an aid to meet those ends then it is indeed a worthwhile contribution to the cause.

One potentially positive, although paradoxical, outcome is that there seems to be evidence that girls are not quite as negative about physics as they think they are. When questioned using the term 'physics', the girls (as did the boys) responded in much more negative terms than when given the same statements in relation to biology or chemistry. However, when asked about the specific content of physics, those same girls rated physics much more positively than either the similar questions about other biology or chemistry, or the direct questions about the term physics. This may reflect an image problem where, according to Murphy et al. (2006), there is an age-related increase in the perception of physics being difficult, and an accompanying increase in the sense of subject inadequacy. These effects are more strongly measured in girls. There seems to be some further work that needs to be undertaken in this area in order to identify what it is that learners (and especially girls) perceive to be so difficult about physics and what drives these perceptions.

Strengths and Limitations of this Research

This study represents the first attempt to explore the multidimensional aspects of academic self-concept in science. In addition, it is the only empirical study to examine the construct validity for science self-concept and to check for invariance across age and gender. It represents real and genuine advances in knowledge for self-concept studies and accordingly represents an important addition to existing self-concept research.

The limitations of the study can be grouped with four broad areas, (i) instrument design, (ii) sampling, (iii) breadth of study, (iv) and transferability into other cultural settings. These will be explored below.

Although much thought and preparation went into the instrument design there are certain features which would benefit from re-evaluation. A decision was taken early in the research to include a balance of positively and negatively worded statements. Although there is no evidence that this decision had a detrimental effect on the data, a more prudent approach would have developed statements items in only the positive orientation. Marsh (1990c) and others, e.g. Lau et al. (1999) have measured what they perceive to be a negative item effect, and this present instrument may

have suffered similarly. The negative effect would be extremely difficult to detect with this instrument as each item pair contained the average of a negatively and positively worded statement. If a negative effect did exist then the questionnaire structure might well mitigate against it, limiting its effect since the negatively worded items are spread evenly through the instrument and included with every item pair. There is a precedent for this particular arrangement of items. Marsh (1993b) in his research project to test self-concept invariance over gender and age used multiple instruments in which each self-concept scale consisted of 'responses to 10 self-report items (half of which were negatively worded items that were revered scored)' (p 848).

Earlier in the thesis the issue of 'identically worded statements' was explored. The lack of identically worded statements across items prevented the direct comparison between the scores on individual statements. This was not too much of an issue because the analysis was undertaken at the either the level of item pairs or at the level of whole factors, however, in any subsequent developments of this instrument, this issue will be re-explored and a different outcome may be reached.

The choice of items to be paired at the time of decision making seemed to be an unproblematic operation. However, since there has been opportunity to examine more closely the outcome of the factor analysis, there might in some cases be an argument for a limited change to a small number of the item pair arrangements. This may well increase the level of fit between the model and data as more robust pairing criteria may well be able to smooth off some of the rough edges of model fit.

A more fundamental issue is concerned with the interpretation of items by KS3 pupils. By its very nature the instrument was designed to measure a response from the pupils to topics across the whole curriculum within that Key Stage. This introduced a possible problem for pupils who were just beginning the Key Stage and had received little exposure to the scientific ideas to inform their judgment. It might be reasonable to expect that under these circumstances the topics which were more unfamiliar to the pupils might have been scored proportionately more negatively on the response grid. However, this does not seem to be the case and the issue does not seem to have manifested itself as a problem. One way to check if there was a non-curriculum exposure effect was to compare the pattern of responses between the KS3 pupils and KS4 pupils for the generic items which shouldn't have been affected by specific teaching (e.g. SCI1 and SCI2) with the high specificity items

which were more likely to be teaching specific. Reference to the two sets of scores (see Table 4.10b) shows that there does not seem to be a change in pattern between the two sets of items. That is, the KS3 pupils generally responded more positively to items SCI1-3 and NOS1-3, which were all general in nature; and also equally positively to those specific items (BIOL2-4, PHYS2-4 and CHEM2-4) where it might have been thought that a lack of teaching might well have facilitated a more negative response. Table 4.10b clearly shows however, that the KS3 pupils remained slightly more positive than the KS4 pupils across all the items, with a differences in their positiveness always about the same with a size effect of around 0.15.

The second limitation was concerned with the nature of sampling. Schools were selected to be a part of the study on the basis that they were representative of the different types of school found in England. In this way the participating pupils would have been drawn from schools with differences in pedagogic, social, geographic and economic contexts. These school types were; a fee paying independent grammar school, a science specialist school, a technology college, a Roman Catholic School, an urban comprehensive (in special measures), a suburban comprehensive and a rural comprehensive. Two schools were situated within a poor social/economic area and one school had a high proportion of ethnic minority pupils. All of these schools were situated in the North West of England. The primary principle of selection for participation occurred at the level of the school. Following this, classes were chosen on the basis of convenience for those schools participating, with all pupils in a particular class being asked to take part in the study. As such, the sampling frame contained pupils that were neither randomly selected nor representative of pupils either at a local or national level.

The sample size for each Key Stage was 507 (KS4) and 981 (KS3). This allowed each sample to be divided into gender groups for the purpose of covariance structure analysis. This is strongly recommended by, for example, Browne and Cudeck (1989) for cross-validation comparisons. This was carried out as reported above, however, the size of each Key Stage sample precluded any further subgroup testing where the sample contained three or more subgroups, as the extreme minimum sample size is recommended to be 100 and preferably 200 participants (Boomsma, 1982). A larger sample size might have more validly facilitated age selection rather than Key Stage selection; school selection; or English as a first language selection.

Given both the methodological sampling (and associated analysis) weaknesses described above, there is need for further research to ensure that the patterns and outcomes of this study can be demonstrated to be securely generalized to the whole population.

The third limitation was concerned with the breadth of the study. This current research concerned itself with an exploration of an expansion of the academics-self concept of science into its multidimensional and hierarchical elements. This has been carried out thoroughly and systematically. However, this exploration was confined solely to the science or general school elements. There was no opportunity, within this research, to study the relationship between the multidimensional facets of science self-concept and the self-concept measures in other curriculum areas, or indeed performance/achievement data in other areas. This would have been particularly useful in generating evidence for the hierarchical nature of science self-concept, as correlational analysis might have taken place to evaluate correlation coefficients between different hierarchical levels of different subjects.

The final limitation is the issue of transferability of the instrument into international settings. This instrument was explicitly designed to measure academic self-concept in science of secondary aged pupils in England and Wales. In terms of achieving this target, the evidence suggests that this instrument has been more than reasonably successful. Although data has yet to be gathered to show statistical evidence of the valid applicability to all pupils in England and Wales, the methodology was robust enough for there to be high confidence that the evidence when collected would be favourable. However, the external validity should be able to go one step further and be able to say something about the instrument's likely success in other English speaking countries.

Teachers, researchers and clinicians all want to be able to learn something about how different groups of individuals from different cultural settings differ (or not) in their self-concepts in various subjects. Comparison of mean scores has proved to be a fertile ground for enquiry. However, in any such comparison a strong case must first be made that the instrument is measuring the same construct in exactly the same way in each cultural setting. If this assumption is not valid then the comparisons are of much less value. For Byrne (2003), the greatest potential problem in transferability from one culture to the next is bias. She identifies three

types of bias; construct bias, method bias and item bias. The likelihood is that the instrument developed here would suffer on all three counts.

In this particular context, one strand of *construct bias* was concerned with the definition of 'science' and in particular, the broadness of definition and the relative important of the numerous facets. Recall that for this study the definition of science has been taken from the Non-Statutory Guidance from the original National Curriculum (DES, 1988) which stated:

> . . . science is enquiry-led and concerned with the pursuit of better investigative strategies and more reliable information about the physical and biological word (p.A6).

This particular flavour of science places enquiry at the heart of the scientific enterprise, and furthermore, values the human, creative aspects as represented by the items relating to the nature of science. Other cultural representations will not match this conceptual profile and therefore are unlikely to have a factor pattern and weighting of loadings which is invariant across the cultural groups. This may well have an irreconcilably negative effect on the *construct bias.*

This instrument is equally likely to come into difficulty with *method bias.* This is likely to manifest itself in two ways. First, there is *sample bias,* where participants from different countries of equal age are not necessarily of equal experience or have had equal opportunity to develop there ideas, skills and emotions. In the UK science is a statutory part of the National Curriculum, and as such most pupils spend around 20% of their curriculum time on the study of science. The situation in the UK even as recently as 1988 was vastly different from this (pre-National Curriculum) and is equally vastly different to the whole host of different curriculum models which have grown up all around the world. Universality of curriculum structure and content is fortunately an event which is unlikely ever to see the light of day, but this cultural richness brings with it the problems of sample bias. A small price to pay, I think!

Instrument bias, by the same token, is likely to be another element method bias which is likely to manifest itself. The Likert scale is familiar territory for UK pupils who are well used to such formats in popular culture quizzes, curriculum teaching material (in arts and humanities teaching) and in school evaluation documents. The notion of

the 4, 5 or 7 point scale is well known and well practiced. This bias reveals itself in two ways, either because of inexperience, or because of cultural differences in the way that individuals from different cultural settings respond to the task. Marín, Gamba and Marín (1992) report that Hispanic pupils, for example, are much more likely to use the extreme ends of a multi-category scale than non-Hispanic pupils.

The final difficulty is likely to arise from the social-comparison contribution to self-concept. Byrne (2003) reports that there are widely different cultural perceptions of 'others' and these have consequential effects on the relationships between 'self' and 'others'. This is particularly relevant in comparisons between Western and Eastern societies and is likely to lead to a differential perception of self in relation to others, which in turn, is likely to lead to *item bias.*

These difficulties are not meant to provide a reason not to take this instrument into other cultural settings, and indeed until such work is undertaken the issues and problems remain at the theoretical stage. The issues were discussed such that it is recognized how difficult the process can in transferring a perfectly good instrument (if one exists) from one cultural place to another.

Proposals for Further Research Possibilities

This present research is very much near the beginning of multidimensional self-concept research in science. Self-concept research in science per se is important because it has a contribution to make to the knowledge of the discipline. It also helps to reinterpret the known boundaries, such that everything which is within those boundaries is better understood. However, research which connects this understanding to other knowledge areas beyond itself is arguably even more important. New research should move in two directions and concern itself both with 'within network' and 'between network' research. According to Cronbach et al. (1955), the valid use of a construct, within the measurement theory paradigm, requires that both within network and between network characteristics be investigated; a process which validates a construct's 'nomological network'. Such within network research might include an examination of the relationship between multidimensional facets of science self concept structure and either; multidimensional facets of other curriculum areas, particularly mathematics, or general academic self-

concept, or global self-concept. This would, amongst other things, provide more evidence as to the nature of the hierarchical structure of self-concept in science. In contrast to the within network research, the between network investigation could centre on the relationships between the various facets of science self-concept and performance and/or achievement in science. This might be further developed as a diagnostic or predictive tool to cast light on such pedagogical themes as underachievement. Two pilot studies into between network research have been undertaken as an extension of the research reported here. One of these studies relates to the development of a model to explain the interaction effect of 'importance' to science self-concept, general school self-concept, and achievement. The other pilot study is concerned with the pedagogical utility of the use of academic self-concept in science as an aid to better predictive models of future performance in such a way as to preemptively identify potential under-achievement. Provisional outcomes are presented below.

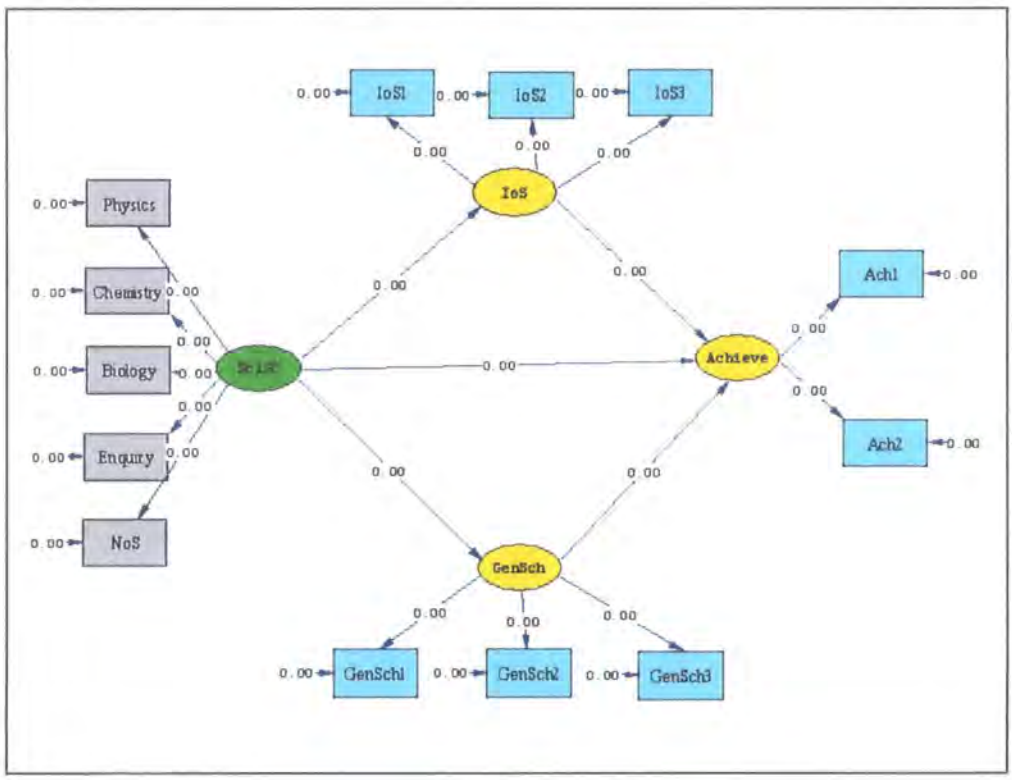
(i) Model for Importance of Science

It is becoming increasingly recognized (see Vispoel, 2003) that there is considerable variation across individuals in the relations of particular facets of self-concept. This may be explained by the mediating factor of domain importance. Marsh (1993b) has suggested two models to explain domain importance; the mediating or additive model where importance ratings are added to the model, and the moderating or multiplicative model where the domain specific self-concept score is multiplied by the importance interaction term and then added to the model. Harter (1986) proposed a similar notion to Marsh's mediating model with her 'discrepancies of self worth'. Here an individual's low evaluation of skill or knowledge would have a negative impact on self-concept if the individual also placed a high value on the specific skill or knowledge. Vispoel (2003) has more recently undertaken model testing of Marsh's moderating model and has secure evidence to the mediating affect of 'importance' within the specific domain of artistic self-concept. This is an area of research which ripe is for further study.

Within this present research project four Importance of Science items were collected within the science self-concept instrument. The normal scale testing procedures were carried out at the same time as the other statistical work on the self-concept scale. Consequently data have been collected and processed on academic self-

concept in science, importance of science, general school self-concept and school performance data (from two of the participating schools).

Model testing is ongoing at the present time and one of the models being tested at the moment resembles the path diagram below. Further developments will be achieved in the near future.



(ii) Cross-network Studies of Science and Achievement

A strand of research which is emerging from this research project is the utility value for teachers of the predictive powers of academic self-concept in science, in particular to preemptively identify potential underachievement in science. One of the schools in this present research project utilized MIDYIS and YELLIS[1] information systems in order to assess their pupils' present achievement and predict future achievement. These information systems are then used by teachers to better

[1] MIDYIS and YELLIS are information systems for monitoring pupils progress, see http://www.cemcentre.org/

monitor and evaluate the progress of their pupils together with the effectiveness of their teaching. Predictions from the beginning of Y10 explained around 60% of the variance in final GCSE grades which was very effective indeed. However, results from this school from a small sample of 60 pupils shows that the self-concept data can add around a further 10% to the predictive power of their information systems. That is, pupils with a high pre-test score and a low self-concept have been shown to underachieve, whilst pupils with a low pre-test and a higher self-concept score have been shown to overachieve relative to their predicted performance. When the Importance of Science was added into the regression calculations the explained variance further increased. This pilot was undertaken on a relatively small sample, however, the results were consistent enough to warrant a full sample undertaking.

(viii) Concluding remarks

The research carried out here is consistent with the pattern of other multidimensional self-concept research undertaken in other areas (see Lau et al., 1999; Vispoel,1995). There is strong evidence to support the notion that self-concept in science is multidimensional, and other evidence although less strong, that is supportive of self-concept in science being hierarchical. The global science self-concept is multifaceted in nature and clearly derives from self-concepts which include physics, chemistry, biology, enquiry and the nature of science. Structural equation modelling has demonstrated that secondary aged learners clearly discriminate between these individual factors and that variances within these factors can be accounted for by a single general higher order science factor. There was also an extremely strong association between this derived general science factor and a higher order global measured factor. However, in the final analysis, this model, with a higher order global factor, was dropped in preference for a more parsimonious model with a simpler higher order structure.

The purpose of this research was not to arrive at a definite final model for academic self-concept in science, but rather (i) to demonstrate that the principle of a multidimensional structure of self-concept in science (and its measurement) was both plausible and desirable, and (ii) to explore a range of models, to test their viability and to make some tentative suggestions as to which model may tell the best story about science self-concept given the evidence gathered from the data.

The theoretical contribution of this study to the field of self-concept research has been through the demonstration of a secure self-concept hierarchy in the domain of science in which the higher order science construct was capable of explaining the variance of the lower order factors. This model showed itself to be invariant across gender and age range groups and has satisfied some of the most stringent model testing procedures. This instrument provides a robust means of measuring self-concept for secondary aged learners and may provide a valuable additional instrument for other self-concept researchers interested in the science domain.

At a practical level, there is in school a need for an easily applicable but robust science self-concept instrument. There is currently no means by which teachers of science can gauge the level of self-concept of their learners, in their subject, and in their classes. Additionally, we find ourselves in a school culture where performance and results are sadly the most important educational outcome by any measure. If we are to aid teachers in being more effective practitioners, and to aid pupils as being more effective learners, then a higher consciousness of self-concept in our schools, along with a better understanding of its contribution to the learning and teaching process may be of positive benefit. As Hamachek (1995) noted:

> Research does not permit us to say that a high self-concept will automatically lead to high achievement, but it does allow us to conclude that high achievement rarely occurs in the absence of a reasonably high self-concept. Although we cannot say definitively which comes first, good schoolwork or high self-regard, we can say that they are mutually reinforcing to the extent that a positive change in one encourages a positive change in the other (p. 364).

One thing is for sure, a better understanding of our pupils, their perceived strength and weaknesses, their fears and motivations, their difficulties and delights must be of positive benefit to all of us concerned with the education process.

REFERENCES

Alva, S.A. and de Los Reyes, R. (1999) Psychological Stress, internalized symptoms, and the academic achievement of Hispanic adolescents. *Journal of Adolescent Research, 98, 102-108.*

Anderson, K. (1992) Self-complexity and self-esteem in middle childhood. In Lipka, R. and Brinthaupt, T (Eds.) Self-perception across the life span, 11-52. Albany: State University of New York Press.

Angell, C., Guttersrud, Ø. and Henriksen, E.K. (2004) Physics: Frightful, But Fun Pupils' and Teachers' Views of Physics and Physics Teaching. *Science Education Volume 88, Issue 5, 683-706*

Bagozzi, R.P. and Yi, Y. (1988) On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16, 74-94.*

Bandura, A (1986) Social *Foundations of Thought and Action: A Social Cognitive Theory.* Prentice-Hall, Englewood Cliffs, NJ.

Bandura, A. (1997) *Self Efficacy: The Exercise of Control.* New York: Freeman.

Bauer, C.F. (2005) Beyond "student Attitude": Chemistry self-concept inventory for assessment of the affective component of student learning. *Journal of Chemical Education, Vol.82, No.12, 1864-1870.*

Baumeister, R.F., Campbell, J.D., Krueger, J.I. Vohs, K.D. (2003) Does high self-esteem cause better Performance, interpersonal success, Happiness, or healthier lifestyles? P*sychological science in the public interest, vol. 4 no. 1, May 2003.*

Bearden W.O., Sharma, S. and Teel, J.E. (1982) Sample size effects on chi square and other statistics used in evaluating models. *Journal of Marketing Research, 19, 425-430.*

Bekhit, N.S., Thomas G.V. and Jolley, R.P. (2005) The use of drawing for psychological assessment in Britain: Survey findings. *Psychology and Psychotherapy: Theory, Research and Practice, 78, 205–217.*

Bennett, J. (2001) The development and use of an instrument to assess students attitudes to the study of chemistry. *International Journal of Science Education Vol. 23, No. 8, 14 6, 833-845.*

Bennett, J. and Lubden, F. (2006) Context based chemistry: The Salters' approach. *International Journal of Science Education Vol. 28, No. 9, 999-1015.*

Bentler, P.M. (1992) On the fit of models to covariances and methodology to the *Bulletin. Psychological Bulletin, 112, 400-404.*

Bollen, K.A. (1989) *Structural equations with latent variable.* Wiley-Interscience

Bong, M. and Clark, R. (1999) Comparison between self-concept and self efficacy in academic motivation research. *Educational Psychologist, 34, 139-154.*

Branden, N. (1994) *Six Pillars of Self-Esteem*, New York, Bantam.

Bracken, B.A. (1992) *Multidimensional self-concept scale.* Austin, TX: Pro-Ed

Bracken, B.A. (1996) Clinical applications of context-dependent, multidimensional model of self-concept. In B.A. Bracken (Ed.), *Handbook of self-concept* (pp. 463-504). New York: Wiley.

Bracken, B.A., Bunch, S. Keith T. Z. and Keith, P. B. (2000) Child and Adolescent Multidimensional Self-Concept: A Five-instrument Factor Analysis. *Psychology in Schools, Vol. 37 (6), 483 – 493.*

Brannick, M.T. (2006) *Path Analysis.* University of South Florida, College of Arts and Science. Accessed online at http://luna.cas.usf.edu/~mbrannic/files/regression/Pathan.html on 7/8/2006

Browne, M.W. and Cudeck, R. (1989) Single sample cross-validation indices for covariance structures. *Multivariate Behavioural Research, 24 (4), 445-455.*

Browne, M.W. and Cudeck, R. (1993) Alternative ways of assessing model fit. In Bollen, K.A. and Long, J.S. (Eds.) *Testing structural equation models* (445-455). Newbury Park, CA: Sage.

Buck, J. (1948). *The house–tree–person technique.* Los Angeles: Western Psychological Services.

Burnes, R. (1979) *The self-concept in theory, measurement, development and behavior.* New York: Longman.

Byrne, B.M. (1990) Methodological approaches to the validation of academic self-concept: The construct and its measures. *Applied Measurement in Education, 3, 185-207.*

Byrne, B.M. (1994) *Structural equation modelling with EOS and EQS/windows: Basic concepts, applications and programming.* Thousand Oaks, CA: Sage.

Byrne, B.M. (1996a) Academic self-concept: Its structure, measurement, and relation to academic achievement. In B.A. Bracken (Ed.) *Handbook of self-concept: Development, social, clinical considerations.* (287-316). New York: John Wiley and Sons, Inc.

Byrne, B.M. (1996b) *Measuring self-concept across the life span: Issues and instrumentation.* Washington, D.C : American Psychological Association.

Byrne, B.M. (1998) *Structural equation modelling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications and programming.* Lawrence Erlbaum Associates, New Jersey.

Byrne, B.M (2003) Testing for the equivalent self-concept measurement across culture: Issues, caveats and applications. In, Marsh, H.W., Craven R.G. and McInerney, D.M. (Eds) International Advances in self research, (pp 291-313). International Age Publishing Inc, Connecticut.

Byrne, B.M and Worth Gavin, D.A. (1996) The Shavelson Model revisited: Testing of the structure of academic self-concepts across pre-, early, and late adolescents. *Journal of educational Psychology, 88(2) 215-228.*

Byrne, B.M. and Shavelson, R.J. (1986) On the structure of adolescent self-concept. *Journal of Educational Psychology, 78, 474-481.*

Byrne, B.M. and Shavelson, R.J. (1987) Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Educational research Journal, Fall 1987, Vol 24, No. 3 pp 365-385.*

Byrne, B.M. and Shavelson, R.J. (1987) Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Education Research Journal, 24, 365-385.*

California Department of Education (2002) *School Safety, Discipline, & Attendance State Board of Education Policy #01-02.* http://www.cde.ca.gov/be/ms/po/policy01-02-mar2001.asp (28/05/06)

Calsyn, R.J. and Kenny, D.A. (1977) Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69, 136-145.*

Carmines, E.G. and McIver, J.P. (1981) Analysing models with unobservable variables. In G.W. Bohrnstedt and E.F. Borgatta (eds) *Social Measurement: Current Issues* (pp. 65-115). Beverly Hills, CA: Sage.

Carr, M. And Kurtz-Costes B.E. (1994) Is being smart everything? The influence of student achievement on teachers' perceptions. *British Journal of Educational Psychology Jun; 64 (Pt 2):263-76.*

Carsrud, A.L. (1986) Effects of structured social interaction with geriatric mentally retarded clients: An exploratory study. *Journal of Psychology and Aging. Vol.1 No.1 78-79*

Chapman, J. (1988) Cognitive-motivational characteristics and academic achievement of learning disabled children: a longitudinal study, *Journal of Educational Psychology, 80, 357-365.*

Chen, P.P. (2006) Relationship between students' self-assessment of their capabilities and their teachers' judgements of the students' capabilities in mathematics problem-solving. *Psychological Review June; 98(3): 765-778.*

Choi, N. (2005) Self efficacy and self-concept as predictors of college students' academic performance. *Psychology in schools, Vol. 42(2), 197-205.*

Coe, R. (2000) *What is an effect size? A guide for users.* Accessed online at http://www.cemcentre.org/File/CEM%20Extra/EBE/ESguide.pdf on 21/07/06.

Coffield, F., Moseley, D., Hall, E. and Ecclestone K (2004) *Should we be using learning styles? What research has to say to practice.* The Learning and Skills Research Centre.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences (2nd edition).* Hillside, NJ: Lawrence Erlbaum associates, Inc.

Coleman, J. M., & Fults, B. A. (1985). Special class placement, level of intelligence, and the self-concept of gifted children: A social comparison perspective. *Remedial and Special Education,* 6, 7–11.

Crane, R. M. and Bracken, B. A. (1994) Age, race and gender differences in child and adolescent self-concept: Evidence from behavioural-acquisition, context dependent model. *School Psychological Review, 23(3), 496-511.*

Crawley, F.E. and Koballa, T.R. (1994) Attitude research in science education: Contemporary models and methods. *Science Education, 78, 35-56.*

Crocker, J. and Wolfe, C.T. (2001) Contingencies of self-worth. *Psychological Review, vol. 108, no. 3, 593-623.*

Crocker, J. and Knight, K.M. (2005) Contingencies of Self-Worth. *Current directions in psychological science, Volume 14—Number, 200-203.*

Cronbach, L. J. and Meehl, P. E. (1955) Construct validation in psychological tests. *Psychological Bulletin, 52(4), 281 - 302*

Cudeck, R. and Henley, S.J. (1991) Model selection in covariance structure analysis and the 'problem' of sample size: A clarification. *Psychological Bulletin, 109: 512-519.*

Curran, P.J., West, S.G. and Finch, J.F. (1996) The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1 16-29.*

Cuttance, P. and Ecob, R. (1987) *Structural modeling by example: Applications in education, sociology and behavioural research.* Cambridge university press.

Dai, D. Y. (2004). How universal is the big-fish–little-pond effect? *American Psychologist, 59,*267–268.

Damon, W. (1995) *Greater Expectations.* New York, Free Press.

Deci, E. L. and Ryan, R.M. (1985) *Intrinsic motivation and self-determination in human behaviour*: New York: Plenum Press.

Denis, D.J. and Legerski, J. (2006) *Theory & Science: Causal Modeling and the Origins of Path Analysis.* University of Montana, ISSN: 1527-5558. Accessed online at http://theoryandscience.icaap.org/content/vol7.1/denis.html on 7/8/2006.

DfES (2004) Key Stage 3 National Strategy 2004–05: Raising standards and supporting whole-school improvement. Date of issue: 03-2004 Ref: DfES 0124-2004 G.

DfES (2006) Science National Curriculum for England and Wales. Assessed online on 3/09/06 at http://www.nc.uk.net/download/alldownloads.html

DES (1988) Science in the national curriculum. HMSO.

NCC (1989) *Science: Non-statutory guidance.* The National Curriculum Council

Diamantopoulos, A. and Siguaw, J.A. (2000) *Introducing LISREL.* Sage Publications, London.

DiStefano, C. (2006) Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal, vol 13, no.3, 440-464.*

Doran, R., Lawrenz, F., & Helgeson, S. (1994). Research on assessment in science. In Gabel, D. (Ed.), *Handbook of research on science teaching and learning* (pp. 388–442). New York: Macmillan.

DuBois, D.L., Felner, R.D., Brand, S., Phillips, R.S.C. and Lease, A.M. (1996) Early adolescent self-esteem: A developmental-ecological framework and assessment strategy. *Journal of research on Adolescence, 6, 543-579.*

Duncan, O. D., & Hodge, R. W. (1963). Education and occupational mobility: A regression analysis. *The American Journal of Sociology, 68, 629-644.*

Eapen, V. and Abbas, N. (2000) Cross-cultural validation of Harter's Self-Perception Profile for Children in the United Arab Emirates. *Annals of Saudi Medicine, Vol 20, No.1, 8-11.*

Edwards, A. (1953) *Techniques of attitude scale construction.* New York: Appleton-Century-Cross.

Elliott, J., Hufton, N., Illushin, L. & Lauchlan, F. (2001) Motivation in the junior years: International perspectives on children's attitudes, expectation s and behaviour and their relationship to educational achievement , *Oxford Review of Education*, 27, pp. 37–68.

Ellis, L. A., Marsh, H. W. and Richards, G .E. (2002) A Brief Version of the Self Description Questionnaire II. *Proceedings of the 2nd International Biennial Conference, University of Western Sydney.*

EMBO Reports (2005) Aptitude or attitude? *European Molecular Biology Organisation (EMBO) reports vol. 6, No.5.*

Fetsch, R.J. and Yang, R.K. (2002) The Effects of competitive and cooperative learning preferences on children's self-perceptions: A comparison of 4-H and non-4-H members. Journal of extension, Vol 40 No. 3. Accessed online at http://www.joe.org/joe/2002june/a5.html. on 16/6/2006.

Keith, L. K. and Bracken, B. A (1996) *Self-concept instrumentation: A historical and evaluative review.* In Bracken (Ed.), Handbook of self-concept: Development, social and clinical considerations, 91-170, New York: Wiley.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7, 117-140.*

Finney, S.J. and Schraw, G. (2003) Self-efficacy beliefs in College statistics courses. *Contempory Educational Psychology, 28, 161-186.*

Fornell, C., and Larcker, D.F. (1981) Structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18: 39-50. Cited in Diamantopoulos, A. and Siguaw, J.A. (2000) *Introducing LISREL.* Sage Publications, London.

Gardner, P.L. (1975) Attitudes to science: a review. *Studies in Science Education, 2, 1-41.*

Garson, D. (2006) Structural Equation Modeling. Accessed online at http://www2.chass.ncsu.edu/garson/pa765/structur.htm on 23 July 2006.

Gilbert, J.K. (2006) On the nature of "context" in chemical education. *International Journal of Science Education Vol. 28, No. 9, 957-976.*

Gonzalez, E.J. and Miles, J.A. (Eds.) (2001) *User guide for the TIMSS 1999 international database.* Boston: International Study Centre.

Guay, F., Marsh, H.W. and Boivin, M. (2003) Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology. 95 (1): 124-136.*

Hamachek, D. (1995) *Psychology in teaching, learning and growth (5th edition).* Sydney: Allan and Bacon.

Hansford, B.C., and Hattie, J.A. (1982) The relationship between self and achievement/performance measures. *Review of Educational Research, 52, 123-142.*

Harter, S. (1982) The perceived competence scale for children. *Child Development, 53, 87-97.*

Harter, S (1985) *Manual for the Self Perception Profile for Children.* Denver, University of Denver

Harter, S. (1988). *The Self-Perception Profile for Adolescents.* Unpublished manuscript, University of Denver.

Harter, S (1990) Issues in the assessment of self-concept of children and adolescents La Greca A (ed.) *Through the Eyes of the Child. Obtaining Self-Reports from Children and Adolescents*, USA, Allyn and Bacon.

Harter, S.and Pike, R. (1984) The pictorial scale of perceived self-confidence and social acceptance for young children. *Child development Vol. 55 No. 6 1969-1982*

Haussler, P. and Hoffmann, L (2002) An Intervention Study to Enhance Girls' Interest, Self-Concept, and Achievement in Physics Classes. *Journal of Research in Science Teaching, Vol. 39, No.9, 870-888.*

Hattie, J. (1992) *Self-Concept.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hay, I., Ashman, A., and Van Kraayenoord, C. (1997). Investigating the influence of achievement on self-concept using an intra-class design and a comparison of the PASS and SDQ-I self-concept tests. *British Journal of Educational Psychology, 67, 311-321.*

Hayduk, L.A. (1987) Structural equation modeling with LISREL: Essentials and advances. Baltimore, MD: John Hopkins University Press.

Hu, L. and Bentler, P.M. (1995) Evaluating model fit. In Hoyle, R.H. (Ed.) *Structural equation modelling: concepts issues and applications,* (pp 158-176). Thousand Oaks, CA:Sage.

Hufton, N.R., Elliott, J.G. and Illushin, L. (2002) Educational Motivation and Engagement:
qualitative accounts from three countries. *British Educational Research Journal, Vol. 28, No. 2, pp 256 – 289.*

Institute of Physics (2006) Evidence from the Institute of Physics for House of Lords Select Committee on Science & Technology on Science Teaching in Schools.

Ireson, J., Hallam, S and Plewis, I. (2001) Ability Grouping in Secondary Schools: Effects on Pupils Self-Concepts *British Journal of Educational Psychology 71, 315-326.*

Jaccard, J. and Wan, C.K. (1996) *LISREL approaches to interaction effects in multiple regression.* Thousand Oaks, CA: Sage Publications.

James, W. (1963) The principles of psychology. New York: Holt, Rhinehart and Winston. (Original work published 1890).

Jerusalem, M. (1984). Reference group, learning environment and selfevaluations: A dynamic multi-level analysis with latent variables. In R. Schwarzer (Ed.), *The self in anxiety, stress and depression* (pp. 61–73). Amsterdam: Elsevier Science.

Jöreskog, K.G. and Sörbom, D. (1993) *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Kline, R.B. (1998) *Principles and practices in structural equation modelling.* Guilford Press: London.

Lau, I.C-y., Yeung, A.S., Jin, P. and Low, R. (1999) Toward a hierarchical, multidimensional English self-concept. *Journal of Educational Psychology, vol. 91, No.4, 747-755.*

Lawrence, D. (1996) *Enhancing Self-Esteem in the Classroom,* London, Paul Chapman.

Leach, J. and Scott, P. (2003). *Individual and sociocultural views of learning in science education.* Science and Education, 12(1), 91-113

Lodewyk, K.R. and Winne, P.H. (2005) Relations among the structure of learning tasks, achievement and changes in self-efficacy in secondary students. *Journal of Educational Psychology, Vol.97, No.1, 3-12.*

Long, J.S. (1983) *Covariance Structure Models: An introduction to LISREL.* Beverly Hills, CA: Sage.

Ludtke, O. Koller, O, Marsh, H.W., Trautwein, U.T. (2005) Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology 30, 263-285.*

Pajares, F. (1996) Self efficacy beliefs in academic settings. *Review of Educational Research, 66, 543-578.*

Pajares, F., Britner, S. L., Valiante, G. (2000) Relation between Achievement Goals and Self-beliefs of Middle School Students in Writing and Science, *Contemporary Educational Psychology 25, 406-422.*

Papanastasiou, E.C. (2004) Differential effects of science attitudes and science achievement in Australia, Cyprus, and the USA. *International Journal of Science Education, vol 26, No.3 259-280.*

Pietsch, J., Walker, R. and Chapman, E. (2003) The relationship among self-concept, self-efficacy, and school performance in mathematics during secondary school. *Journal of educational psychology, Vol. 95, No.3, 589-603.*

MacCallum, R. (1995) Model specification: procedures, strategies and related issues. In Hoyle, R.H. (Ed.) *Structural equation modelling: concepts issues and applications,* (pp 16-39). Thousand Oaks, CA:Sage.

MacCallum, R., Browne, M.W. and Sugawara, H.M. (1996) Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods 1, 130-149.*

Machover, K. (1949) *Personality projection in the drawing of a human figure: A method of personality investigation.* Springfield IL: Charles C. Thomas.

Marín, G. Gamba, R.J. and Marín, B.V. (1992) Extreme response style and acquiescence among hispanics. *Journal of Cross-Cultural Psychology, 23, 498-509.*

Markus, H (1977) Self-schemata and processing information on the self. *Personality and Social Psychology. 35(2), 63-78.*

Markus, H. and Nurius, P. (1986) Possible Selves. *American Psychologist. 41, 9, 954-969.*

Markus, H. and Wulf, E. (1987) The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology, 38, 299-337.*

Marsh, H.W. (1985) Age and sex effects in multiple dimensions of preadolescent self-concept: A replication and extension. *Australia Journal of Psychology, 37,(2), 197-204.*

Marsh, H.W. (1986) Verbal and maths self-concepts: an internal/external frame of reference model, American *Educational Research Journal, 23, 129-149.*

Marsh, H. W. (1987). The big fish little pond effect on academic self-concept. *Journal of Educational Psychology,79*, 280–295.

Marsh, H.W. (1989) Sex differences in the development of verbal and mathematics constructs: The high school and beyond. *American Educational Research Journal, 26, 191-225.*

Marsh, H.W. (1990) *Self-Description Questionnaire 1*, Sydney, University of Western Sydney

Marsh, H.W (1990a) A multidimensional, hierarchical self-concept theoretical and empirical justification, *Educational Psychology Review, 2, 77-172.*

Marsh, H.W. (1990b) Confirmatory factor analysis of multitrait-multimethod data: The construct validation of multidimensional self-concept responses. *Journal of Personality, 58(4) 661-692.*

Marsh, H.W. (1990c) The structure of academic self-concept. The Marsh/Shavelson Model. *Journal of Educational Psychology, 82, 623-636.*

Marsh, H.W. (1991). The failure of high ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations. *American Educational Research Journal, 28,* 445–480.

Marsh, H.W. (1992a) *Self-Description Questionnaire (SDQ-I): A theoretical and empirical basis for the measurement of multiple dimensions of pre-adolescent self-concept: A test manual for research monograph.* Macarthur, NSW Australia: Faculty of Education, University of Western Sydney.

Marsh, H.W. (1992b) *Self-Description Questionnaire (SDQ-II): A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept: An interim test manual for research monograph.* Macarthur, NSW Australia: Faculty of Education, University of Western Sydney.

Marsh, H. W. (1992c) *Self-Description Questionnaire (SDQ-III): A theoretical and empirical basis for the measurement of multiple dimensions of late adolescent self-concept: An interim test manual for research monograph.* Macarthur, NSW Australia: Faculty of Education, University of Western Sydney.

Marsh (1992d) The content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology, 82 646-656.*

Marsh, H.W. (1993a) Academic self-concept: Theory measurement research. In J. Suls (Ed.), *Psychology perspectives on the self* (Vol.4, p.59-98). Hillsdale, NJ: Erlbaum.

Marsh, H.W. (1993b) The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal, Winter 1993, Vol. 30, No. 4, pp. 841-860.*

Marsh, H. W. (1999). *Academic Self Description Questionnaire – I: ASDQ I,* University of Western Sydney, Self-concept Enhancement and Learning Facilitation Research Centre, Macarthur, Australia.

Marsh, H.W., Barnes, J., Cairns, L. and Tidman, M (1984) The self-description questionnaire (SDQ): Age and sex effects in the structure and level of self-concept for preadolescent children. *Journal of Educational Psychology, 76, 940-956.*

Marsh, H.W., Byrne, B. and Shavelson, R. (1988) A multi-faceted academic self-concept its hierarchical structure and its relation to academic achievement, *Journal of Educational Psychology 80, 366-380*

Marsh, H.W., Byrne, B.M. and Yeung, A.S. (1999) Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist, 34, 154-157.*

Marsh, H. W., Chessor, D., Craven, R. G., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Education Research Journal, 32,* 285–319.

Marsh, H.W. and Craven, R.G. (2005) A reciprocal effects model of the causal ordering of self-concept and achievement in Marsh, H.W., Craven R.G. and McInerney, D.M. (Eds) New Frontiers for self research, (pp 17-51). International Age Publishing Inc, Connecticut.

Marsh, Craven and McInerney (2005) Overview: New Frontiers for self research in Marsh, H.W., Craven R.G. and McInerney, D.M. (Eds) New Frontiers for self research, (pp 3-13). International Age Publishing Inc, Connecticut.

Marsh, H.W. and Gouvernet, P.J. (1989) Multidimensional self-concepts and perceptions of control: Construct validation of responses of children. *Journal of Educational Psychology, 81(1), 57-69.*

Marsh, H.W. and Hattie, J. (1996) Theoretical perspectives on the structure of self-concept. In B.A. Bracken (Ed.), *Handbook of self-concept* (pp. 38-90). New York: Wiley.

Marsh, H. W. and Hau, K.T (2002) Internal/external frame of reference model. *SELF Research Centre: Proceedings of the 2nd International Biennial Conference.*

Marsh, H. W. and Hau, K.T. (2003) Big-fish– little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376.

Marsh, H.W. and Hau, K.T. (2004) The Big-Fish–Little-Pond Effect Stands Up to Scrutiny. *American Psychologist, May-June 2004.*

Marsh, H.W. and Hocevar, D. (1985) Application of confirmatory factor analysis to the study of self-concept: First – and higher-order factor models and their invariance across groups. *Psychological Bulletin, 97, 562-582.*

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish–little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong high schools. *Journal of Personality and Social Psychology, 78,* 337–349.

Marsh, H. W., Koller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big fish little pond effect on academic self-concept. *American Educational Research Journal, 38, 321–350.*

Marsh, H.W. and McDonald Holmes, I.W. (1990) Multidimensional self-concepts: Construct validation of responses of children. *American Educational Research Journal, 27 (1), 87-117.*

Marsh, H.W. and O'Neil, R. (1984) Self-description questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21, 153-174.*

Marsh, H.W., Parker, J. and Smith, I.D. (1983) Preadolescent self-concept: Its relation to self-concept as inferred by teachers and to academic ability. *British Journal of Educational Psychology, 53, 60-78.*

Marsh, H. W., and Parker, J. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology, 47,* 213–231.

Marsh, H.W., Relich, J.D. and Smith, I.D. (1983) Self-concept: The construct validity of interpretations based upon the SDQ. *Journal of personality and social psychology, 16, 270-305.*

Marsh, H.W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept— An application of multilevel modeling. *Australian Journal of Education, 40,* 65–87.

Marsh, H.W. and Shavelson, R.J. (1985) Self-concept: Its multifaceted and hierarchical structure. *Educational Psychologist, 20, 107-125.*

Marsh, H.W., Smith, I.D. and Barnes, J. (1985) Multidimensional self-concept: Relations with sex and academic achievement. *Journal of Educational Psychology, 77, 581-596.*

Marsh, H.W., Trautwein, U., Ludtke, O., Koller, O., and Baumert, J. (2005) Academic self-concept, interests grades and standardized test scores: Reciprocal effects models of causal ordering. Sydney: SELF Research Centre, University of Western Sydney.

Marsh, H.W., Walker, R. and Debus, R. (1991) Subject-specific components of academic self-concept and self-efficacy. *Contemporary Educational Psychology 16, 2331-345.*

Marsh Dowson Pietsch and Walker (2004) Why Multicollinearity Matters. *Journal of Educational Psychology*

Mattern. N. and Schau, C. (2002) Gender differences in science attitude-achievement relationship over time among white middle-school students. *Journal of Research in Science Teaching, vol. 39, No. 4 324-340.*

Miller, P.H., Blessing, J.S. and Schwartz, S. (2006) Gender differences in high school students' view about science. *International Journal of Science Education Vol. 28, No. 4, 363-381.*

Mimi Bong and Einar M. Skaalvik *(2003)* Academic Self-Concept and Self-Efficacy: How Different Are They Really? *Educational Psychology Review, Vol. 15, No. 1, March 2003*

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (1999). The Adelaide Declaration on National Goals for Schooling in the Twenty first Century.

Mitchell, J., Trent, R. and McArthur, R. (1993) *Human Figure Drawing Test.* Los Angeles: Western Psychological Services

Moore, R., & Foy, R. (1997). The Scientific Attitude Inventory: A revision (SAI II). *Journal of Research in Science Teaching,* Vol.34, No.4, 327–336.

Moore, R., & Sutman, F. (1970). The development, field test and validation of an inventory of scientific attitudes. *Journal of Research in Science Teaching, 7,* 85–94.

Motl, R.W. and DiStefano, C. (2002) Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling: A Multidisciplinary Journal, vol 9, no.4, 562-578.*

Mulaik, S.A., James, L.R., Van Altine, J., Bennett, N., Lind, S. and Stilwell, C.D.. (1989) Evaluation of goodness to fit indices for structural equation models. *Psychological Bulletin, 105, 430-445.*

Munby, H. (1983a). *An investigation into the measurement of attitudes in science education.* Columbus, OH: ERIC Science, Mathematics and Environmental Education Clearinghouse, Center for Science and Mathematics Education, Ohio State University.

Munby, H. (1983b). Thirty studies involving the Scientific Attitude Inventory: What confidence can we have in this instrument? *Journal of Research in Science Teaching, 20,* 141–162.

Munby, H. (1997) Issues of validity in science attitude measurement. *Journal of Research in Science Teaching, vol. 34, No. 4 337-341.*

Murphy, C., Ambusaidi, A and Beggs, J. (2006) Middle East meets West: Comparing children's attitudes to school science. *International Journal of Science Education Vol. 28, No. 4, 405-422.*

Murphy, P. and Whitelegg, E. (2006) *Girls in the Physics Classroom: A review of the research evidence on the participation of girls in physics.* Institute of Physics Report, June 2006.

Naglieri, J. A., McNeish, T. J., & Bardos, A. N. (1991). *Draw A Person: Screening Procedure for Emotional Disturbance.* Austin, TX: Pro-Ed.

Naglieri, J. A.,& Pfeiffer, S. I. (1992). Validity of the Draw A Person: Screening Procedure For Emotional Disturbance with a socially emotionally disturbed sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 4,* 156-159.

NfER. 2004. *Qualitative Study of the Early Impact of On-Track.*

Nunnaly, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

OCR (2006) Twenty First Century Science GCSE, Exam Specification. Accessed online 4/09/06 at http://www.gcse-science.com/teachers_subpage.php?pg_id=282

Osborne, J. and Collins, S. (2000) Pupils' and Parents' Views of the School Science Curriculum: A study funded by the Wellcome Trust. Pub: King's College, London.

Pajares, F. (1996) Self-efficacy beliefs in educational settings. *Review of educational research, 66, 543-578.*

Pajares, F. (2002) *Overview of Social Cognitive Theory and of Self-Efficacy*: Retrieved 7 July 2006, from http://www.emory.edu/EDUCATION/mfp/eff.html

Pajares, F. and Miller, M.D. (1994) Role of Self-Efficacy and Self-Concept beliefs in Mathematical problem solving: A path Analysis. *Journal of Educational Psychology, 86, 193-203.*

Pajares, F., Britna, S.L. and Valiante, G. (2000) Relation between achievement goals and self-beliefs of middle school students in writing and science. *Contemporary Educational Psychology 25, 406-422.*

Pajares, F. and Schunk, D. H. (2002) Self and Self Being in psychology and Education. In J. Aronson and D Cordova (Eds.), Improving academic achievement: Impact of psychological factors on education (pp3-21). New York: Academic Press

Piers, E.V. (1984) *Piers-Harris Children's Self-Concept Scale: Revised manual.* Los Angeles: Western Psychological services.

Pietsch, J., Walker, W. and Chapman, E. (2003) The Relationship among Self-Concept, Self-Efficacy, and Performance in Mathematics During Secondary School. *Journal of Educational Psychology.* Vol 95, No. 3, 589-603.

Pilot, A and Bulte, A.M.W. (2006) Why do you "need to know"? Context-based education. *International Journal of Science Education Vol. 28, No. 9, 953-956.*

Popper, K. (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge, London.

Rani, G. (2006) A cross-domain analysis of change in students' attitudes toward science and attitudes about the utility of science. *International Journal of Science Education Vol. 28, No.6, 571-589.*

Reynolds, A.J. and Walberg, H.J. (1992) A structural model of science achievement and attitude: An extension to high school. *Journal of Educational Psychology, Vol. 84, No.3, 371-382.*

Rosenberg, M. (1979) *Conceiving the Self.* New York. Basic books.

Rosenberg, M. (1989) *Society and the Adolescent Self-Image.* Revised edition. Middletown, CT: Wesleyan University Press.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans & D. S. G. Pollock & A. Satorra (Eds.), *Innovations in Multivariate Statistical Analysis* (pp. 233-247). Dordrecht: Kluwer Academic Publishers.

Satorra, A. and Bentler, P.M. (1994) Correction to test statistic and standard errors in covariance structure analyses. In A. Von Eye and C.C. Clogg (Eds), *Analysis of latent variables in developmental research* (pp. 399-419). Newbury Park, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66, 507-514.*

Schwartz, A.T. (2006) Contextualized Chemistry Education: The American experience. *International Journal of Science Education Vol. 28, No. 9, 977–998*

Schwarzer, R., Lange, B., & Jerusalem, M. (1982). Selbstkonzeptentwicklung nach einem Bezugsgruppenwechsel [Self-concept development after a reference-group change]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 14*, 125–140. Reported in Ludtke et al. (2005) Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology 30, 263-285.*

Shavelson, R.J. and Bolus, R. (1982) Self-concept: The interplay of theory and methods. *Journal of Educational Psychology, 74, 3-17.*

Shavelson, R.J., Hubner, J.J. and Stanton, G.C. (1976) Validation of construct interpretations, *Review of Educational Research. 46, 407-441.*

Shavelson, R.J. and Marsh, H.W. (1986) On the structure of self-concept. In R. Schwarzer (Ed.) *Anxiety and Cognitions* (p.303 – 330). New York: Erlbaurm.

Shrauger, J.S and Schoeneman, T.J. (1979) Symbolic interactionist view of self-concept; Through the looking glass darkly. *Psychological Bulletin, 86, 549-573.*

Skaalvik, E.M. (1997a). Issues in research on self-concept. In: Maehr, M., and Pintrich, P. R. (eds.), *Advances in Motivation and Achievement* (*Vol. 10*), JAI Press, New York, pp. 51–97.

Skaalvik, E.M., and Rankin, R.J. (1990) Math, verbal and general academic self-concept: The internal/external frame of reference model and gender differences in self-concept structure. *Journal of Educational Psychology, 82, 546-554.*

Sobel, M.F. and Bohrnstedt, G.W. (1995) The use of null models in evaluating the fit of covariance structure models. In Diamantopoulos, A. and Siguaw, J.A. (2000) *Introducing LISREL*. Sage Publications, London.

Software Review (1998). *Journal of Psychoeducational Assessment (16): 434-364.*

Song, I.S. and Hattie, J.A (1984). Home environment, self-concept, and academic achievement: A causal modelling approach. *Journal of Educational Psychology, 76, 1236-1281*

Soares, A.T. and Soares L.M (1979) *The Affective Perception Inventory – Advanced Level.* Trumbell, CT: ALSO.

Smith, I.D. and Marsh, H.W. (1985) A cross-cultural study of self-concept in primary school children. *Collected papers of the 1985 Annual Conference of the Australian Association for Research in Education, (350- 353).*

Smithers, A. and Robinson, P. (2005) *Physics in schools and colleges: Teacher deployment and student outcomes.* ISBN 1 90 1351 74 2. Carmichael Press, University of Buckingham

Smithers, A. and Robinson, P. (2006) *Physics in schools and universities II. Patterns and policies.* ISBN 1 90 1351 85 8. Carmichael Press, University of Buckingham

Stevenson, H.W. & Lee, S. (1990) A study of American, Chinese and Japanese Children, *Monographs of the Society for Research in Child Development*, No. 221, 55 (whole issue).

Stevenson, H.W. & Stigler, J. W. (1992) *The Learning Gap: why our schools are failing and what we can learn from Japanese and Chinese education.* New York, Summit Books.

Stephenson, W. (1953) *The study of behaviour – Q-technique and its methodology.* Chicago: University of Chicago Press.

Swann, W.B. (1997) The trouble with change: Self verification and allegiance to the self. *Psychological Science, 8, 177-180.*

Tracey, D. K., Marsh, H. W., & Craven, R. G. (2003). Self-concepts of preadolescent students with mild intellectual disabilities: Issues of measurement and educational placement. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self research* (Vol. 1, pp. 203–230). Greenwich, CT: Information Age.

Trautwein, U., Ludke, O., Marsh, H.W., Koller, O. and Braumert, J. (2004) Tracking, grading and student motivation: How reference groups impact on self-concept and interest I, ninth grade mathematics. Manuscript submitted for publication. Cited in Trautwein et al, 2005.

Trautwein, U., Koller, O., Ludtke, O and Braumert, J. (2005) Student tracking and the powerful effects of opt-in courses on self-concept: Reflected glory effects do exist after all. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.) *International advances in self research: New frontiers for self research* (Vol. 2, pp. 307-328). Greenwich, CT: Information Age.

Tsaia, C-C. and Liub, S-Y. (2006) Developing a Multi-dimensional Instrument for Assessing Students' Epistemological Views toward Science. *International Journal of Science Education Vol. 27, No. 13, 1621–1638.*

UCAS (2006) University and Colleges Admissions Service: Statistical enquiry. Accessed online at: http://www.ucas.ac.uk/figures/enq/index.html, July 2006.

Valentine, J.C., DuBois, D.L. and Cooper, H. (2004) The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39(2), 111-133.*

Van Langena, A., Rekers-Mombargb, L. and Dekkersa, H. (2006) Sex-related Differences in the Determinants and Process of Science and Mathematics Choice in Pre-university Education. International Journal of Science Education, vol. 28, no. 1, 71–94

Van Dongen-Melman, J.E.W.M., Koot, H.M. and Verhulst, F.C. (1993) Cross-cultural validation of Harter's self-perception profile for children in a Dutch sample. *Educational and Psychological Measurement, 53, 739-753.*

Veenman, M.V.J., Prins, F.J. and Verheij, J.(2003) Learning styles: Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology, 73, 357–372*

Vispoel, W.P. (1995) Self-concept in artistic domains: an extension of the Shavelson, Hubner and Stanton (1976) model. *Journal of Educational Psychology, Vol.87, No.1, 134-153.*

Vispoel, W.P. (2003) *Measuring and understanding self-perceptions of musical ability.* In, Marsh, H.W., Craven R.G. and McInerney, D.M. (Eds) International Advances in self research, (pp 151-179). International Age Publishing Inc, Connecticut.

Yeung, A. S., Chui, H-S and Lau I. C-Y (1999) Hierarchical and Multidimensional Academic Self-Concept of Commercial Students, *Contemporary Educational Psychology 24, 376-389.*

Watkins, D and Akande, A. (1992) The internal structure of the self-description questionnaire: A Nigerian investigation. *British journal of Educational Psychology, 62, 120-125.*

Watkins, D., Fleming, J.S. and Alfon, M.C. (1989) *A test of Shavelson's hierarchical multifaceted self-concept model in Filipino college sample. International Journal of Psychology, 24, 367-379.*

Watkins, D and Gutierrez, M (1989) The structure of self concept: Some Filipino evidence. *Australian Psychologist, 24, 401-410.*

Waugh, R.F. (2001). Measuring ideal and real self-concept on the same scale, based on multifaceted, hierarchical model of self-concept. *Educational and Psychological Measurement, 61, 85-101.*

Wells, L and Marwell, G. (1976) *Self-esteem: Its conceptualization and measurement.* London: Sage.

West, S.G., Finch, J.F. and Curran, P.J. (1996) Structural equation models with non-normal variables: Problems and remedies. In Hoyle, R.H. (Ed.) *Structural equation modelling: concepts issues and applications,* (pp 56-75). Thousand Oaks, CA:Sage.

West, S.G., Fish, A.J. and Stevens, R.J. (1980) General self-concept, self concept of academic ability and school achievement: Implications for "causes" of self-concept. *Australian Journal of Education, 24, 194-213.*

Wheaton, B. (1987) Assessment fit in overidentified models with latent variables. Sociological Methods and Research, 16 118-154.

Wheaton, B., Muthen, B., Alwin, D.F. and Summers, G.F. (1977) Assessing reliability and stability of panel models. In D.R. Heise (ed.) *Sociological methodology 1977* (pp. 84 – 136). San Francisco: Jossey-Bass.

Zohar, A. and Bronshtein, B. (2005) Physics teachers' knowledge and beliefs regarding girls'low participation rates in advanced physics classes. *International Journal of Science Education, vol. 27, no. 1, 61–77.*

Zeidner M., & Schleyer, E. J. (1999). The big-fish–little-pond effect for academic self-concept, test anxiety and school grades in gifted children. *Contemporary Educational Psychology, 24,* 305–329.

Zimmerman, B. J. (1996). *Misconceptions, Problems, and Dimensions in Measuring Self-Efficacy*, Paper presented at the annual meeting of the American Educational Research Association, New York.

APPENDIX ONE

Faculty of Education
The University of Manchester, Humanities Building, Oxford Road, Manchester M13 9PL

Learning, Teaching and Assessment

Telephone 0161-275 3200/8472 Fax 07092 221966
email graham.hardy@man.ac.uk

Graham Hardy PGCE Science Education Tutor

THE UNIVERSITY of MANCHESTER

Ref: Pupils' academic self-concept in science

Dear Parent,

The School of Education at the University of Manchester is carrying out a piece of research into the self-esteem of pupils when learning science. XXXXXXXXXXX School has agreed to be one of the schools taking part.

The aim of the research is to develop an instrument that measures academic self-concept (self-esteem) in science. The devised self-concept instrument will be helpful for teachers as it will provide a way to measure how pupils feel about themselves in their learning of science related subjects. The information will be gathered from the pupils by means of a questionnaire which includes questions such as: 'I am good at planning how to do an experiment'. The questionnaire should take about 20 to 25 minutes to complete and we will be asking all pupils in your child's class to complete it in the next few days.

The results will be treated anonymously and all information will be treated in the strictest of confidence. I hope you feel that you are able to allow your child to take part in this exercise. Should you NOT wish your child to complete the questionnaire please tick (√) the box, sign below and return to school. If you have no objection, you do not need to return the form.

If you require further information you can speak directly to Graham Hardy at the university on 0161 275 3200.

Many thanks,

Graham Hardy

School of Education, University of Manchester

☐ I DO NOT AGREE for my child to fill in the self-concept questionnaire

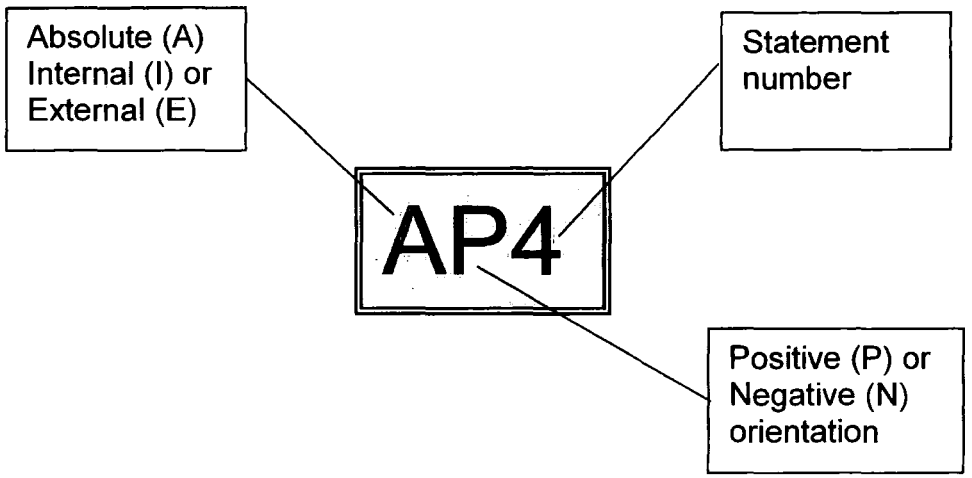
Name of the child:.....

Parent's/carer's signature: .....

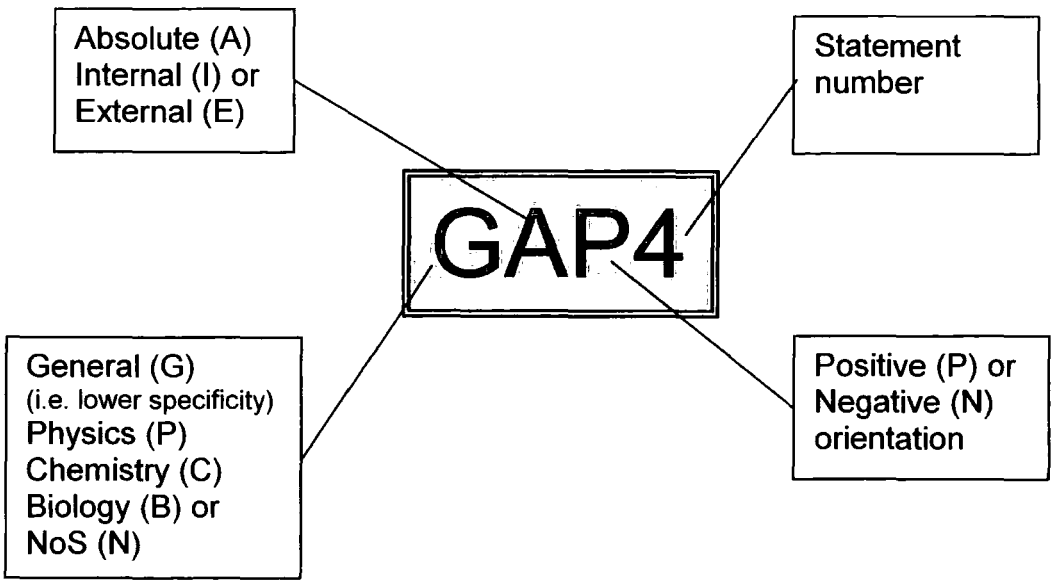
APPENDIX TWO

Understanding the Coding of Statements

Pre-pilot Statements



Final Questionnaire Statements



APPENDIX THREE

Construction of Item-Pairs

The CFA and SEM analysis was carried out on ITEM-PAIRS. The item-pairs were constructed from the combination of the following individual items.

SCI1 = GAP5, GAN6
SCI2 = GIP5, GIN1
SCI3 = GEP7, GEN5
ENQ1 = GAP3, GAN2
ENQ2 = GIP2, GIN1
ENQ3 = GEP2, GEN7
ENQ4 = EAP1, EAN2
ENQ5 = EAP2, EAN6
ENQ6 = EAP4, EAN1
BIOL1 = GAP2, GAN5
BIOL2 = BAP5, BAN5
BIOL3 = BAP2, BAN2
BIOL4 = BAP3, BAN3
PHYS1 = GAP1, GAN3
PHYS2 = PAP1, PAN1
PHYS3 = PAP2, PAN2
PHYS4 = PAP3, PAN3
CHEM1 = GAP4, GAN1
CHEM2 = CAP1, CAN6
CHEM3 = CAP4, CAN1
CHEM4 = CAP5, CAN5
NoS1 = NAP3, NAN3
NoS2 = NAP4, NAN5
NoS3 = NAP5, NAN6
IoS1 = IOSP1, IOSP2
IoS2 = IOSP3, IOSP4

APPENDIX FOUR

List of Abreviations

Abbreviation	Explanation
ACH	Academic Achievement
AGFI	Adjusted Goodness of fit index
AIC	Akaike's Information Index
ASC	Academic Self-Concept
ASD	Items beginning ASD are the science statements from Marsh's (1990b) ASDQ II instrument
ASDQ	Academic Self Description Questionnaire (Marsh, 1999)
BFLPE	Big Fish Little Pond Effect (Marsh, 1987)
BRGE	Basking in Reflected Glory Effect (Trautwein et al., 2005)
CAIC	Consistent Akaike's Information Index
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
DAP	Draw a Person test (Machover, 1949)
DWLS	Diagonally Weighted Least Squares
ECVI	Expected Cross-Validation Index
EFA	Exploratory Factor Analysis
EPC	Expected Parameter Change
GenSci	An inferred latent variable constructed from the latent variables of
GFI	Goodness of Fit index
GlobalSc	A latent variable constructed from the directly observable
GLS	Generated Least Squares
GSM	Items beginning GSM are general school statements from Marsh's (1990b) ASDQ II instrument
HFDT	Human Figure Drawing Test (Mitchell et al., 1993)
HTP	House Tree Person test (Buck, 1948)
I/E Model	Internal/External frame of reference model (Marsh, 1986)
IoP	Institute of Physics
IoS	Importance of Science items
KS3	Key Stage 3
KS4	Key Stage 4
LISREL	Linear Structural Relations (Jöreskog and Sörbom, 1993)
MSCS	Multidimensional Self-Concept Scale (Bracken, 1992)
NNFI	Non-Normed Fit Index (also known as TLI)
NoS	Nature of Science
PCSC	Perceived Confidence Scale for Children (Harter, 1982)
PGFI	Parsimony Goodness to Fit index
PIRLS	Progress in International Reading and Literacy Study
PISA	Program for International Study
PNFI	Parsimony Normed Fit Index
RMR	Root Mean Square Residual
RMSEA	Root Mean Square Error of Approximation
SDQ-I	Self Description Questionnaire 1 (Marsh, 1992a)
SDQ-II	Self Description Questionnaire 2 (Marsh, 1992b)
SDQ-III	Self Description Questionnaire 3 (Marsh, 1992c)
SEM	Structural Equation Modeling
SES	Self-Esteem Scale (Rosenberg, 1989)
SPED	Screening Procedure of Emotional Disturbance (Naglieri et al., 1991)
SPPC	Self-Perception Profile for Children (Harter, 1985)
TIMMS-R	Repeat Third International Mathematics and Science Study
TLI	Tucker Lewis Index (also known as NNFI)
TSLS	Two Stage Least Squares
ULS	Unweighted Least Squares

Science Questionnaire KS 4

Name ..... *Boy or girl* .....

Year Group ..... *Date* .....

Languages Spoken.....

GCSE Science Course(s) .....

THE UNIVERSITY of MANCHESTER

What is this questionnaire about?

This is a questionnaire about what you think about your science lessons. It is not a test. There are no right and wrong answers, just your opinions. It is important, that you try to be as honest as possible when you fill in your answers.

How do you fill it in?

On the next four pages there are lots of statements. There are five answers to each statement and you need to think about the one that fits YOU the best. I will give you some examples of the way I could fill out a questionnaire about things outside of school.

	True of Me				False for Me
I really enjoy watching East Enders on TV	(1)	2	3	4	5
I enjoy eating chocolate more than chips	1	2	(3)	4	5
McDonalds is my favourite meal	1	2	3	(4)	5

I have answered the questions with the small circles. I think East Enders is great and so I have scored it with a (1.) I like chocolate and chips about the same so I have scored it in the middle with a (3.) I don't think McDonalds' food tastes very good, but I don't hate it either, so I have scored it (4.)

		True for Me				False for Me
EAP1	I am good at planning how to do an experiment	1	2	3	4	5
GEP2	I am better at science investigations than my friends	1	2	3	4	5
ASDP6	I have always done well in Science classes	1	2	3	4	5
EAN6	I have difficulty in recording results from experiments	1	2	3	4	5
GEP7	I know more about science than my friends	1	2	3	4	5
ASDP4	I learn things quickly in Science	1	2	3	4	5
EAP4	I have no trouble in thinking up ideas to investigate	1	2	3	4	5
IOSP3	It is important for me to understand the work in Science	1	2	3	4	5
CAN5	I have poor knowledge of atoms, molecules and bonding	1	2	3	4	5
ASDP1	I get good marks in Science classes	1	2	3	4	5
NAP5	I have good knowledge of what scientists do	1	2	3	4	5
IOSP2	Doing well in Science is important for me	1	2	3	4	5
NAN3	I am pretty bad when it comes to knowing how science can help us.....	1	2	3	4	5

		True for Me				False for Me
PAN3	I am pretty bad when it comes to planet Earth and space	1	2	3	4	5
GSMP6	I get good marks in most school subjects	1	2	3	4	5
PAP1	I am good at understanding about electricity and how it is made and used	1	2	3	4	5
PAN2	I usually do poorly at understanding what energy can do and how we use it.	1	2	3	4	5
GSMP3	I have always done well in most school subjects	1	2	3	4	5
GIP2	I am better at science investigations than my other subjects	1	2	3	4	5
CAP1	I am good at understanding about changes in the Earth and its atmosphere.	1	2	3	4	5
PAN1	I have difficulty in understanding about light, waves and sound	1	2	3	4	5
GSMN1	I am hopeless when it comes to most school subjects	1	2	3	4	5
GAP2	I usually do well at biology	1	2	3	4	5
CAP5	I have good knowledge of chemical reactions and equations	1	2	3	4	5
GAP3	I am pretty good when it comes to science investigations	1	2	3	4	5
ASDP7	It is important for me to do well I Science classes	1	2	3	4	5

		True for Me				False for Me
BAN2	I usually do poorly at understanding systems of the human body, e.g. digestion, circulation and respiration	1	2	3	4	5
CAN1	I have difficulty in understanding about rates of reactions	1	2	3	4	5
IOSP1	It is important for me to gain high marks in science tests or exams	1	2	3	4	5
GIN1	Science is my weakest subject	1	2	3	4	5
ASDN3	I am hopeless when it comes to Science	1	2	3	4	5
GAP1	I am good at understanding about physics	1	2	3	4	5
BAP3	I am pretty good when it comes to learning about variation and inheritance ...	1	2	3	4	5
NAP4	I have no trouble in knowing how scientists work	1	2	3	4	5
ASDP2	Work in Science classes is easy for me	1	2	3	4	5
GAN2	I usually do poorly at science investigations	1	2	3	4	5
GAN5	I have poor knowledge of biology	1	2	3	4	5

		True for Me				False for Me
GIN5	I enjoy my other subjects much more than science investigations	1	2	3	4	5
GSMP2	I learn things quickly in most school subjects	1	2	3	4	5
PAP3	I am pretty good when it comes to learning about calculations and formulae ..	1	2	3	4	5
GAN1	I have difficulty in understanding about chemistry	1	2	3	4	5
ASDP5	Compared to others of my age I am good at Science classes	1	2	3	4	5
NAN6	I have difficulty learning about how we get science knowledge	1	2	3	4	5
GAN6	I have difficulty learning new things about science	1	2	3	4	5
GSMP4	Compared to others my age I am good at most school subjects	1	2	3	4	5
GAN3	I am pretty bad when it comes to physics	1	2	3	4	5
BAP2	I usually do well at photosynthesis and the transport of substances in green plants	1	2	3	4	5
GAP4	I have no trouble in learning about chemistry	1	2	3	4	5
GSMP5	Work in most school subjects is easy for me	1	2	3	4	5

		True for Me				False for Me
GAP5	I have good knowledge of science	1	2	3	4	5
EAP2	I usually do well at collecting the data of an experiment	1	2	3	4	5
GIP5	I enjoy science much more than my other subjects	1	2	3	4	5
BAP5	I have good knowledge of evolution	1	2	3	4	5
CAP4	I have no trouble in learning about the periodic table	1	2	3	4	5
ASDP8	I am satisfied with how well I do in Science classes	1	2	3	4	5
NAN5	I have poor knowledge of where theories come from	1	2	3	4	5
BAN3	I am pretty bad when it comes to cells	1	2	3	4	5
CAN6	I have difficulty learning new things about useful products from oil and rocks..	1	2	3	4	5
GEN7	My friends know more about science investigations than me	1	2	3	4	5
NAP3	I am pretty good when it comes to knowing why we need scientists	1	2	3	4	5
IOSP4	It is important for me to make good progress in Science	1	2	3	4	5

		True for Me				False for Me
GEN5	My friends have a clearer understanding than me about science	1	2	3	4	5
PAP2	I usually do well at learning how forces can move and change things	1	2	3	4	5
EAN2	I usually do poorly at drawing and explaining tables and graphs	1	2	3	4	5
BAN5	I have poor knowledge of living things in their environment, ... e.g. adaptation, competition and food chains	1	2	3	4	5
EAN1	I have difficulty in evaluating an experiment	1	2	3	4	5

Thank you for completing this questionnaire

Science Questionnaire KS3

Name ..... *Boy or girl* .....

Year Group ..... *Date* .....

Languages Spoken.....

THE UNIVERSITY of MANCHESTER

What is this questionnaire about?

This is a questionnaire about what you think about your science lessons. It is not a test. There are no right and wrong answers, just your opinions. It is important, that you try to be as honest as possible when you fill in your answers.

How do you fill it in?

On the next four pages there are lots of statements. There are five answers to each statement and you need to think about the one that fits YOU the best. I will give you some examples of the way I could fill out a questionnaire about things outside of school.

	True of Me				False for Me
I really enjoy watching East Enders on TV	(1)	2	3	4	5
I enjoy eating chocolate more than chips	1	2	(3)	4	5
McDonalds is my favourite meal	1	2	3	(4)	5

I have answered the questions with the small circles. I think East Enders is great and so I have scored it with a (1.) I like chocolate and chips about the same so I have scored it in the middle with a (3.) I don't think McDonalds' food tastes very good, but I don't hate it either, so I have scored it (4.)

		True for Me				False for Me
EAP1	I am good at planning how to do an experiment	1	2	3	4	5
GEP2	I am better at science investigations than my friends	1	2	3	4	5
ASDP6	I have always done well in Science classes	1	2	3	4	5
EAN6	I have difficulty in recording results from experiments	1	2	3	4	5
GEP7	I know more about science than my friends	1	2	3	4	5
ASDP4	I learn things quickly in Science	1	2	3	4	5
EAP4	I have no trouble in thinking up ideas to investigate	1	2	3	4	5
IOSP3	It is important for me to understand the work in Science	1	2	3	4	5
CAN5	I have poor knowledge of atoms, elements and compounds	1	2	3	4	5
ASDP1	I get good marks in Science classes	1	2	3	4	5
NAP5	I have good knowledge of what scientists do	1	2	3	4	5
IOSP2	Doing well in Science is important for me	1	2	3	4	5
NAN3	I am pretty bad when it comes to knowing how science can help us.....	1	2	3	4	5
PAN3	I am pretty bad when it comes to planet Earth and solar system	1	2	3	4	5

		True for Me				False for Me
GSMP6	I get good marks in most school subjects	1	2	3	4	5
PAP1	I am good at understanding about electrical circuits	1	2	3	4	5
PAN2	I usually do poorly at understanding what energy can do	1	2	3	4	5
GSMP3	I have always done well in most school subjects	1	2	3	4	5
GIP2	I am better at science investigations than my other subjects	1	2	3	4	5
CAP1	I am good at understanding about rocks and weathering	1	2	3	4	5
PAN1	I have difficulty in understanding about light and seeing, sound and hearing ..	1	2	3	4	5
GSMN1	I am hopeless when it comes to most school subjects	1	2	3	4	5
GAP2	I usually do well at biology	1	2	3	4	5
CAP5	I have good knowledge of chemical reactions	1	2	3	4	5
GAP3	I am pretty good when it comes to science investigations	1	2	3	4	5
ASDP7	It is important for me to do well I Science classes	1	2	3	4	5
BAN2	I usually do poorly at understanding how our bodies work, e.g. digestion and reproduction	1	2	3	4	5

		True for Me				False for Me
CAN1	I have difficulty in understanding about reactivity of metals	1	2	3	4	5
IOSP1	It is important for me to gain high marks in science tests or exams	1	2	3	4	5
GIN1	Science is my weakest subject	1	2	3	4	5
ASDN3	I am hopeless when it comes to Science	1	2	3	4	5
GAP1	I am good at understanding about physics	1	2	3	4	5
BAP3	I am pretty good when it comes to learning about inheritance and selection	1	2	3	4	5
NAP4	I have no trouble in knowing how scientists work	1	2	3	4	5
ASDP2	Work in Science classes is easy for me	1	2	3	4	5
GAN2	I usually do poorly at science investigations	1	2	3	4	5
GAN5	I have poor knowledge of biology	1	2	3	4	5
GIN5	I enjoy my other subjects much more than science investigations	1	2	3	4	5
GSMP2	I learn things quickly in most school subjects	1	2	3	4	5
PAP3	I am pretty good when it comes to learning about heating and cooling	1	2	3	4	5

		True for Me				False for Me
GAN1	I have difficulty in understanding about chemistry	1	2	3	4	5
ASDP5	Compared to others of my age I am good at Science classes	1	2	3	4	5
NAN6	I have difficulty learning about how we get science knowledge	1	2	3	4	5
GAN6	I have difficulty learning new things about science	1	2	3	4	5
GSMP4	Compared to others my age I am good at most school subjects	1	2	3	4	5
GAN3	I am pretty bad when it comes to physics	1	2	3	4	5
BAP2	I usually do well at putting living things into groups and looking at differences between them	1	2	3	4	5
GAP4	I have no trouble in learning about chemistry	1	2	3	4	5
GSMP5	Work in most school subjects is easy for me	1	2	3	4	5
GAP5	I have good knowledge of science	1	2	3	4	5
EAP2	I usually do well at collecting the results of an experiments	1	2	3	4	5
GIP5	I enjoy science much more than my other subjects	1	2	3	4	5
BAP5	I have good knowledge of respiration and photosynthesis	1	2	3	4	5

		True for Me				False for Me
CAP4	I have no trouble in learning about solids, liquids and gases	1	2	3	4	5
ASDP8	I am satisfied with how well I do in Science classes	1	2	3	4	5
NAN5	I have poor knowledge of where theories come from	1	2	3	4	5
BAN3	I am pretty bad when it comes to cells	1	2	3	4	5
CAN6	I have difficulty learning new things about acids and alkalis	1	2	3	4	5
GEN7	My friends know more about science investigations than me	1	2	3	4	5
NAP3	I am pretty good when it comes to knowing why we need scientists	1	2	3	4	5
IOSP4	It is important for me to make good progress in Science	1	2	3	4	5
GEN5	My friends have a clearer understanding than me about science	1	2	3	4	5
PAP2	I usually do well at learning how forces can move and change things	1	2	3	4	5
EAN2	I usually do poorly at drawing and explaining graphs	1	2	3	4	5
BAN5	I have poor knowledge of living things in their environment, e.g. food chains and habitats	1	2	3	4	5
EAN1	I have difficulty in explaining the results of an experiment	1	2	3	4	5

Thank you for completing this questionnaire