CHAPTER 4

CONCEPTUAL FRAMEWORK FOR THE STUDY

4.1 INTRODUCTION

In this chapter, a conceptual framework for the study is built, based on school effectiveness models and factors indicated in literature that influence science achievement of students. The current research project requires a conceptual framework to classify factors influencing achievement in science and to assume relationships between the clusters of the factors. The IEA has offered its own research framework for its international comparative studies since SIMS, and as a matter of course, TIMSS has designed and developed its instrument based on the IEA framework (Travers & Westbury, 1989). The main focus of the IEA research framework is placed on the intended, implemented, and attained curriculum. The collection of data was designed to identify factors likely to influence student learning and to explain international variation in student achievement, reflecting the IEA's main interest of curriculum per se. It has, however, been pointed out that the factors in each unit of the IEA framework are not strictly categorized or concretely defined to operationalize questionnaire items addressed in TIMSS, and that it lacks a theoretical and empirical basis (Bos & Kuiper, 1999). Some researchers who used TIMSS to explore factors likely to influence achievement tried to supplement the IEA framework with other models (Bos, 2002; Howie, 2002).

Furthermore, research concerning TIMSS should take account of the multilevel structure of the data, which consists of achievement in science, and background information obtained on four levels, namely, student, classroom/teacher,

school/principal, and context level, which means mainly national level. Taking the points mentioned above and the current research questions together, the requirement of the conceptual framework can be met in a school effectiveness model which explains hierarchically the variance in educational outcomes. SER has identified many factors influencing student achievement as reviewed in Chapter 3. School effectiveness models have been built on these findings, reflecting the hierarchical structure of the educational systems. They include the Scheerens model, the Slavin/Stringfiel model and the Creemers model (Scheerens & Bosker, 1997), all of which share commonalities as they are based on input-process-output, multilevel, and complex causal structure (Scheerens & Bosker, 1997).

In order to build the conceptual framework, in particular, the current research referred to the Creemers (Creemers, 1994) and the Scheerens (Scheerens, 1990) models. The research adopted the Creemers model as it explains variance in outcomes in terms of essential factors of learning theory, viz., time, opportunity, and quality. In Section 4.2, it is comprehensively explored as it forms the main basis of the research framework. In Section 4.3, the Scheerens model is introduced, its factors associated with student outcomes in school in terms of education production function. The research also consulted the Shavelson, McDonnell, and Oakes model (hereafter referred to as the Shavelson et al. model), detailed in Section 4.4. Shavelson et al. (1989) formulated a model to ascertain the state of science and mathematics education in school and to improve student outcomes. Their model accounts for the relationship among clusters within the educational system. Considering all above, the research conceptual framework is proposed in Section 4.5 and conclusions are drawn in Section 4.6.

4.2 CREEMERS' MODEL

Creemers' model has often been used in research, and has been modified to reflect the context of various studies (Bos & Kuiper, 1999; van der Werf et al., 2000; Bos, 2002; Kyriakides & Charalambous, 2005). However, according to Bos and Kuiper (1999), the TIMSS data was based on a weak theoretical framework as clarification and definition of factors are not clear enough to operationalize with questionnaire items, and they had to consolidate the IEA's framework with Creemers' theoretically and empirically well-defined factors. Kyriakides and Charalambous (2005) pointed out that TIMSS' attempt to find factors likely to influence achievement in a student-classroom and teacher-school context is in line with the multilevel models of school effectiveness. They proposed that international comparative studies such as TIMSS could be based on educational effectiveness research, e.g., Creemers' model, although the limitations of the TIMSS data lie in testing final outcomes rather than valued-added progress, with a lack of prior knowledge. Using the multilevel modelling of the TIMSS data and identified factors based on Creemers' model, the results showed that the country-level factors had a greater effect than the student-level and teacher-level factors, as seen in the international comparison. This means that more attention should be paid to the vast differences between the various educational systems rather than the results of TIMSS highlighted in a perspective of summative assessment ranking orders.

The Creemers comprehensive model of educational effectiveness (1994) was developed from a review of the empirical research on effective instruction and consideration of Carroll's learning model. The scope of the two models, those of Creemers and Carroll, differ (De Jong et al., 2004), but they do both attempt to explain variances in student outcomes by the same factors of aptitude, time on task, and opportunity to learn. Placing more emphasis on the classroom and teacher, Creemers (1994) focuses on the teaching-learning process in the classroom, where all factors or variables that contribute to educational outcomes

exist. The quality of instruction in the classroom depends on three components, namely curriculum, grouping procedure and teacher behaviour, as shown in Figure 4.1 (below). Amongst them, the most important factor is teacher behaviour, because all those factors depend on how a teacher runs his or her lesson. In other words, it is how teachers implement the curriculum that determines student outcomes, not the curriculum itself, and even grouping which positively influences outcomes can be realized by the teacher's capacity.
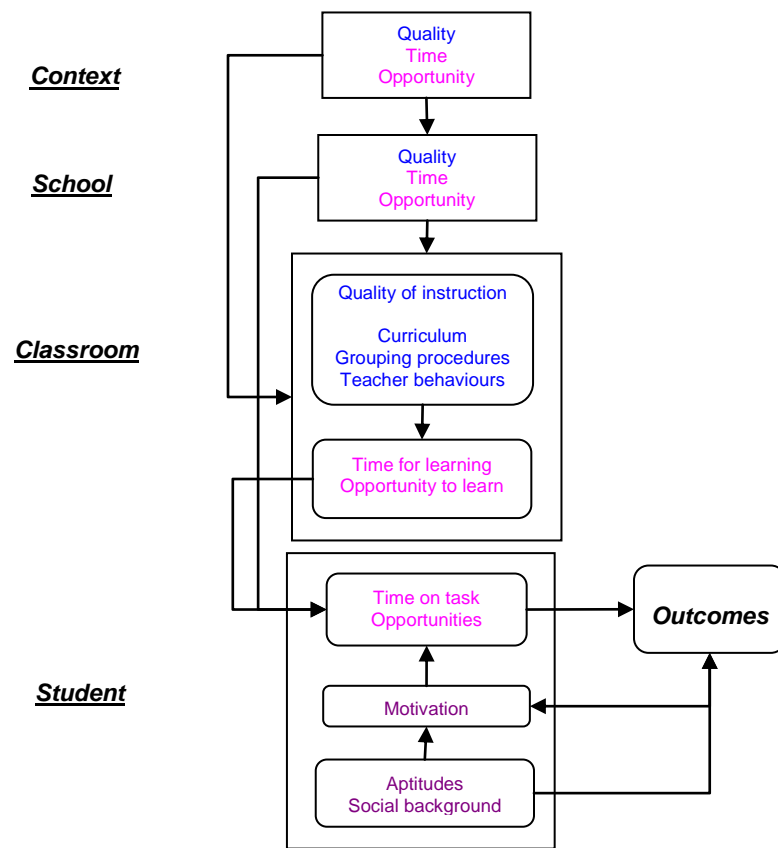
Figure 4.1 Creemers' comprehensive model of educational effectiveness (1994)

Another feature of the model is the three components, viz., quality, time and opportunity, all of which influence achievement across levels. These components emerged from Carroll's (1963) five factors, namely students' aptitude, perseverance, ability to understand instruction, quality of instruction and

opportunity to learn. While Carroll attempted to consider these as time required and explain student learning in terms of individual factors, Creemers places more focus on educational factors, especially quality of instruction within class, which in the form of curriculum, grouping procedure, and teacher behaviour at the classroom level influences time and opportunity at classroom level. They in turn influence time on task and opportunity to learn at the student level and eventually student outcomes. Furthermore, school-level and context-level factors are defined in terms of quality, time, and opportunity, and they in turn influence the classroom level. This attempt can lead to consistently viewing educational effectiveness from a teaching-learning point of view.

In attempting to account for the variance of achievement among students of the two countries, Korea and South Africa, this study assumes that the students sampled in the two countries are learners who grow up through common psychological development, although their contexts such as culture and socio-economic situation are different. Therefore, the Creemers model, based as it is on teaching-learning theory, suits the current research. Analysis from the teaching-learning perspective may offer another insight into the variances, since TIMSS focuses on the effectiveness of curriculum and instruction on student learning, and thereby provides relevant data (Martin et al., 2004). The different levels of the educational system accounted for by Creemers are student, classroom, school and context, all related to each other and contributing to educational outcomes. More detail on each level is now provided separately, for greater clarity.

4.2.1 THE CONTEXT LEVEL

Creemers' model places the context level above the school level, as contextual conditions influence the school level and the classroom level. Context-level factors are differentiated into time, opportunity, and quality, as in other levels. Conditions to develop and enhance quality at the context level are national

policies for effective education, indicator systems or national policies on evaluation, national testing systems, training and teacher support systems, and funding of schools based on outcomes. Conditions for optimal time at the context level are national guidelines for the time schedules in school and supervision of the maintenance of schedules. Conditions for opportunity to learn at the context level include the national guidelines and rules for the national curriculum.

At the context level, consistency, constancy, and control for effective instruction should be guaranteed as formal characteristics. It is of interest that although Creemers defined and acknowledged as important the availability of materials, teachers and other components that support education in schools and classrooms, he did not emphasise it as much Scheerens (Section 4.3).

4.2.2 THE SCHOOL LEVEL

At the school level, Creemers restricted the scope of school-level factors to conditions for the classroom level factors, with conditions for effective instruction at the school discerned in terms of time, opportunity to learn and quality. Conditions for time at the school level are the time schedule for subjects and topics, rules and agreement about time use, and the maintenance of an orderly and quiet atmosphere, so that learning time can be increased in an orderly climate. Conditions for opportunity to learn at the school level are the development and availability of a curriculum, school working plan or activity plan, consensus about the mission of the school, and rules and agreements about implementation of curriculum with respect to transition from one class to another or from one grade to another.

As indicated by the arrows in Figure 4.1 (above), the school level is influenced by the context level. In the same way, quality, time and opportunity at the school level influence education at the classroom level and at the student level. However, as Creemers stated, in order for the factors to effectively operate for

the better outcomes, characteristics should be identified. As far as effective characteristics at the school level are concerned, four are presented to produce effective education: 'consistency' between three main components at the school level, 'cohesion' of all members of school, 'constancy' for the total school career of students, and 'control' to assess student and teacher, as well as a well-organized school climate.

4.2.3 THE CLASSROOM LEVEL

The Creemers model emphasizes the classroom level in particular, since this is where learning and teaching take place and effectiveness of education created. For this reason, the model is developed around the instructional conditions in a classroom. A central component, quality of instruction consists of curriculum, grouping procedure, and teacher behaviour, which interrelate and thus maximize quality of instruction. This quality of instruction in turn influences time for learning and opportunity.

The model offers more distinguishing factors from reviews of research in relation to curriculum, grouping procedure, and teacher behaviour (Table 4.1, below). Firstly, curriculum refers to the documented materials at the classroom level used by teachers and students in the instructional process. Creemers argues that the degree of the implementation of curriculum by teachers is more influential on student achievement than curriculum itself. Secondly, grouping based on mastery learning is strongly related to evaluation, feedback and corrective instruction to overcome deficiencies in learning. The grouping of students can also influence the allocation of time and opportunities for learning. On the other hand, the effect of grouping depends on the capacity of teachers. Grouping procedures seem to reflect the Dutch, US, and UK education systems, which practise streaming or tracking in primary and secondary education to overcome the difference between students. Thirdly, teacher behaviour can be translated into two sub-components,

viz., management behaviour to control the class, to prepare the students for learning and to maintain learning, and instructional behaviour related to teaching.

Apart from the main components explained above, Creemers stated that there are effective interactions between the components at each level. For example, at the classroom level, 'high consistency' between three components can cause a synergistic effect and eventually improve student achievement. Without consistency, the results could be worse. For example, when C2005, a new outcomes-based curriculum, was implemented in South Africa, teachers did not have the chance to adapt to it or adopt it, causing considerable confusion among teachers (Hoadley & Jansen, 2002). The curriculum has been revised subsequently but that a similar problem still exists. This shows the importance of developing consistency between the components in order to successfully realize effective education.

4.2.4 THE STUDENT LEVEL

Individual factors such as aptitude, background, and motivation determine student outcomes at the student level. In addition, outside-controllable factors such as time on task and opportunity to learn also influence student outcomes. In particular, the two factors are important as they might be controllable in the educational system. All those factors are derived from Carroll's model, but it should be noted that opportunity to learn is defined as supply of learning material, experiences and exercises, as opposed to time dimension as in Carroll's model. As defined in Chapter 3, aptitude indicates what a student already knows, and involves general ability and prior learning. The background factor reflects socio-economic status (SES), an important factor in explaining student outcomes. Creemers considered motivation only at the student level, while some researchers see motivation as also resulting from teacher's or school's expectation (Papanastasiou, 2002; Lyons, 2006). Motivation affects student achievement and vice versa, as shown by the two-way arrow in Figure 4.1

(above). In the conceptual model for the study built in Section 4.5, the term 'motivation' was replaced by 'attitudes' to encompass broader meaning, since attitudes imply more comprehensive meaning, such as feeling, cognition, and behaviour (Simpson et al., 1994).

In summary, Creemers (1994) classified four levels, viz., context, school, teacher, and student, and three components, i.e., quality, time, and opportunity to learn. The details of components in each level are described in Table 4.1 (below).
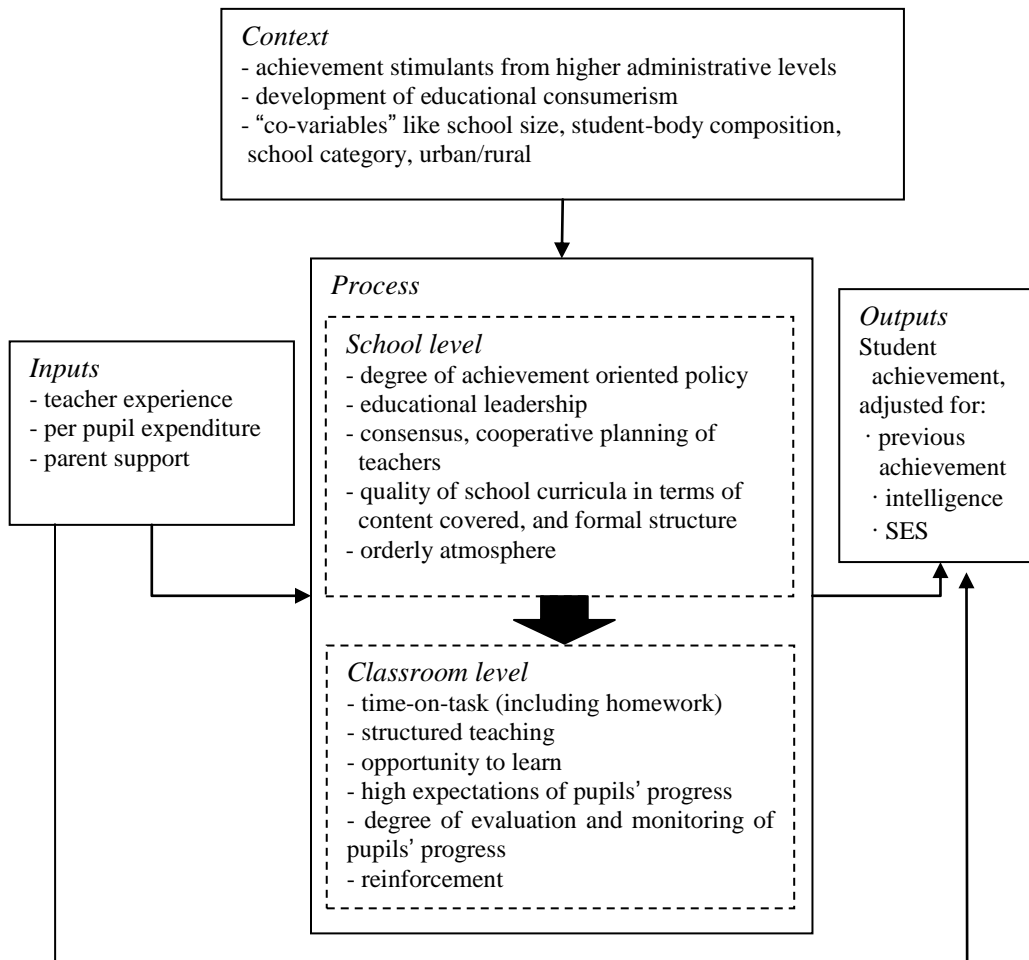
Table 4.1 *The Creemers' comprehensive model of educational effectiveness (1994)*

Levels	Components		Details of components	Formal criteria
Context	Quality		Policy focusing on effectiveness Indicator system/policy on evaluation /national testing system Training and support system Funding based on outcomes	Consistency Constancy Control
	Time		National guidelines for time schedules Supervision of time schedules	
	Opportunity		National guidelines for curriculum	
School	Quality (educational)		Rules and agreements about classroom instruction Evaluation policy /evaluation system	Consistency Cohesion Constancy Control
	Quality (organizational)		Policy on intervision, supervision, professionalization School culture inducing effectiveness	
	Time		Time schedule Rules and agreements about time use Orderly and quiet atmosphere	
	Opportunity		School curriculum Consensus about mission Rules and agreements about how to implement the school curriculum	
Classroom	Quality of instruction	Curriculum	Explicitness and ordering of goals and content Structure and clarity of content Advance organizers Evaluation Feedback Corrective instruction	Consistency

		Grouping procedures	Mastery learning
			Ability grouping
			Cooperative learning highly dependent on
			    Differentiated material
			    Evaluation
			    Feedback
			    Corrective instruction	

		Teacher behaviour	Management /orderly and quiet atmosphere
			Homework
			High expectations
			Clear goal setting
			  Restricted set of goals
			  Emphasis on basic skills
			  Emphasis on cognitive learning and transfer
			Structuring the content
			  Ordering of goals and content
			  Advance organizers
			 Prior knowledge
			Clarity of presentation
			Questioning
			Immediate exercises
			Evaluation
			Feedback
			Corrective instruction	

	Time for learning Opportunity to learn			
Student	Time on task Opportunities used Motivation Aptitudes Social background			

4.3 SCHEERENS' MODEL

The Scheerens model (1990), an integrated, multilevel school effectiveness model, attempts to explain school effectiveness from a systematic point of view, as opposed to Creemers' model which places emphasis on classroom processes from a perspective of teaching and learning theory. Although Scheerens acknowledged the importance of classroom level factors in school effectiveness, his model places more emphasis on the functioning of a school as an organizational system, following his definition of effectiveness as productivity. Such an emphasis on organizational structures or managerial processes is more appropriate for meeting demands of policymakers or decision-makers, who find manipulative factors to promote school effectiveness, than for practitioners such as teachers, who want to improve teaching and learning in the classroom.

As depicted in Figure 4.2 (below), Scheerens adopts a two-dimensional analytic scheme which contains context-input-process-output and educational multi-levels, viz., pupil, classroom, school, and environment. The context-input-process-output dimension reflects economic productivity and the pupil-classroom-school-environment dimension indicates a hierarchical and nested structure of the school system. The context and input cluster mainly involves resource-related factors, whereas the school and classroom level concerns attitudes, ethos, climates, and teaching practice. From an additional review of previous literature and research (Scheerens & Bosker, 1997), overarching factors related to educational effect at the classroom and the school level were explored, together with the definitions.

Context
- achievement stimulants from higher administrative levels
- development of educational consumerism
- "co-variables" like school size, student-body composition,
  school category, urban/rural

Process

School level
- degree of achievement oriented policy
- educational leadership
- consensus, cooperative planning of
  teachers
- quality of school curricula in terms of
  content covered, and formal structure
- orderly atmosphere

Inputs
- teacher experience
- per pupil expenditure
- parent support

Outputs
Student
 achievement,
adjusted for:
· previous
  achievement
· intelligence
· SES

Classroom level
- time-on-task (including homework)
- structured teaching
- opportunity to learn
- high expectations of pupils' progress
- degree of evaluation and monitoring of
pupils' progress
- reinforcement

Figure 4.2 Integrated model of school effectiveness (Scheerens, 1990)

The main factors identified from the aforementioned review are as follows:

1. Achievement orientation, high expectations, teacher expectations

2. Educational leadership

3. Consensus and cohesion among staff

4. Curriculum quality, opportunity to learn

5. School climate

6. Evaluative potential

7. Parental involvement

8. Classroom climate

9. Effective learning time (classroom management)

10. Structured instruction

11. Independent learning

12. Differentiation, adaptive instruction

13. Feedback and reinforcement

Some of these factors, such as educational leadership, school and classroom climate, or parent involvement, are not identified in Creemers' model, and so may be considered as making up for gaps in it.

4.4 SHAVELSON, MCDONNELL, AND OAKES' MODEL

Shavelson, McDonnell, and Oakes (1989) developed a comprehensive model of the educational system, aiming at an indicator system that would measure the state of mathematics and science education. Their model could help policymakers determine the nature of current problems, evaluate the factors influencing educational trends, monitor the effects of policy, and identify interventions to improve student performance.

Shavelson et al.'s model features many arrows, which indicate the direction of influence, as shown in Figure 4.3 (below). Looking at the arrows around 'achievement', educational outputs or outcomes are directly influenced by instructional quality, together with student background. The instruction quality, in turn, is affected by the school, the curriculum, teaching quality, and student background. The school quality can influence the instructional quality by working conditions, including class size, classroom resources, occupational support, and school-wide standards. The curriculum quality can have influence on the instructional quality by giving students the opportunity to learn, and the teaching

quality can affect the instructional quality by teacher qualifications and general patterns of teaching practices.
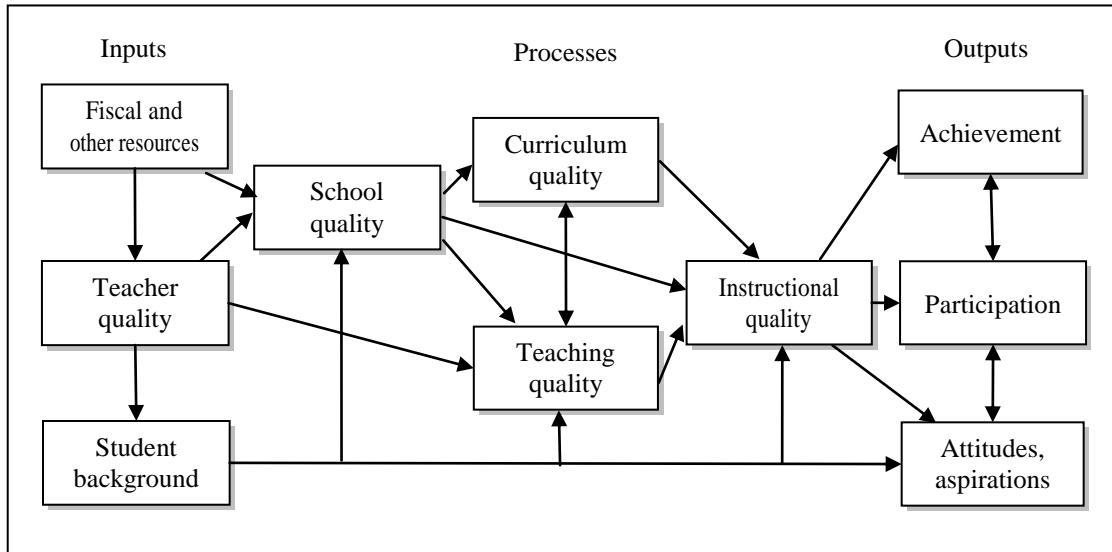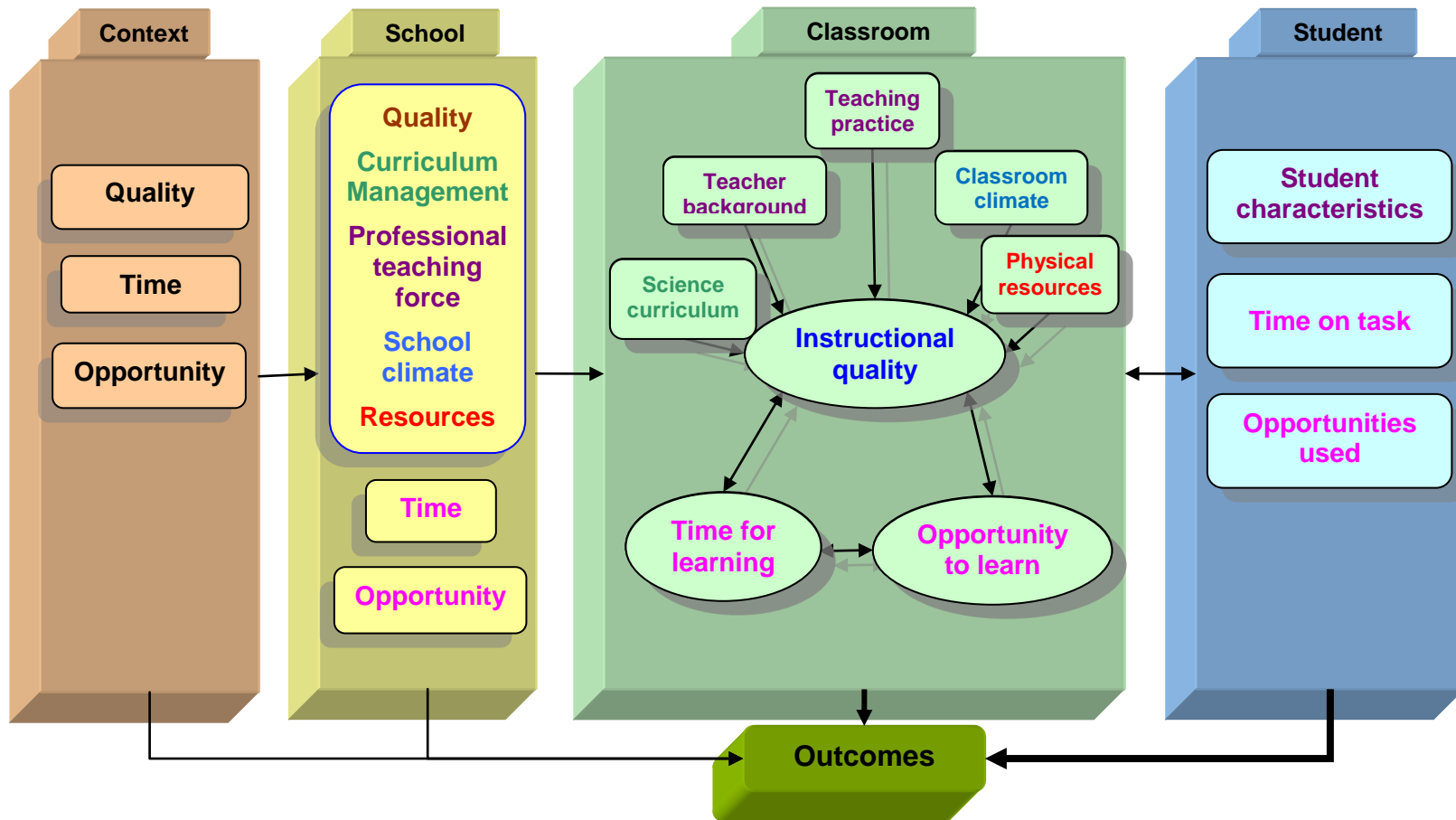

Figure 4.3 **A comprehensive model of the educational system (Shavelson et al., 1989)**

What should be highlighted in the model compared to the two presented above is that the instructional quality is affected by student background as well as school, teacher and curriculum quality. School effectiveness models such as those of Scheerens and the Creemers assume that the instructional quality in classroom is affected only by curriculum, grouping procedures, teacher behaviour or higher-level factors such as educational leadership, but not by lower-level factors such as student attitudes. There is no doubt that the quality of science education rests on the quality of instruction that students receive. This in turn is largely determined by such teacher factors as the qualifications of science teachers, and school factor such as the conditions under which these teachers work. Taking into consideration that there are not only teachers but also students in the classrooms, and that teachers should focus on both teaching subject matter and enforcing classroom discipline with the dual responsibilities (Shavelson et al., 1989), student factors such as attitudes, family background and previous

133

performance can be related to the instructional quality. Consequently, relationship between factors and levels should be considered in a reciprocal way, not just as one directional.

4.5 CONCEPTUAL FRAMEWORK FOR THE RESEARCH

The conceptual framework for the research is mainly built on the three models explored previously, namely Creemers', Scheerens', and Shavelson et al.'s. Furthermore, factors specific to science achievement are combined in the framework. Broadly stated, the new model differs from that of Creemers in two important aspects, namely inclusion of resources at each level, and reclassification of sub-components in quality at classroom and school level. As pointed out above, Creemers examined all the factors likely to influence student achievement in terms of time, opportunity, and quality, thereby limiting possible factors to these three categories, consequently, risking missing important factors such as resources and leadership.

Context
- Quality
- Time
- Opportunity

School

Quality
- Curriculum Management
- Professional teaching force
- School climate
- Resources

Time

Opportunity

Classroom
- Teaching practice
- Teacher background
- Classroom climate
- Science curriculum
- Physical resources
- Instructional quality
- Time for learning
- Opportunity to learn

Student
- Student characteristics
- Time on task
- Opportunities used

Outcomes

Figure 4.4 **A proposed model of effectiveness of science education**

In terms of resources, Creemers (1994) and Scheerens (1992) regarded them as an input factor. Creemers did not include resources in the structured model although he mentioned resources at context level, together with the definition as mentioned above. Scheerens showed resources in the input and the context unit. In the new model, mainly based on the Creemers model, as shown in Figure 4.4 (above), the resource factor is added to quality component at classroom, school, and context level. The resources correspond to infrastructure or instructional materials to support teaching and learning, such as the library, laboratory, experimental apparatus, or computers.

In particular, resources are important in the current research for two reasons. One is that this research concerns developing countries, in which resources are more likely to influence student achievement than in developed countries (Fuller, 1987; Glover, 1992; de Feiter et al., 1995; Scheerens, 2001; Reddy, 2006). The other is that the research focuses on the subject of science, which is thought to depend much more on physical resources than other subjects (Rogan, 2000). This is confirmed by the previously reviewed literature.

As the second aspect of the new conceptual framework, a reclassification of sub-components was made to encompass some effective factors which are not emphasized in the Creemers model. The quality component at the school level contains curriculum management, professional teaching force, school climate, and resources, including time and opportunity to learn. Instructional quality at the classroom level consists of science curriculum, teacher background, teaching practice, classroom climate, and physical resources. Based on the literature review in Chapter 3, the climate factor and resource factor are included at both levels. As compared to Table 4.1 (above), other sub-components were shown in more detail in Table 4.2 (below).

Table 4.2 *Proposed factors at the student, classroom, school, and context level*

Levels	Factors		Details of factors
Student	Time on task		The time spent on homework, private tutoring, and outside-school activities related to science
	Opportunities used		Absenteeism, Participation in science course, Homework, Tutoring
	Student characteristics	Aptitudes towards science	Prior achievement
		Attitudes towards science	Self-confidence, Motivation, Enjoying science, Valuing science
		Social context	SES, Parent education, Books in home, Parent involvement, Peer environment, Ethnicity, Language, Gender
Classroom	Instructional quality	Science curriculum	Science textbook and workbook
		Teacher background	Academic skills, Teaching assignment, Teacher experience, Degree, Certification, Major area of study, Professional development, Gender
		Teaching practice	Direct instruction-Structured teaching, Questioning, Manipulation-practical work, Enhanced material, Assessment-test, feedback, reinforcement, Inquiry-problem solving, Enhanced context-linkage with daily life, Collaborative learning
		Classroom climate	High expectations, relationships between teachers and students, and among students, Management /orderly and safety atmosphere Teachers' attitudes towards student and science teaching Students' attitudes towards class, Student disruption, intrusion, and interruption
		Physical resources	Labatory, Equipments and materials for science experiments, Computer, Software, Internet access, Video-audio facility, Teaching condition, Class size
	Time for learning		The time assigned by science teacher to teach science contents, Instructional time
	Opportunity to learn		The science contents taught by science teachers
School	Quality	Curriculum management	Rules and agreements about classroom instruction, science-related extracurricular activities, Curriculum-related task or decision-choosing textbook, determining course content, course offerings, student grading policies, assigning teachers to science classes, and instructional days or hours per year
		Professional teaching force	Educational leadership, Consensus or cohesion among school staffs including teachers, Stable body of teachers, Regular meeting of teachers
		School climate	High expectation, Achievement orientation, Community SES or School location, School discipline-student disrespect for teachers, absenteeism, tardiness, bullying, fighting, and theft, Higher student body mobility, Orderly and safety atmosphere
		Resources	Building, Grounds, Gymnasia, Library, Heating/cooling and lighting, Budget for science supplies, General instructional material, Budget-related resources-teacher salary, student-teacher, expenditures per pupil, administrative inputs, and facilities
	Time		Time schedule per week and per year, Duration of class, Rules and agreements about time use, Frequency of field trips
	Opportunity		School science curriculum offered, Science field trips Rules and agreements about how to implement the school science curriculum Curricular differentiation in science
Context	Quality	Curriculum	Policy focusing on effectiveness Indicator system/policy on evaluation / National testing system Training and support system, Policy on science curriculum
		Resources	Expenditures per pupil, Expenditures as a percentage of per capita income Average teacher salary, Pupil/ teacher ratio , Funding based on outcomes
	Time		National guidelines for time schedules Supervision of time schedules
	Opportunity		National guidelines for curriculum

Another focus should be placed in particular on the classroom level in the new model shown in Figure 4.4 (above). Capacity to be successful in terms of implementation depends on various factors such as resources, teacher, student, and school support (Rogan & Grayson, 2003; Rogan & Aldous, 2005). Therefore, the point made here is that teaching practice and teaching conditions mainly determine instructional quality, but it cannot depend only on teacher behaviour, curriculum or grouping factors as described in the Creemers model. As in the model of Shavelson et al. (see Figure 4.3, above), instructional quality can be influenced by school factors, student factors, teaching quality, and curriculum. The arrow drawn from the student level to the classroom level in Figure 4.4 (above) reflects this point. The studies based on constructivism have evidenced that the active role of the learner is important to good subject matter teaching (Brophy, 1992; Scheerens, 1997). This is decidedly true in science teaching in particular. The main assumption in multi-level educational effectiveness models, namely that higher levels facilitate operations of lower levels (Scheerens, 1997), should therefore be re-considered. Accordingly, the linkages between clusters are not simple one-way processes as in the Creemers model.

On the other hand, there is a need to examine the deficiencies of the model mainly adopted, the Creemers model, and to explain the reason for modification. Creemers considers the classroom as key to effective instruction and looks upon the school and context levels as conditions to facilitate effective instruction in the classroom. When these levels are defined as organizational conditions to support classroom instruction, a lack of interrelationship between them occurs. To avoid this lack and to investigate factors other than those found at the classroom level in the same vein as them, he defined school-level and context-level factors in light of three components, namely, time, opportunity to learn, and quality.

Nonetheless, this definition restricts the selection of school level factors only to those factors conditional for, and directly related to, quality of instruction, time or

opportunity to learn (Creemers et al., 2000). Actually, his attempt seems to miss some important factors related to student outcomes, especially at other levels. For example, educational leadership, which has been acknowledged to have an effect on achievement (although not in the Netherlands) (Creemers, 1994), is not found clearly in his model, even though he stated teacher's management at the classroom is in the similar vein. As for resources, he mentioned these but did not specify them in the model, as mentioned above. Given that the TIMSS data to be examined in the research offers information about these factors, more should be added to the Creemers model. Exhaustive factors shown in Scheerens and Bosker (1997) supplemented this deficiency in the model built for the study.

Although each component is distinguished as shown in Figure 4.4, they interact and are interrelated. Furthermore, it should be kept in mind that those interactions or interrelation are not causal but correlational. In designing the new model, some effort was made to avoid a labyrinthine scheme and to parsimoniously use arrows, which show linkages across the factors and the levels. It is therefore understood that there are more linkages between factors than shown by the arrows in Figure 4.4.

4.6 CONCLUSION

In this chapter, the research framework was formulated as described up to this point. The framework consulted some representative educational effectiveness models, including the Scheerens model and the Creemers model. The models referred to shared commonalities as based on input-process-output, multilevel, and complex causal structure (Scheerens & Bosker, 1997). In particular, the current research investigated thoroughly the Creemers and reflected some factors proposed in the Scheerens model. It is recognized that Scheerens offers factors associated with student outcomes in school in terms of education production function, while Creemers explains variance in outcomes in terms of

essential factors of learning theory, viz., time, opportunity, and quality. Also, the Shavelson, McDonnell, and Oakes model contributed to building the model when accounting for the relationship among clusters within the educational system.

Accordingly, the theoretical framework of the research is based on the Creemers model and considers the factors shown in the Scheerens model. Finally, factors specific to science explored above are incorporated and, as a result, some modifications are made reflecting the model of Shavelson et al. (1989), where relationships between levels and factors in educational systems are clear. Just as effective teacher behaviour must be qualified in different grades or contexts (Brophy & Good, 1986), the possibility exists that the conceptual framework for the research may need to be modified depending on systems or subjects. The model developed here, which emphasizes resource factors, can contribute to SER in developing countries and in science.

CHAPTER 5

RESEARCH DESIGN AND METHODOLOGY

5.1 INTRODUCTION

The purpose of this research is to explain the difference of science achievement between Korea and South Africa by undertaking a secondary analysis of existing data, namely TIMSS 2003. As this is a secondary analysis, and given that it involves quantitative methods and seeks to explore the nature of relationships in social phenomena, this research falls within a post-positive paradigm. Post-positivism emerged after World War II, as an alternative to positivism, which is applicable to the natural science but not to the social sciences (McMillan & Schumacher, 2006; Cohen et al., 2007).

Positivists believe that causes that explain effects or outcomes can be acquired by scientists' observations and measurements, since knowledge is objective and tangible. Accordingly, one would understand, explain, and predict phenomena in the world. However, while this may work well in the natural sciences, the social sciences research complex human behaviour that reflects the unconscious mind and is therefore not subject to such rational methodology (Cohen et al., 2007).

Post-positivism and positivism have in common quantitative approaches, such as observation and measurement, but the degree of accuracy is different. Knowledge of objective reality can be identified by means of careful observation and numerical measurements but, under post-positivism the objectivity, or generalisability, can be ensured by multiple measurements. As human behaviour and quality of social phenomena are complex and often intangible,

the measurement or research may be imperfect. Therefore, post-positivism allows for some limitations, and contextual factors that account for the complexity of human nature (Cohen et al., 2007). The current research is located in a post-positivism paradigm, based on findings identified previously and examining data from a perspective of a theory that is attempts to enforce, then modify, and finally improve in terms of its generalisability and objectivity.

The rest of the chapter is as follows: Secondary analysis based on survey data is discussed in more detail in Section 5.2, after which the research questions are examined in Section 5.3. Sections 5.4, 5.5, and 5.6 briefly introduce the sampling, data collection and instruments. Thereafter, data analysis procedures for the study are described along with presenting appropriate statistical processes (5.7). Thereafter, correlation analysis and multilevel analysis are introduced in Section 5.8 and 5.9. Methodological norms in TIMSS are discussed (5.10). Ethical considerations related to the research are discussed (5.11), and conclusions are drawn in Section 5.12.

5.2 SECONDARY ANALYSES OF DATA

Secondary analyses can be defined as any further research that studies diverse problems with the same data as previously collected by other researchers to study a problem (Herrnson, 1995). The intention of using secondary analysis is to obtain a more in-depth understanding of the subject matter at hand, or present interpretations, conclusions or knowledge additional to, or different from, those presented in the primary study (Dale, Arber & Procter, 1988). Given the aforementioned points, secondary analyses can be seen as a good way to increase knowledge in research (Herrnson, 1995). In some cases, secondary analyses may aim at generating hypotheses and identifying critical areas of interest that can be examined during primary data gathering activities.

Regardless of the intention of secondary analysis, there are some advantages and disadvantages that should be acknowledged. In light of advantages, secondary analysis may be conducted for a number of reasons, including data quality, adequate sample size, time efficiency, and cost effectiveness (McMillan & Schumacher, 2006). The advantages are explored in more detail as follows:

Data quality: Taking into consideration the aspect of data quality, the TIMSS data was proved to offer a high degree of validity and reliability as TIMSS developed instruments ensuring well-designed processes, careful fieldwork and attention to methodological norms such as validity and reliability, consistency over time, and national representativeness of their large sample size (Dale et al., 1988). When primary data has a high degree of validity and reliability, it is evident that this applies to their use in a secondary analysis. The reliability and validity of research analysis can be enhanced, particularly with a large sample.

Sample size: In addition large nationally representative samples, TIMSS presents data at many levels, including student, class, school, and context, and so provides the most appropriate data to answer the current research questions. The secondary analyst may choose only one among various levels, or examine individuals within the context of the larger group or organization. In particular, the advantage of studying individuals nesting in a group is preferred in education research, where student achievement or attitudes can be examined within the school context.

Time and cost: Another advantage of secondary analysis is saving time and costs. Unlike formal primary data collection and analysis processes, secondary analysis can be carried out more quickly and, therefore, the cost is reduced. This kind of financial saving encourages the secondary analyst who has no source of funding or few resources for carrying out the primary data collection to become involved in secondary analysis. While large-scale and worldwide studies, such as IEA, take place over a considerable time span and require

substantial funding, conducting a secondary analyses on these studies is easily effected by the global availability of the data.

On the other hand, there are some disadvantages to secondary analysis that should be noted. Occasionally information or data is not always what might be needed to answer the research questions, and so might have the potential to bias the study. Furthermore, secondary analysts are more likely to become overwhelmed by the considerable volume of data available, making it difficult to determine its quality (McCaston, 1998). Therefore, it is important to develop a strong theoretical framework beforehand, and secondary data should be examined and presented with in-depth interpretation and analysis within the theoretical framework.

This is a secondary analysis of the survey for TIMSS 2003 and as such is research in which the researcher has not been involved in the actual collection of the data (Bryman, 2004). Survey research can provide an analytical framework for research while less structured forms of methods, such as interviews or ethnographic observations, may enhance insights and understandings on the social phenomena in question (Dale et al., 1988). Survey research tends to be regarded as suitable for exploratory and confirmatory analysis which may lead to, where possible, a modification of the original theory. Therefore, secondary analyses using survey data can benefit from different methods and theoretical perspectives to answer the identified research questions.

5.3 DISCUSSION OF RESEARCH QUESTIONS

The current study brought up the two research questions to ascertain the difference and similarity between Korean and South African science achievement. A research design and method to be used can be determined

depending on the two questions to be answered. Therefore, the main two questions were examined in terms of method.

The first main question is '***To what extent does TIMSS 2003 reflect factors related to effective science education?***' The first main question aims to identify, from literature, factors influencing performance in science. In order for this question to be answered, the study examined previous research comprehensively and adapted the conceptual framework from the literature consulted. As a result, many factors specific to science were identified, as explored in Chapter 3. In the process of the exploration, SER has played a role in directing the way in which the factors should be identified. However, these factors and the framework should be verified empirically using the current analysis.

Since TIMSS collected data at student, teacher, classroom, school, and context levels, it would be reasonable to examine factors corresponding to each level. Accordingly, the first research question can be broken down into three sub-questions: '*Which factors at the school level influence science achievement?*', '*Which factors at the classroom level influence science achievement?*', and '*Which factors at the student level influence science achievement?*'. An exploration of variables at different levels was carried out to address these questions.

As the first step of verification, the TIMSS data related to information was examined in terms of the framework developed in Chapter 4, specifically that regarding the background of the students, the teachers and the classroom, and about the schools and their principals. This was then compared to the constructs in the conceptual framework depicting teachers' characteristics, including classroom practice, students' characteristics (e.g., SES and attitudes), and school characteristics, (e.g., facilities). Thereafter, items in the TIMSS questionnaires were explored and selected for further analyses, namely factor, reliability, and correlation analyses. This resulted in a number of specific issues

influencing student performance in science, the results of which are presented in Chapter 6 and 7.

The second question is '***To what extent do the factors derived from the analysis explain the differences in the achievement of Korean and South African students?***' This arose from a perspective of international comparative studies, which, like the current research, attempts to find similarities and differences in background factors related to student achievement (Bos, 2002). In order for this to be investigated, it needs to be answered step-by-step as follows: '*Which factors influencing achievement are generic when comparing Korea and South Africa?*', '*Which factors influencing achievement are specific to Korea?*', '*Which factors influencing achievement are specific to South Africa*?', and '*How do these generic and specific factors explain the difference in the performance of the two countries?*'

For the study to reflect on such hierarchical structure of the data influencing student achievement, the research method adopted a multilevel approach to analysis, making it possible to examine influences between the levels as well as each level's impact on student achievement. In addition, the multilevel analysis involves the interaction between and within each level, allowing factors specific to students, classroom, and school to be studied simultaneously. The results of this multilevel analysis are presented in Chapter 8.

In terms of multilevel aspects, previous research has shown that school-level factors are more likely to influence student achievement in science in developing than developed countries (Heyneman & Loxley, 1983; Fuller, 1987; Fuller & Clarke, 1994). It is therefore believed that school-level factors are more likely than student-level factors to play a significant role in South African Grade 8 student achievement in science. In contrast, given the highly competitive educational zeal displayed by Korean students (Martin, Mullis, Gonzalez & Chrostowski, 2004), it is plausible that student-level factors will be found to influence student achievement more than other level factors.

Cross-national comparative research such as the current study seeks to understand similarities and differences in background factors related to student achievement measured in TIMSS. When these questions are answered, one can then investigate whether factors that are important for consideration in South Africa are also important to consider in a Korean context, or which factors generally apply to both countries. However, commonly identified factors may lead to generalizations about the effect of particular student, teacher, classroom or school-level across educational systems in both Africa and Asia. Ultimately, answering these questions also assists in understanding the educational contexts in each country and the reasons for cross-national differences in achievement.

5.4 SAMPLE

TIMSS studied achievement in two target populations, namely, population 1, consisting of mostly 9-year-olds at the time of testing, and population 2, consisting of mostly 13-year-olds at the time of testing. This study has focussed on population 2, consisting of Grade 8 learners. As South Africa did not participate in TIMMS 2007, the TIMMS 2003 results were the most recent data that could be used for this study.

The Korean sampling frame for TIMSS 2003 included 607,123 students with the teacher sample being selected from the class of the sampled school automatically. Some schools were excluded for various reasons, such as their being situated on far-away islands, or in remote areas, or because they were too small. Accordingly, the sampling frame resulted in 601,123 students as of April in 2002 (Park, Hong, Lee & Cheon, 2003). Korea adopted an administrative district as an explicit stratification variable and constructed 16 sampling frames from which the sample was drawn. In addition, an implicit stratification was identified, namely urbanization and gender (Park et al., 2003). As a result, Korea sampled 151 schools with 16 explicit strata by province and

83 implicit strata by urbanization and gender, resulting in 5,300 learners participating in the study (Martin, Mullis & Chrostowski, 2004).

South Africa stratified the sample by two dimensions, viz., by province and language of teaching and learning. Consequently, 265 schools were sampled with 9 explicit strata by province and 19 implicit strata by language, resulting in approximately 9,000 learners being tested across the provinces (see Table 5.1, below). Where class size was over 50 learners, 40 learners from the whole class were sub-sampled with probability-proportional-to-size (PPS) (Martin, Mullis & Chrostowski, 2004).

It should be noted that the number of schools sampled in South Africa exceeds those of most countries, which sampled around 150 schools. This oversampling was designed to produce provincial statistics across the nine provinces (Reddy, 2006) as it was suspected that a broader range of gaps within the country in terms of education, race, and social-economic status would emerge, and thus it was intended to get in-depth and precise insight into these gaps (Howie, 2001).

Table 5.1 *Schools sampled in Korea and South Africa*

	Schools sampled	Sampled schools participating	Replacement school	Total schools	Total learners
Korea	151	149	0	149	5309
South Africa	265	241	14	255	8952

Source: Martin, Mullis, Gonzalez & Chrostowski, 2004

5.5 DATA COLLECTION

The TIMSS 2003 data was collected at the end of the school year. In countries in the Northern Hemisphere, where the school year typically ends in June, the

assessment was administered in April, May, or June 2003. In the Southern Hemisphere, including South Africa, the school year typically ends in November or December so the assessment in these countries was conducted in October or November 2002 (Martin, Mullis & Chrostowski, 2004). Korea tested from 14 to 19 April 2003 (Park et al., 2003) and South Africa tested from 21 October to 1 November 2002 (Reddy, 2006).

Each participating country was responsible for the data collection, using standardized procedures developed for the study and based on training manuals created for school co-ordinators and test administrators. As explained in Chapter 2, a Quality Control Monitor (QCM) appointed in each country monitored the procedures for her/his country. Additionally, the QCM interviewed the National Research Coordinator (NRC) and visited a selection of the sampled schools. The school co-ordinators and field workers took part in the training course run by the Korea Institute of Curriculum and Evaluation (KICE) two weeks in advance of the administration. The training course involved 34 to 40 schools at a time and took place four times across the country. A similar training procedure took place in South Africa.

Specifically in Korea, the test was supposed to be administered to Grade 8 in February 2003, however, there was a problem because at that time Korean schools had only just been open for a few days, the academic year having ended in February and the new one started in March. The Korean administration date (14 to 19 April 2003) was negotiated with the sampled schools and is shown in the international report. However, because the academic year had only just begun, in agreement with the international study centre, the test was administered to Grade 9 students but the students tested were to be reported as Grade 8, and all science teachers were to respond to the questionnaire reporting the Grade 8 classes taught by them the previous year (Park et al., 2003).

South Africa conducted the test under the auspices of the Assessment Technology and Education Evaluation Research Programme at the Human Sciences Research Council (HSRC). Just as in other southern hemisphere countries, South Africa administered TIMSS instruments to Grade 8 students at the end of their academic year, which is from 21 October to 1 November 2002. The HSRC assigned *AC Nielsen and Mictert*, an outside agency, for the administration of the instruments in schools. This body trained their data collectors using a manual prepared by the TIMSS International Study Centre (ISC) to assist when TIMSS is administered in the sampled schools (Reddy, 2003). School staff was also supposed to help with logistical arrangements, such as identifying testing locations.

In both countries, sampled students each used one booklet containing both mathematics and science items for 90 minutes and responded to the questionnaire for 30 minutes, taking a break between the assessment and the questionnaire.

The TIMSS 2003 data for Korea and South Africa was accessed from the IEA website, which is in the public domain[8].

5.6 INSTRUMENTS

TIMSS 2003 consisted of mathematics and science achievement test items, as well as questionnaires. The achievement test was designed to assess mathematics, science knowledge and skills based on school curricula for Grade 8 learners. As explored in Chapter 2, assessment is addressed by the form of a booklet containing both mathematics and science items, and each student takes one booklet. For the purposes of this study only the assessment and questionnaires concerning science will be focussed on. The assessment items were developed using the TIMSS assessment framework and specifications, as

---

[8] http://www.iea.nl/iea_studies_datasets. html.

well as depending on the contribution of NRCs during the entire process of the regular meetings in which the NRCs could add their inputs (Martin, Mullis & Chrostowski, 2004). The questionnaires were designed to provide a context for the performance scores, focusing on students' backgrounds and attitudes towards science, the science curriculum, teachers of science, classroom characteristics and instruction, and school context (Martin, Mullis, Gonzalez & Chrostowski, 2004). The details of two instruments are provided in the following sections respectively.

5.6.1 THE SCIENCE ASSESSMENT

The science assessment was framed by two organizing dimensions, a content dimension and a cognitive dimension. The content dimension subsumes five content domains: life science, chemistry, physics, earth science, and environmental science, and consists of three cognitive domains: factual knowledge, conceptual understanding, and reasoning and analysis (Mullis et al., 2003).

The five content domains are described in more detail in Table 5.2 (below). Concepts related to matter and energy overlap considerably in both the physics and chemistry domain and Grade 4 does not separate them as opposed to Grade 8. Environmental science, as a field of applied science concerned with environmental and resource issues, involves concepts from the life, earth, and physical sciences.

Table 5.2 *Science content domains and target percentage in TIMSS 2003 for Grade 8*

Domains	Main topic areas	Target percentage devoted
Life science	understanding of the nature, function of living organisms, the relationships between them, and their interaction with the environment	30%
Physics	general physical states of matter and their transformation	25%
Chemistry	the properties, composition, classification, and particular structure of matter	15%
Earth science	earth structure and physical features, the earth's processes, cycles and history, and the earth in the solar system and the universe	15%
Environmental science	changes in population, use and conservation of natural resources, and changes in environments	15%

Source: Adapted from Mullis et al., 2003

The cognitive dimension involves the sets of behaviour expected of students as they engage with the science content. This domain is divided into the three areas of factual knowledge, conceptual understanding, and reasoning and analysis (Mullis et al., 2003). *Factual knowledge* refers to students' knowledge base of relevant science facts, information, tools, and procedures. When students solve problems and develop explanations in science, accurate and broad-based factual knowledge enables them to engage successfully in doing, understanding, and interpreting science. Therefore, factual knowledge is a prerequisite to students' in-depth learning process. *Conceptual understanding* involves perceiving the relationships between the phenomena of the physical world and drawing more abstract or more general scientific concepts from the observations. It can be measured by the way students using and applying it perform specific tasks. *Reasoning and analysis* are related to the more complex tasks occurring in unfamiliar or more complicated contexts in which students should reason from scientific principles to provide an answer. The process of

engaging with such tasks may involve a variety of approaches or strategies (Mullis et al., 2003). The details are described along with other student skills and abilities defining the cognitive domains in Table 5.3 (below).

Within the content and cognitive domains, scientific inquiry, which included knowledge, skills, and abilities as well as problem solving and inquiry tasks, was assessed overall in various content-related contexts. Identifying the impact of each of these factors is important since it can inform one of where education and learning can be improved.

Table 5.3 *Science cognitive domains and target percentage in TIMSS 2003 for Grade 8*

Domains	Main activities	Target percentage devoted
Factual knowledge	Recall/recognize, define, describe, use tools & procedures	30%
Conceptual understanding	Illustrate with examples, compare/contrast/classify, represent/model, relate, extract/apply information, find solutions, explain	35%
Reasoning and analysis	Analyze/interpret/solve problems, integrate/synthesize, hypothesize/predict, design/plan, collect/analyze/interpret data, draw conclusions, generalize, evaluate, justify	35%

Source: Adapted from Mullis et al., 2003

It should however, be noted that a large-scale international assessment like TIMSS may not cover all the content taught in science in each country. Some of the topics tested in TIMSS 2003 may be part of other curricula, such as those for geography or social studies. This was the case with South Africa and, where the topic coverage was the lowest amongst the participating countries (Martin, Mullis, Gonzalez & Chrostowski, 2004).

Using the aforementioned framework, the development of the assessment items was effected through the cooperative efforts of the NRCs and the science task

forces composed of the science coordinator and two experienced science item writers. Once the items were developed, they were reviewed by the task forces, and later by an Item Review Committee and a group of experts. Subsequently, the items reviewed were field-tested in participating countries and again reviewed by the Science and Mathematics Item Review Committee. Finally, the items were endorsed by the NRCs of the participating countries to ensure that the assessments represented the curricula of the participating countries and that the items exhibited no bias toward or against particular countries, along with an opportunity to match the content of the assessment to each specific country's curriculum (Martin, Mullis & Chrostowski, 2004).

The resulting TIMSS 2003 Grade 8 assessment contained 194 items in mathematics and 189 in science. In order to ensure broad subject-matter coverage without overburdening individual students, TIMSS used a matrix-sampling technique. Each assessment item was assigned to one of 14 mathematics or 14 science item blocks and these were distributed across 12 booklets. Each student took one booklet containing both mathematics and science items (Mullis et al., 2003). The science assessment at Grade 8 contained 109 multiple-choice and 80 constructed-response types where students were asked to generate and write their own answers. Among constructed-response questions, some asked for short answers while others required extended responses requiring students to offer explanations for their answers. Additionally, the assessment included 7 problem-solving and inquiry tasks for Grade 8, reflecting the importance placed by the assessment framework on problem-solving, reasoning and scientific inquiry (Martin, 2004).

In line with the purpose of TIMSS, the assessment encompassed items used in the 1995 and 1999 administrations to guarantee reliable measurement of trends over time. With this intention, 74 items in science and 79 items in mathematics, for both multiple-choice and constructed-response items, were trend items that had already been used in 1995 and 1999.

5.6.2 THE CONTEXTUAL QUESTIONNAIRES

TIMSS 2003 developed 11 questionnaires across the two grades and two subjects, with NRCs completing four. Grade-8 students who were tested answered questions pertaining to mathematics and science. The mathematics and science teachers of sampled students responded to questions about teaching. Questionnaires for mathematics and science teachers were administered separately at Grade 8. The principals responded to questions about schools at Grades 4 and 8 (Martin, Mullis, Gonzalez & Chrostowski, 2004). The purpose of the questionnaires was to gather information about five broad areas, viz., curriculum, school, teachers and their preparation, classroom activities and characteristics, and students at various levels of the educational system (Mullis et al., 2003). All questionnaires were based on Likert-type scales to record the self-reported information. Three questionnaires on student, science teacher, and principal were examined in the current study, and are described in more detail below.

Principal questionnaire: The school questionnaire addressed to the principal of each sampled school covered school-quality-related issues such as school organization, roles of the principal, and resources to support mathematics and science learning, parental involvement, and a disciplined school environment. Some of the main topics addressed in the school questionnaire were as follows: school climate, stability and mobility of the student body, parental involvement, professional development, instructional resources, and principal's experience (see Appendix D). The school questionnaire comprised 25 items and various sub-items that constituted item sets and was designed to be completed in about 30 minutes.

Science teacher questionnaire: The science teacher of the class tested was asked to complete a science teacher questionnaire. The questionnaire for teachers was composed of information about the classroom contexts for teaching and learning, and actually about the implemented curriculum in

science. Teacher preparation and professional development, and the use of technology were newly added to the TIMSS 2003 teacher questionnaire. The main areas included teaching experience, preparation to teach, teacher interactions, attitudes toward subject, time spent teaching subject, content-related activities, factors limiting teaching, topic coverage, homework, and assessment (see Appendix C). The science teacher questionnaire was made up of 34 items, some of which consisted of various sub-items. The teacher questionnaire was designed to be completed within 45 minutes, reflecting the greater number of items.

Student questionnaire: Each student of the class sampled for the TIMSS 2003 study was asked to complete a student questionnaire, designed on the basis of factors thought to influence student achievement in science and so focusing on home background and resources for learning, prior experiences, and attitudes toward learning. The main question areas covered language, books in the home, home possessions, parents' education, educational expectations, liking and valuing science, learning activities in science, safety in school, out-of-school activities, and extra lessons or tutoring (see Appendix B). It comprised 23 items with some items including sub-categories. The TIMSS 2003 student questionnaire was designed to take about 30 minutes to complete. Across the aforementioned questionnaires, parallel questions were used to measure the same construct from different sources.

Since the instruments were developed in English, they were translated by the participating countries into 34 languages of instruction. The full set of instruments were translated into Korean for application in Korea, while the assessment in South Africa was contextualised for South Africa, adapted to international English and also translated into Afrikaans. The IEA Secretariat in Amsterdam used a rigorous process of translation verification to ensure that instruments and questionnaires were translated accurately and were internationally comparable (Martin, Mullis & Chrostowski, 2004).

5.7 DATA ANALYSIS

Data analysis began with exploring the TIMSS data sets to preliminarily identify sets of items and single items which relate to the factors of the conceptual framework. The constructs related to effective science education had already been defined when developing the conceptual framework in the previous chapters (3 and 4). First, the item examination was paralleled by descriptive statistics to get a brief view of the two countries' sets of data and to explore them for suitability for further analyses (5.7.1), with missing data also scrutinized (5.7.2). Next, the statistical processes such as factor and reliability analysis were used to build construct validity (5.8.3).

5.7.1 EXPLORING THE DATA SETS

In order to explore the data sets, the definition of the factor was compared to the contents of the items from the TIMSS questionnaires (see Appendix B, C, and D). The items which corresponded to the definitions were selected and recoded to suit the current study. The codes were reversed when an item was negatively phrased. Corresponding to the factors of the conceptual framework, variables were renamed, labels assigned to the codes given and measurement scale allocated. Once the data were recoded and checked for errors, the file was converted into a Statistical Package for the Social Sciences (SPSS) for further analysis.

Using the SPSS programme, the descriptive statistics analysis for the items was undertaken in order to describe, organise and make understandable data for the study (Minium, King & Bear, 1993). The descriptive statistics involved the identification of the mean, standard deviation, range of scores, skew and kurtosis. Frequencies were run for the selected items and the output was examined for any missing cases and values in the data, as well as the percentage of respondents who checked each answer option.

The evaluation of descriptive statistics, like running frequencies, shows the characteristics of the sample tested, and allows the researcher to check if the data violates any assumptions underlying the statistical techniques for further analysis and addresses specific research questions (Pallant, 2007). For instance, histograms generated from the descriptive statistics can provide a visual representation of the normality of the data. Usually parametric tests have four basic assumptions to ascertain the accuracy of the tests: normally distributed data, homogeneity of variance, interval data, and independence (Field, 2005). Since this particular research involved factor analysis and multilevel analysis to answer the research questions, it is important to check if any assumptions underlying those statistical techniques were violated. At the first stage of testing the assumptions, the research searches for missing case and data. Next, the distribution of scores is explored to check the normality. Apart from the information about the distribution of variables, the descriptive statistics provided the central tendency of the data, variability around the mean, deviations from normality, the spread of the distribution and information about stability or sampling error in the data.

The distribution of scores was explored by checking the skew and kurtosis. The skew indicates the symmetry of distribution, which is whether the data is normally distributed. A positive skew has scores clustered to the left of the centre while a negative skew indicates the reverse. The kurtosis indicates peaks of distribution, with a positive value of kurtosis having the peak of distribution in the centre and a negative value indicating a flat distribution (Field, 2005). A zero value of skew and kurtosis means that the distribution is normal, which rarely happens in the social sciences. A larger sample (more than 200 cases) tends to lessen the effects of skew and kurtosis. While the skewed distributions should be transformed so that the scores are normally distributed for further analysis, checking the shape of the distribution by means of a histogram is recommended in a large sample, since the tests used for skew and kurtosis are too sensitive for a large sample (Pallant, 2007). It should be borne in mind that violation of the assumption of normality is common in a larger

sample, and skewed distributions reflect the underlying nature of the construct being measured, not a problem with the scale (Field, 2005).

The exploration of the data set also identifies outliers, cases with scores well above or well below the majority of other cases. Since the outliers influence mean and standard deviation as well as distribution (Field, 2005), there is a need to decide how to deal with the outliers by removing, transforming, or changing the value. If the mean and the trimmed mean values are very similar, the decision could be taken to include the outlier. If the outliers identified are the main cause of the skewed distribution, the transformation can reduce the impact as described above. Change to the scores, if transformation fails, can be made by the next highest score plus one, converting back from a z-score, or the mean plus two standard deviations (Field, 2005).

5.7.2 MISSING DATA

Generally, research shows three types of missing data, viz., missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Croninger & Douglas, 2005). It is understood that in a large-scale study like TIMSS, the pattern tends to include all these kinds of missing data. It was reported that there are many factors influencing the relative performance of most missing data procedures: sample size, number of variables missing, mechanism of missing data, proportion of missing data, average inter-correlation among variables, characteristics of the variables, and psychometric properties of the measures. Despite all these factors, the proportion and pattern of missing data are most likely to influence the relative performance of missing data procedures (Dodeen, 2003).

Some methods may be employed to deal with missing data in analysis. SPSS has two methods to deal with it, the *listwise method*, which deletes any case that has missing values and accordingly, it results in a loss of sample and statistical power; and *pairwise method*, which uses all the data available in an

analysis and deletes the specific missing values from the analysis, generating different sample sizes for each parameter. Accordingly, it is practical when the sample size is small or missing values are large (Croninger & Douglas, 2005).

Another way to deal with missing values or data is to replace them through *imputation*, which includes *mean substitution*. Dodeen (2003) documented that valid mean substitution was more effective than multiple regression replacement in terms of producing parameters like R^2, the coefficient of determination, or F value. The favoured type of imputation is an estimation method such as *Full-Information Maximum Likelihood*, which is considered superior when missing data is non-random.

The study excluded missing cases prior to dealing with missing data (values) as the two countries had a large enough sample size, even after removing them. Thereafter missing data in each of the remaining cases was taken care of, being replaced by mean or median, given that the sample sizes in question were large in contrast to the amount of missing data at each level, and not so serious. It should be noted that this way is a very traditional approach, although it is documented that it would be acceptable to consult other sources of secondary analysis (Bos, 2002; Howie, 2002). If more than 5% of the data was missing for an item which seemed important for analysis, then it was replaced using the mode, mean or median. The mean was used where the distribution of frequency was not skewed, the median where the distribution of frequency was skewed, and the mode was used to replace missing data specifically for yes-no format items (Allison, 2002; O'Rourke, 2003; McKnight, McKnight, Sidani & Figueredo, 2007).

Once all missing data was replaced, as explained above, frequencies were run again and finally reviewed before proceeding, in order to ensure that the data was ready for further analysis and to construct scale scores.

5.7.3 Constructing scale scores and variables

A good instrument depends on internal consistency and unidimensionality of items constituting scales in nature (Gardner, 1995). Whereas internal consistency is commonly determined by calculating Cronbach alpha as put forward above, the unidemensionality of scales can be tested using a statistical technique such as factor analysis (Osborne et al., 2003). The study calculated reliability of a set consisting of more than three items in order to find internally consistent items. Once the sets were satisfied with reliability criterion (alpha=0.5), those items were then examined along with the results of factor analysis. Items finally extracted from the analyses were summed to make up a scale. The details are discussed as follows.

5.7.3.1 Factor analysis

Factor analysis is used to determine the underlying conceptual structure in a set of items (Coolidge, 2000). Since it is concerned with grouping together items that have the same construct, it can help researchers reduce a set of items to a smaller number of underlying factors, form a conceptually understood set of data, and ultimately ensure construct validity of the research (Cohen et al., 2007). There are two main forms of factor analyses, namely *exploratory* and *confirmatory*. In this study, exploratory factor analysis, also referred to as 'principal component analysis', was used, and involved exploring previously unidentified groupings of variables for underlying patterns (Cohen et al., 2007). The factor analysis (Devellis, 1991) was carried out in the following steps:

- It determined whether a set of items were suitable for factor analysis by investigating sample size and the strength of inter-item correlation. Correlation matrix from items was constructed in order to examine pure item homogeneity. The inter-item correlation for the optimal level of homogeneity should range from 0.2 to 0.4 (Briggs & Cheek, 1986). The research adopted the two other measures generated by SPSS, which are

the Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity. The KMO index ranging from 0 to 1 should be at least 0.6 or above, and Barlett's Test of Sphericity should be significant (p<.05) (Pallant, 2007).

- Once the matrix was established, latent variables were identified by means of factor extraction, which explained the patterns of co-variation among items. The method of factor extraction refers to such different procedures as principal component analysis, principal axis factoring, and maximum likelihood. The current research used principal component analysis, which is the most commonly used analysis procedure. Research has shown that different procedures tend to yield similar solutions, regardless of which are used (Briggs & Cheek, 1986).

- First factors related to the most shared co-variation among items that best account for the total variance amongst the entire set were identified. Successively, other factors with the next most remaining co-variation amongst items were identified. The second factor is likely to account for less variance than the first. This process was continued as far as factors classified in the model were met. Factor loadings, which represent correlations between each item and a factor, were also generated and examined (Kline, 1993). The value of loading ranges from +1.00 to -1.00 and the higher absolute value indicates the stronger relationship (Crowl, 1986). For the purposes of this research, loadings of 0.3 and above were considered as acceptable (Kline, 1993).

- Even though factors are extracted as explained above, they are still arbitrary. By performing a factor rotation, one can make the picture of the relationships among the items simpler and clearer. Factor rotation identifies items with high factor loadings on one factor but low on the others, and draws a meaningful and understandable factor structure. There are various methods to rotate factors, such as *Varimax,* which involves the factors being orthogonal or independent, and *Direct Oblimin* and *Promax* which allow factors to correlate. Orthogonal rotation with *Varimax* rotation, where

the independence of factors is sustained, was practised for simplicity in the study. *Varimax* rotation is considered useful to maximize the variance between factors and thus more likely to distinguish from each other (Cohen et al., 2007). As any factors to emerge would presumably be somewhat correlated, an oblique approach like *Direct Oblimin* would be more appropriate. As mentioned with the extraction method, the rotation method did not change the results in any meaningful way, regardless of variations (Briggs & Cheek, 1986).

- Finally, an approach known as *Kaiser's criterion* used the eigenvalue rule and scree test techniques to confirm the proper number of factors to retain. A minimum eigenvalue of 1 was utilized while *Catell's Scree* test was used and, as Catell recommends, all factors above the elbow or break in the plot are retained (Pallant, 2007). Additionally, parallel analysis could be used to compare the size of the eigenvalues with those derived from a randomly generated data set of the same size. Only those eigenvalues that exceed the corresponding values from the random data set are retained. Eigenvalues indicate how much variance a factor accounts for in terms of the average original variable. An eigenvalue of 1.0 indicates that a factor accounts for as much of the variance as the average original variable. However, since an eigenvalue greater than 1.0 is likely to result in overestimating the number of underlying factors, researchers tend to reject this procedure (Briggs & Cheek, 1986).

As explored up to this point, the study examined the internal structure of the many items of the TIMSS data of Korea and South Africa. As factor analyses identified latent variables underlying a set of items offered in TIMSS, and substantive meaning of the latent variables (DeVellis, 1991), the different results between the two countries mean underlying patterns on the variables sought are different (Cohen et al., 2007).

5.7.3.2 Reliability analysis

Reliability is concerned with consistency of scale, which is also referred to as stability and equivalence, over time, over samples, and over forms. As many items constitute a scale in the study, it assessed in particular the internal consistency of the scale prior to further analysis being made. Once items were confirmed as suitable constructs for the research by means of factor analysis, and problems identified as well as rectified where possible, the reliability analysis was carried out to examine internal consistency of the remaining items that made up the scales.

The degree of internal consistency reliability was calculated by Cronbach's coefficient alpha (α), which is most widely used for items that are not answered 'right' or 'wrong' but with a range of possible options (McMillan & Schumacher, 2006). The reliability coefficient ranges from 0.00 to 1, but where the coefficient of a scale is high, the scale is highly reliable, and vice versa. For the most part, 0.70 to 0.90 is acceptable (McMillan & Schumacher, 2006). DeVillis (1991) suggests a scale of 0.65 for questionnaire data. However, as this is an exploratory study, a coefficient as low as 0.5 for the questionnaire is considered acceptable (Bos, 2002; Howie, 2002). For achievement data, however, a coefficient above 0.8 is preferable (Kline, 1993).

Apart from considering Cronbach alpha values above 0.5, the mean inter-item correlation and the correlation between each item and the total score were also examined as another means of item homogeneity. A high item-total correlation would be expected if items measure the same construct, which then would contribute to the total score of a test (Kline, 1993). Furthermore, where the items comprising a scale have a strong relationship to a latent variable, they are likely to have a strong relationship within themselves as well (DeVellis, 1991). Therefore, high correlations between the items reflect strong links between the items and the latent variable and indirectly imply the internal consistency of the factor.

In the case where the number of items that make up a scale is as small as less than 10, the mean inter-item correlation for the items can be calculated and reported. The mean inter-item correlation values ranging from 0.2 to 0.4 were acceptable for the study (Briggs & Cheek, 1986). When a set of items has a correlation value of less than 0.25 between items and the total score, they were excluded from the research.

5.8 CORRELATION ANALYSIS

Once factor and reliability analyses confirmed the items are uni-dimensional and internally consistent, the scores were added together to make scales, and variable names and labels were assigned for further analysis. Thereafter, correlation analysis was undertaken to ascertain the relationship between the scales or factors identified. First, preliminary examination was made to ensure no violation of the assumptions of normality, linearity and homoscedasticity[9] occurred but not in a strict way as the study involves large sample and regression analyses in nature. Next, bivariate correlations calculated for the scales constructed using the items from the questionnaire and science achievement. The inter-correlations between the scales were also examined to ensure that multicollinearity[10] is not present in the data.

The study calculated a correlation coefficient, the bivariate Pearson product-moment correlation coefficient γ. The product-moment correlation has been known to be appropriate when both variables have continuous scales as in achievement tests or self-concept inventories (McMillan & Schumacher, 2006). Correlation coefficients range from -1.00 to +1.00 and indicate the strength and direction of a relationship. A plus sign indicates a positive relationship and a minus sign a negative one. Where the coefficient is below plus or minus 0.35,

---

[9] Homoscedasticity indicates that the variance of the error terms for the independent variable is constant and one of assumptions in regression analysis (Miles & Shevlin, 2001).
[10] Multicollinearity exists where independent variables are highly correlated in regression analysis, (Mendenhall, & Sincich, 1996).

the relationship is low and an inference that the variables are not related can be drawn. Coefficients between plus or minus 0.35 and 0.65 indicate the variables are moderately related. When the coefficient is higher than plus or minus 0.65, the variables are highly related (Gay & Airasian, 2003).

The study adopted a correlation coefficient of an absolute value above 0.2 and the significance level, 0.01 (0.99, confidence interval) as criterion to include the scales for further analysis. The criterion for cut-off seems low, a slight relationship, considering the strength of a relationship to coefficient value described above. Nonetheless, when correlations are ranging from 0.20 to 0.35, and if the number of cases is more than 100, it may be statistically significant and valuable enough to explore the interconnection of variables in particular in explanatory studies such as this (Cohen et al., 2007; Cresswell, 2008). As for the significance level, the level of statistical significance of a correlation tends to depend largely on the sample size. The greater the sample size, the smaller the correlation needs to be in order to be significant at a given level of confidence (Cohen et al., 2007).

In some instances, variables were constructed from a number of items as a result of factor analyses and reliability analyses. However, in other instances and based on literature, single items were used as a variable, such as level of education of mothers and fathers. Once it is considered that the items make sense conceptually, those items were analyzed by correlation analysis as well. Although correlation analysis does not guarantee causal relationships, it enables one to preliminarily identify the causes of important educational outcomes and to predict the score on a dependent variable (McMillan & Schumacher, 2006).

In addition to correlation coefficients, the coefficient of determination and the significance level could be investigated through regression approach in the correlation analysis. The coefficient of determination can be calculated by squaring and multiplying the γ value by 100 to make a change into percentage

of variance. It represents how much variance is shared. A correlation of 0.2 means that only 4% of the variance is shared, but it can not be ignored in large-sampled and exploratory studies (Cohen et al., 2007).

5.9 MULTILEVEL ANALYSIS

Multilevel analyses refers to any analysis that involves data sets with a nesting structure, such as students in classes, classes in schools, or schools in districts (Snijders & Bosker, 1999). Multilevel modelling has a hierarchical data set collected at all existing levels, but with one single outcome that is measured at the lowest level, namely student level. The term 'multilevel regression model', dealing with such multilevel data sets, is used interchangeably with 'random coefficient model', 'hierarchical linear model', or 'variance component model' (Hox, 2002).

Since TIMSS collected data in a multilevel structure, viz., student, classroom, and school, and the intact class in a school was sampled to allow data to be collected in a natural situation, effects of both individual and group level variables need to be taken into account (Keeves & Sellin, 1997). Multilevel analysis is recommended if research is to focus on correlations between levels as well as within levels, particularly as single level analysis dealing with aggregating data fails to explain within and across-level interaction or relation (Kyriakides & Charalambous, 2005). Kyriakides and Charalambous's comparison between findings of single-level analysis and multilevel analysis into TIMSS 1999 data strongly supports a multilevel approach in analyzing the data of the IEA studies consisting of hierarchical structures.

5.9.1 CHARACTERISTICS OF MULTILEVEL ANALYSIS

The analysis of data, which is structured at several levels, is concerned with compositional effects across levels and takes account of grouping effects (Fitz-

Gibbon, 1996). More specifically, students are grouped in a classroom and again the classroom is nested within the next higher level, the school. Variables that influence student achievement exist at the student level, classroom, and school level respectively. When considering the structure of hierarchy in the data collected, students within a class tend to be more alike than those from different classes, and the same holds for the school level.

In terms of the variance effects, multilevel analysis makes it possible to understand where and how it occurs because it deals simultaneously with the variance components at all levels (Rasbash, Steele, Browne & Goldstein, 2009). It is considered that the ability to estimate between-group variation in an attempt to explain variation is a great strength of multilevel modelling. Ultimately, multilevel analysis is the way to discover the inference made about the variance among all schools, using the schools sampled. As a result, researchers can explain the pattern of variance occurring across the schools of the population in question (Rasbash et al., 2009).

In contrast, linear models used previously deal separately with the variance at each level of the hierarchy and cause problems such as aggregation bias or misestimated precision (Raudenbush & Bryk, 1997). Furthermore, ignoring clustering causes other technical problems, such as the underestimation of standard error of regression coefficient, which makes an incorrect inference about the effect of higher-level explanatory variables by interpreting the effect as being significant when not so (Rasbash et al., 2009).

When considering hierarchy of data collected, the interaction between variables characterizing individuals and variables characterizing groups should also be considered in research which is involved in individuals' achievements (Hox, 2002). Multilevel analysis enables the researcher to investigate the interaction between factors within each level and interaction between levels. Accordingly, multilevel analyses provide researchers with a picture of the variance in achievement in the whole system and of the factors affecting it.

There are several concepts in multilevel analysis that should be kept in mind. The first concept to note is intra-class correlation, which measures the extent to which the achievements of students in the same school resemble each other as compared to those from students in different schools. The intra-class correlation, which indicates the similarity between students in the same school, can be measured by the proportion of school[11] level variance compared to the total residual variation that is attributed to differences between schools (Hox, 2002; Rasbash et al., 2009).

The second concept to consider is random and fixed coefficients that show up as parameters in the multilevel regression equation. Random coefficients operate as a probability function varying within a level in the regression equations. It includes intercept and slope coefficients. Fixed coefficients show up as regression coefficients that are deterministic in regression equations (Hox, 2002). Because fixed coefficients apply within-level, they are not assumed to vary across within-level. Multilevel analysis determines random and fixed coefficients in the regression equation along with residual errors to explain the variance between and within levels.

The third aspect to look at is cross-level interactions that involve interactions between explanatory variables from different levels. The interaction effects in the multilevel equations are formed by multiplying the scores for the variables from the different levels.

Finally, there are various estimation methods when estimating parameters in the multilevel regression equations. The techniques used include Maximum likelihood (ML), Generalized Least Squares, Generalized Estimating Equations, Bootstrapping, and Bayesian methods (Hox, 2002). ML estimates of the population parameters that maximize the probability of observing the actual data are mostly used because they are robust against mild violations of the assumptions, such as having non-normal errors. The ML method has two

---

[11] TIMSS 2003 sampled only one class per school

different functions, such as Full Maximum Likelihood (FML) and Restricted Maximum Likelihood (RML). The current study used FML because it has some advantages over RML, particularly as computing the ML estimates is easier in FML. In addition, the regression coefficients are included in the likelihood function and thus overall the chi-square test based on the likelihood makes it possible to compare two models with different regression coefficients, whilst RML allows one to compare only differences in the variance components (Hox, 2002).

5.9.2 BUILDING THE MULTILEVEL REGRESSION MODEL

Once the factors were confirmed through factor, reliability, and correlation analyses, along with single items, these were used for further analyses, namely multilevel analyses. The factors influencing science achievement in the two countries are different, as shown in the results, and thus multilevel analysis used the different sets of selected latent variables. It is expected that estimating the pattern of variation in the underlying population of the two countries becomes possible, enabling the researcher to explain the pattern in terms of the general characteristics of schools in the two countries.

TIMSS sampled one class per school and although more than one teacher tends to teach one class in Korea, the data from class level and school level cannot be differentiated, unlike the case of more than two classes under a school from a perspective of multilevel analysis. Therefore, the current study modelled the data in a two-level structure, viz., student, and classroom/school level. The two-level model distinguishes the variance specifically explained at the student level and then the variance accounted for at the classroom/school level in light of science achievement, together with the interaction between the two levels.

The model was built starting from the intercept only or null model. The detailed procedures can be summarized as follows:

- Step 1 – building a null model (the intercept-only model) to estimate the total variance. Because there are no explanatory variables in the intercept-only model, random effects depending on the residual variances represent unexplained error variance (Hox, 2002). The intercept-only model gives an estimate of the intra-class correlation ρ and a benchmark value of the deviance, which is a measure of the degree of misfit between the model and the data. The equation of the model is as follows:

  $Yij = \gamma_{00} + u_{0j} + e_{ij}$

  Y_{ij} = dependent variable, science achievement in TIMSS 2003 in this case

  γ_{00} = intercept or regression coefficients, the expected value of the outcome variable when all explanatory variables have the value zero.

  u_{0j} = residual error at the classroom/school level

  e_{ij} = residual error at the student level (Hox, 2002)

- Step 2 – building a lower-level, student-level, model and adding a predictor to the null model one-by-one to examine the deviation in each case. Once the deviations produced by each model have been identified, the researcher can rank all variables in order of largest to smallest in deviation. That order is the reference when the individual variables are entered into the model as the equation of the model is built up extensively. Entering individual variables into the model, by the so-called 'step up method', 'step-by-step', or 'forward steps upward from level-1 method'. When a variable added resulted in a significant effect, it was kept in the model. To evaluate whether a variable is significant or not, the Wald test referred to as the Z-test and was conducted and any change in the deviance was examined by making use of Chi-square if the variable contributes to the model (Hox, 2002). In this step, the improvement of the final model with all lower-level significant explanatory variables can be tested by computing the deviance gap between the final model of the lower-level and the null model. The

equation of the model with student-level explanatory variables can be written as:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + u_{0j} + e_{ij}$$

Where:

X_{pij} = the first-level explanatory variables

Subscript p = explanatory variables at the student level

- Step 3 – building a higher-level, classroom/school, model. All explanatory variables from the lower level to the higher-level were entered into the model. This allows one to examine whether the group-level explanatory variables explain between group variations in the dependent variable. In this step, one can test the improvement of the final model with all lower-level and higher-level explanatory variables significant by computing the difference of the deviance between the final model of the lower level and the final model of the higher level just as in the previous step. The models in steps 2 and 3 are called variance component models since the residual variance is divided into components corresponding to each level in the hierarchy (Hox, 2002; Rasbash et al., 2009). Variance component models assume the fixed regression slopes and the random regression intercept (Hox, 2002). The variance component model with classroom/school-level explanatory variables can be written as:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}Z_{qj} + u_{0j} + e_{ij}$$

Where:

Z_{qj} = the classroom/school-level explanatory variables

Subscript q = explanatory variables at the classroom/school level

- Step 4 – building the full model by putting all the variables identified as significant into the model. The full model can be formulated by adding

cross-level interactions between explanatory group level variables and those individual level explanatory variables that had significant slope variation above. The model built for the full steps is as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}Z_{qj} + \gamma_{pq}Z_{qj}X_{pij} + u_{pj}X_{pij} + u_{0j} + e_{ij}$$

Where:

$Z_{qj}X_{pij}$ = cross-level interaction term

u_{pj} = the classroom/school-level residual of the slopes of the student-level explanatory variables X_{pij}

The researcher started with fixed regression coefficients, as fixed parameters are more likely to be estimated with much more precision than random parameters (Hox, 2002). The random coefficient model can be built to see whether there exist the slopes of explanatory variables of which variance between the groups is significant. Testing for random slope variation on the basis of variable-by-variable might lead to an explanatory variable having no significant average regression slope but having a significant variance component in random coefficient model. After each process of adding explanatory variables, parameters added were examined to see if they are significant, as were the residual errors.

Estimation of parameters, including regression coefficients and variance components in the multilevel models, was mostly made by using the Full Maximum Likelihood (FML) method, referred to as Iterative Generalized Least Square (IGLS) in MLwiN (Hox, 1995). As put forward above, FML is preferred since IGLS is faster and numerically more stable, and the overall chi-square test based on the likelihood makes it possible to compare two models with different regression coefficients in FML and formally test the improvement of fit (Hox, 1995; Hox, 2002). Based on the results of FML, the decision was made as to which should be included in the model based on significance tests, the change in deviance and change in variance components (Hox, 2002).

5.9.3 PROGRAMMES OF THE MULTILEVEL REGRESSION MODEL

There are several kinds of programmes used for analyzing the multilevel regression model including HLM, VARCL, and MLwiN. HLM is considered the easiest to use and the output contains the parameter estimates, their standard errors, the covariance at the two levels, and the deviance. HLM provides p-value as an indicator for their significance. In contrast, VARCL does not provide p-value although using FML, comparing the deviance of different models, or inspecting the estimates and standard errors of various coefficients in one specific model. Hence, it should be computed outside the programme (Hox, 1995).

In addition to the various characteristics featured above, MLwiN contains more build-in provisions and is considered more difficult as such. MLwiN uses the single equation representation when the multilevel models are formulated while the software HLM specifies the separate equations at each available level (Hox, 1995). The single-equation formulation makes the effect of cross-level interactions clear. On the other hand, the single-equation representation hides the effects of the complicated error components as multilevel models have different slopes (Hox, 2002).

For the purpose of this research MLwiN, software developed by the Centre for Multilevel Modelling in the UK, was used. MLwiN has some interesting features. In terms of workplace, the programme has a graphic interface with plotting, diagnostics and data manipulation facilities. Besides, it is spreadsheet-typed which consists of columns and rows (Rasbash et al., 2009). On the other hand, it makes it possible an analysis of non-standard as well as standard multilevel models by allowing all regression coefficients to be random at all levels. In addition, researchers can analyse data with arbitrary levels and estimate FLM and RLM by MLwiN. Furthermore, MLwiN allows researchers to make repetitive computations and the use of residuals derived from analysis for another model.

Accordingly, it is a user-friendly help system in terms of data computations and manipulation. It however should be noted that MLwiN does not handle missing values and one has to deal with them as described in advance before importing them into the programme.

5.10 METHODOLOGICAL NORMS

To confirm the quality of the data and improve the generalisability of the results collected in survey research, reliability and validity need to be achieved. These can be explicated into several kinds respectively, depending on the goal of the research. In this study they were as follows:

5.10.1 VALIDITY CONSIDERATIONS FOR THE STUDY

An assessment's validity is the extent to which it measures what it claims to measure (Goldstein, 1993), property obtained at the end of research. Validity as a property can be expressed by degree (high, moderate or low) and inferred from evidence. Therefore, validity, which is mainly referred to as construct validity, is inextricably linked with the consequences of research involved in assessment or questionnaires. On the other hand, validity can be negatively influenced by inadequate sampling or administration and poorly-constructed items (Gronlund, 1998; Linn & Gronlund, 2000). Therefore, validity should be ensured in all areas of research.

There are different facets of validity which form part of the unitary term 'validity', such as content-related validity, construct-related validity, and predictive validity. TIMSS ensured in particular the content-related validity of instruments, which included face and content validity in the process of designing instruments (see Chapter 2). Despite various aspects of validity, Messick (1981) argues that construct validity takes precedence over other validities from both a scientific and applied point of view in education and psychology. Construct validity was

addressed quantitatively by using inferential statistics such as factor analysis and reliability analysis (Suen, 1990). Factor analysis, in particular, is one of the most useful methods for studying and validating the internal structure of instruments (Schönrock-Adema, Heijne-Penninga, van Hell & Cohen-Schotanus, 2009). If the measurement of a scale is taken as measuring what it is supposed to measure, then the variance would be accounted for by a loading on a single factor (Osborne et al., 2003).

The current study focused specifically on construct validity with respect to the questionnaires by undertaking factor analyses and reliability analyses. The scores on the scales were grouped by the same construct, as items were clustered according to the conceptual framework for the study. From a perspective of the conceptual framework underlying the study, if some variables have to do with other constructs, the scales to measure those constructs can also be expected to have a similar bearing on the same constructs.

Construct validity ultimately leads to validity of inference and the consequence of the study (McMillan & Schumacher, 2006). Given that this research was based on an adapted conceptual framework, construct validity supported by the empirical evidence is important in interpreting the consequence of the research by using the conceptual framework as a lens.

5.10.2 RELIABILITY CONSIDERATIONS FOR THE STUDY

Reliability is a measure for the consistency of instruments (Cohen et al., 2007), involving *stability,* which indicates a consistent measure over time and over similar samples, and *equivalence,* which is the consistency of the results through similar design or researchers (Cohen et al., 2007). *Internal consistency* denotes the homogeneity of the items, that is, the degree to which those that make up a scale all measure the same underlying attribute (DeVellis, 1991).

There are several ways to evaluate reliability of measurements, including test-retest reliability, split-half reliability, and internal consistency reliability. Initially TIMSS enforced test reliability by using a matrix-sampling technique, ensured test-retest reliability by TIMSS 2003, including items used in the 1995 and 1999 assessments, and inter-rater reliability when scoring the constructed responses at the data collection stage.

The study stressed internal consistency as it is useful for multi-item scales and thus considered a pre-requisite for construct validity to be established in building a scale based on multiple-items. Cronbach's coefficient alpha (α) is the most commonly used statistic for internal consistency reliability (Litwin, 1995), and is discussed further under the data analysis sections with the criterion of reliability (alpha=0.5) that are applied in the study (Howie, 2002).

5.11 ETHICAL CONSIDERATIONS

TIMSS 2003 makes available data from over 360,000 students, approximately 25,000 teachers, approximately 12,000 school principals, and the NRCs of each country, which aims at improving mathematics and science education by means of secondary analyses of the data. As part of the ethical considerations of the IEA, NRCs were requested to obtain permission from the respective Ministries of Education and from the schools and other stakeholders to release the data from all participating countries (Martin, 2005). This was done, and permission from the stakeholders was received. As part of the informed consent, anonymity and confidentiality of participants were guaranteed through the whole research process. Normally, as part of secondary analysis, free and informed consent is required to conduct a secondary data analysis. However, as the secondary analysis suggested here falls within the scope of the original consent, this is not deemed necessary.

5.12 CONCLUSION

In this chapter, the information concerning the research design and method was detailed. The intention of the current research was to explore factors influencing science achievement in two countries using quantitative data. Post-positivism grounded the current research, given that it researches the characteristics of relationships in educational contexts other than physical environment. The research also was categorized within secondary analysis in terms of method. The research used the TIMSS 2003 data set that collected by the IEA. The secondary analysis using the TIMSS data was recommended, considering that the data is of high quality, and researchers can save time and cost.

A description of the design issue, such as sampling and data collection, was described briefly and the instruments and methodological norms examined. Aspects of sampling, data collection, instrument, and methodological norms were explored, mainly consulting IEA's report on TIMSS, since the research is secondary analysis.

Data analysis strategies also were discussed. Firstly, the contents of the items from the TIMSS background questionnaires were explored to see the brief pictures in science education in the two countries. Corresponding to the factors of the conceptual framework, variables were identified, labels renamed to the codes given, and measurement scale assigned. Once the data were recoded and checked for any errors, the descriptive statistics was carried out by running SPSS. In particular, frequencies were run for the selected items and the output was examined for any missing cases and values in the data, as well as the percentage of learners who checked each answer option.

Factor analysis was undertaken of the items identified above that comprised sets of items. Extraction of factors made it possible to identify latent variables that can explain the patterns of co-variation among items. Thereafter, performing a factor rotation made the picture of the relationships among the

items much simpler and clearer. In order to confirm the proper number of factors to retain, as Kaiser' criterion the eigenvalue rule and scree test techniques were adopted. Besides factor and reliability analyses were undertaken to confirm whether the items can form the basis for the constructs or variables to be used in the further analyses or not. Correlation analyses followed them to see if items selected or scales made have a significant relationship with achievement.

With factors confirmed for further analysis through factor, reliability, and correlation analyses, the researcher carried out multilevel analyses that involve data sets with a nesting structure such as students in classes. Multilevel modelling can be adopted in the case of a hierarchical data set collected at all existing levels but with one single outcome at the lowest level. It is recommended to undertake multilevel analysis using IEA studies such as TIMSS due to their hierarchically-structured data. In the current study, multilevel analysis was carried out into the two different sets of selected latent variables, since the research showed that the factors influencing science achievement in the two countries are different.

How to build a two-level model, viz., student, and classroom/school level, was elaborated on. The null model or intercept only model which does not include any explanatory variables was explained. Thereafter, how variance component models were established at the lower and higher level was elaborated on. Lastly, the full model was described, including adding cross-level interactions between explanatory group level variables and those individual level explanatory variables. MLwiN, which was used in the research, was discussed in addition to the ethical considerations.