Education. 11 (5): pp. 1157-1181.

DESIGN, DEVELOPMENT AND VALIDATION OF A MODEL OF PROBLEM SOLVING FOR EGYPTIAN SCIENCE CLASSES

ABSTRACT. Educators and policymakers envision the future of education in Egypt as one in which schools provide high-quality education for every learner. The learning environment should be based on active learning to enable learners to acquire self-learning, critical thinking, scientific inquiry and problem solving skills. In this article, we describe the validation of a model for problem solving and the design of instruments for evaluating new teaching methods in Egyptian science classes. The instruments were based on an established model for problem solving and were designed to assess 7th grade students' problem solving, experimental strategy knowledge, achievement and motivation towards science. The test for assessing students' knowledge has been developed based on the topic, density and buoyancy. which will be taught in 7th grade in a later intervention study. The instruments were partly self-developed and partly adapted from newly performed studies on strategy knowledge and problem solving in Germany. All instruments were translated into Arabic; the translation process and quality control is described. In order to determine the quality of the instruments, 44 students in Egypt completed the questionnaires and tests. The study's aim to develop and validate the instruments did require an ad-hoc and typical sample which was drawn from an accessible population. Accordingly, the characteristics of the sample are described. Data were analysed according to the classical test theory, but to underpin the results, the instruments were additionally analysed using the even stronger Rasch model. The findings demonstrated the reliability of the items and aspects of validity. In addition, this study showed how test items can be successfully developed and adapted in an international study and applied in a different language.

KEY WORDS: scientific inquiry, strategy knowledge, problem solving, motivation, reliability, validity

INTRODUCTION

Students in Egypt are not achieving at a high level in science, for example, in TIMSS 2003 and 2007, the average science score of Egyptian 8th grade students was 421 and 408, respectively, which is significantly less than the international means of 474 and 500 (Martin, Mullis, Gonzalez & Chrostowski, 2004; Mullis, Martin, Robitaille & Foy, 2009). Based on the TIMSS 2007 results, Egypt ranked 41st out of the 59 participating countries (Mullis et al., 2009). For many years, educators and researchers have debated how to improve student achievement - with a focus on which school variables influence student outcomes. As policymakers become more involved in school reform, this question takes on new importance because many initiatives rely on the presumed relationships between various education-related factors and learning outcomes (Darling-Hammond, 2000). Some studies have suggested that low student achievement scores are related to student effort, social context and the role of teachers with regard to their expectations, efficacy and support, and structure of activities (Roderick & Engel, 2001; Ronald, 2009). In this respect, the present study assumes that teachers' instructional methods affect student performance (Good & Brophy, 2008).

In addition, the Ministry of Education in Egypt found that secondary school students enrol in the liberal arts increasingly more than in the sciences (UNESCO, 2008). Newburghl (2008) states that the loss of science students, especially in physics, is an alarming trend and is an extremely relevant issue for developing countries where significant changes are taking place in science programs offered to young citizens by focusing on new teaching approaches such as student-centered instruction. Therefore, making science compulsory in schools could increase enrolments**;** in addition**,** reforming science teaching and identifying new teaching approaches are needed to motivate students in earlier stages of schooling towards the study of science (Ministry of Education (MOE), 2008; 2010).

In Egypt, traditional patterns of science education have less emphasis on laboratory and practical experience than acquiring content knowledge but these patterns are changing. New programs and systems with more emphasis on scientific inquiry and problem solving processes are being introduced but reliable and valid models of instruction have not been developed or evaluated. Nevertheless, the primary goal of the new generation science programs focus on active teaching and the national interest in having a scientifically literate population (Hassan, 1997).

The amount of scientific knowledge has increased exponentially over the past several decades (Bloom, 2006). During this period, there has been a rethinking by philosophers of science about the nature of scientific inquiry (SI) (i.e., the practices and process of the growth of scientific knowledge) (Duschl & Hamilton, 2011). Now, most curricular programs are moving away from isolated skills to the integration of several different aspects of inquiry skills such as hypothesis generation and evaluation (Jeong & Songer, 2008). Instruction used to focus either on content knowledge or process skills but in recent years, in many countries, the development of content knowledge and inquiry skills are being increasingly addressed simultaneously (e.g., Kim & Hannafin, 2011; Tang, Coffey, Elby & Levin, 2009) by emphasizing student questioning, observing, inferring, classifying, interpreting, analysing, predicting, investigating, and problem solving (Lederman & Lederman, 2012; Quintana et al., 2004).

SCIENCE TEACHING IN EGYPT

The education system in Egypt includes Basic Education, General Secondary Education, Technical Secondary Education, and University and Higher Education (OECD, 2010). The average class size for all elementary, middle and secondary schools in Egypt is 39.65 students

(Ministry of Education-Egypt (MOE), 2008). The primary language of instruction in Egyptian schools is Arabic; students attend three types of schools− boys only, girls only and co-education.

In Egypt, the common method of science education in all grade levels from 1 to 12 can be described as one of explanation and lecturing. In science class, the teacher presents examples of concepts, writes the rules, generalizations and characters related to these concepts, gives students time to answer some questions individually to evaluate them, and finally he or she draws conclusions on the blackboard. The teacher may also use some concrete materials to explain specific content features of the lesson. One can say that, typically, the teacher acts and speaks much more than his or her students (Hashimoto, Pillay & Hudson, 2008; UNESCO, 2008). Since Egyptian society needs current and future generations of students to be capable of dealing with the requirements of "new knowledge", the Ministry intends that the following skills, knowledge and values be improved -- advanced knowledge, self-learning, citizenship values, dialogue and acceptance of others, enlightened morals, religious values and social cohesion, problem solving, creativity and scientific inquiry (Ministry of Education (MOE), 2008). In order to foster these skills, the Ministry is following several international trends and is pursuing teaching methods that are considered the basis for improving education and its outcomes. These trends and methods can be summarized as learner-oriented instruction – including problem-solving, the encouragement of learning through exploration, linking education and learning to real-life contexts, continual meditation and reflection in educational processes and outcomes, and continuous and comprehensive evaluation of the learner's performance (see Ministry of Education (MOE), 2008; 2010; UNESCO, 2008).

THEORETICAL BACKGROUND

The national science education standards in many countries (Bernholt, Neumann, & Nentwig, 2012) recommend teachers focus on inquiry, predominantly on real phenomena in classrooms, outdoors, or in laboratory settings, where students are given investigations or are guided toward performing investigations that are demanding but within their capabilities. Although it is true that inquiry requires new ways of thinking, developing these habits of mind, especially through experimental experiences, is not yet standard practice in many science classrooms in countries with highly developed school systems, and certainly not in Egypt. In part, this is because teachers lack guidance in designing investigations in ways that facilitate the conditions for students to practice and learn to inquire, and develop their critical thinking about evidence and make discoveries (Sutman, Schmuckler & Woodfieldand, 2008; Zimmerman, 2000).

To enable students to engage in scientific inquiry processes in science classes, student-oriented instruction is needed (Hofstein & Kind, 2012). Therefore, inquiry learning should be structured in such a way that students are provided with information related to a certain problem situation and encouraged and guided to plan, conduct, and evaluate their own investigation (Berg, Bergendahl, Lundberg, & Tibell, 2003). In a country like Egypt, which is at the beginning of developing more student-oriented teaching and learning models, guidance for both students and teachers is an important part of the implementation. In investigating scientific inquiry and problem solving processes, researchers have operationised these concepts in various ways (e.g., Dogru, 2008; Dunbar & Klahr, 1988; Klos et al., 2008; Kneeland, 1999; Oser & Baeriswyl, 2001). In all those models, the step *identifying an idea* is used for solving a problem. Other steps occur only in some of those models, and these steps are *finding a hypothesis* based on prior knowledge, *conducting an experiment*, and *evaluating the results* (e.g., Dunbar & Kahr, 1988; Klos et al., 2008). In other models, *problem presentation*, the *discovery of a problem*, *reformulation of the* *problem task*, and *exploring a possible way of solving the problem* are proposed (e.g., Oser & Baeriswyl, 2001). However, in other models the steps previous knowledge and calculating data were missing (e.g., Dogru, 2008; Kneeland, 1999). To meet the situation in Egypt we therefore choose an integrated model using the Dunbar and Klahr (1988) and the Oser and Baeriswyl (2001) approaches as an attempt to offer teachers and students as much guidance as possible for developing their own ideas on scientific inquiry.

In order to find out which factors impact upon students' outcomes, a model that indicates the quality of the instruction and the effectiveness of teachers' actions is needed (Fischer et al., 2005). One such model describing processes in the classroom is Helmke's (2003) mediation model for quality of instruction. Helmke's model includes teachers' characteristics, particularly with regard to professional knowledge and teaching expertise, the quality of the teaching, and aspects of the lessons including how students are motivated, the way learning activities are organised and the results of student learning. The model of Helmke is developed from a more general pedagogical view but can easily be adapted to describe science lessons. With this model of instruction in mind, science teachers should present learning opportunities that meet students' abilities and motivations related to physics in individual ways. According to Helmke, students' performance in the classroom depends on the criteria for teaching quality, such as clarity, efficacy of classroom management and quality of learning materials, which are mediated by students' motivation and perception of the learning opportunities. All criteria mentioned are relevant for organizing and analysing physics teaching and learning, therefore Helmke's model will be used to develop the teaching unit in an intervention study which is completed but the data are not yet finally analysed. We will report the results in a subsequent article.

RESEARCH AIMS

Bearing in mind that the Egyptian Ministry of Education has high expectations that science teaching be reformed, the effective implementation of new teaching approaches, which focus on scientific inquiry and problem solving processes, is essential. Further, motivating students towards science in the earlier stages of their studies is needed. Hence, the three main aims of the study, based on the theoretical background, were to: 1. develop a model of problem solving for Egyptian science classes, 2. design instruments, and 3. validate the applied model of problem solving. The quality of the instruments evaluated in the study is described in the results section.

A MODEL OF PROBLEM SOLVING

The future vision of education in Egypt focuses on the need for improving self-learning, scientific inquiry and problem solving. Consequently, a model for problem solving that indicates how we can teach problem solving in the classroom is needed (Oser & Baeriswyl, 2001). In order to develop the model of problem solving, as described above it was important to define problem solving and experimental strategy knowledge, in the context of physics instruction (Dunbar & Klahr, 1988; Oser & Baeriswyl, 2001). The following section gives an overview about how the model was developed.

Defining Problem Solving

Posamentier and Krulik (2009) defined a problem as a situation that confronts the learner, that requires resolution, and for which the path to the answer is not immediately known. However, Robertson (2001) assumes that a problem arises when there is a goal but how to reach this goal is not known and figuring out how to do something different is needed (VanGundy, 2005). Eventually problem solving may provide a solution which can then be analysed retrospectively, as stated by Wragg (2005), but solutions often elude the rationale of an a priori approach. Problem solving skills are then said to involve identifying complex problems and reviewing related information to develop and evaluate options and implement solutions. Based on this definition, this study assumed that in a problem-solving situation, the experimentee ideally and typically knows the solution but he or she does not know how to reach it (Oser & Baeriswyl, 2001).

Defining Experimental Strategy Knowledge

Although the importance of teaching students to design, execute, and evaluate controlled experiments is emphasized in the USA national standards, the methods of teaching these concepts and procedures are not well specified (Toth, Klahr & Chen, 2000). To define students' problem solving skills, this study needs to describe students' experimental strategy knowledge and their problem solving abilities because both are used when students conduct experiments. As we are restricted to paper and pencil instruments, variables regarding practical performance cannot be methodically collected. However, Weinstein and Mayer (1986) have defined experimental strategy knowledge as behaviour and thoughts in which a learner engages and which is intended to influence the on-going learning process. Thus, every learner selects, acquires, organizes, or integrates new knowledge to reach a certain conceptual goal of a problem solving process. Thus, if the process of experimentation is meant to be a learning goal, experimental strategy knowledge has to be expressed and identified explicitly (Emden & Sumfleth, 2012). In addition we have to define students' problem solving abilities. Toth, et al. (2000) stated that, besides variables that refer to operability and organizing aspects when conducting real experiments, scientific experimentation mainly is the ability to use the *control of variables strategy* (CVS). Procedurally, CVS refers to the method used when experimenting with many variables. To avoid confounded experiments in which the crucial variable cannot be detected, all variables except those under consideration must be controlled in order to isolate the influencing factor on the outcome. Conceptually, CVS is a necessary strategy in designing easily understood experiments, like those performed at school, and in determining whether or not all variables involved are controlled.

Reiner (1992) classified three main types of knowledge about experimentation strategies, which are specific to generating evidence, interpreting evidence, and managing the memory requirements of the task. Recognizing that both structures of knowledge and strategies of experimentation are fundamental to scientific reasoning (cf., Schauble, Glaser, Ranghavan & Reiner, 1991), several researchers have explored the relations between knowledge and strategy in scientific discovery, using experimental methodologies. Some of this work (e.g., Chen & Klahr, 1999; Klahr & Dunbar, 1988; Kuhn, 2005; Schauble, Glaser, Ranghavan & Reiner, 1991, 1992) has contributed to more detailed and elaborated descriptions of the strategic component, particularly with new findings concerning how strategic processes interact with the students' hypotheses. In addition, those studies also showed how instruction for experimental strategy knowledge led to significant improvement in students' abilities to design simple and easy-to-understand experiments (e.g., Chen & Klahr, 1999).

Based on the above background, in particular Dunbar and Klahr's (1988) model of scientific discovery as dual search (SDDS), this study further defined knowledge about experimenting strategies as knowledge about strategies to systematically identify new information and strategies to systematically integrate new information into one's own knowledge base (see also Wirth and Leutner (2006), and Thillmann, Künsting, Wirth and Leutner (2009).

The Dunbar and Klahr model (SDDS) is a well known psychological model that embraces essential steps for solving a problem in comparison to other models (Emden & Sumfleth, 2012). In addition, SDDS has frequently been taken up by science education researchers and translated into sequences suitable for science classes (cf. Emden & Sumfleth, 2012; Schreiber, Theyßen, & Schecker, 2009), but not explicit enough to plan lessons. On the contrary for our study, the Oser and Baeriswyl model provides concrete teaching steps for problem solving processes at school (cf. Ohle, 2010). However, up until now no attempt has been made to use the aforementioned psychological and practical teaching models of problem solving in Egyptian teaching environments. Consequently, and to provide guidance also for the teachers, Dunbar and Klahr's (1988) model of scientific discovery as dual search (SDDS) was combined with Oser and Baeriswyl's (2001) practical teaching theory into a synthesized model for problem solving. This model of problem solving (see Figure 1) informed several aspects of their models by focusing on an individual's capacity to use cognitive processes in cross-disciplinary situations where the solution path is not immediately obvious and where the content areas or curricular areas that might be applicable are within a single subject area of science. Examples for such cognitive processes are included in Figure 1.

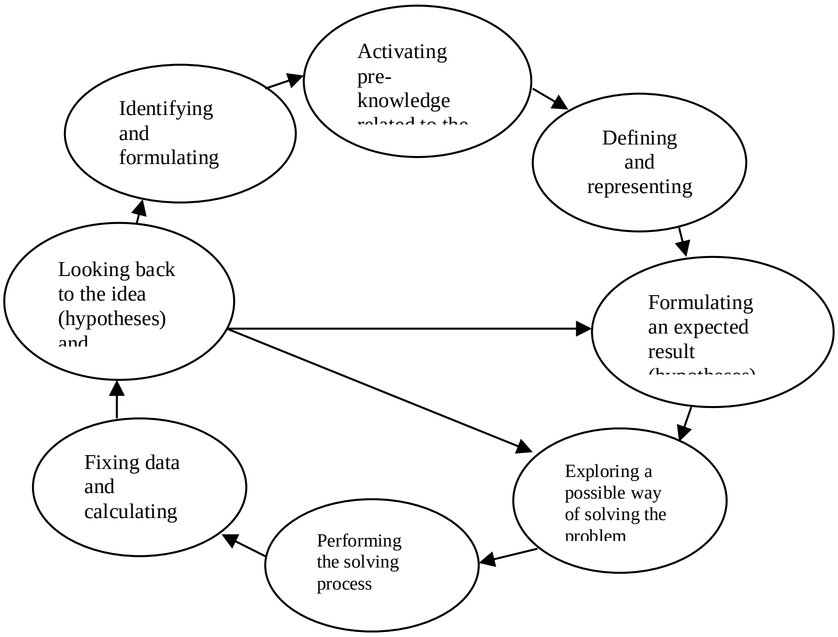

Figure 1. Model of problem solving to be used within the study

DESIGN OF INSTRUMENTS

Further, motivating students towards science in the earlier stages of their studies is needed. The four instruments used in this study are designed to measure 7th grade Egyptian students' experimental strategy knowledge, problem solving, content knowledge and motivation towards science (see Table1). The 7th grade level was selected because it is the first year of the Egyptian middle school stage and problem solving and scientific inquiery is one important part of the curriculum of this grade. In order to control the content, to enlarge the variance (students tend to have difficulties learning about this topic), and to enable the use of already developed instruments, the study focused on a limited area of one topic, namely, density and buoyancy. Content validity was taken into account by comparing the instruments and the curriculum. The content of the instruments is a proper subset of the curriculum. The items were partly adapted from the German and English versions of TIMSS and mainly PISA tests and questionnaires. In order to use the model of problem solving described in Figure 1 and the instruments in an intervention study, the validity of the instruments have to be judged by experts and the reliability has to be checked on an Egyptian sample. As the intervention should be performed in 7th grade of a middle school in Aswan, the sample for the study described here was chosen from the same school and the same grade. The sample size was limited by conditions set by the administration of the Ministry of Education.

TABLE 1

Instruments used in this study

Instrument	Original language
Experimental Strategy Knowledge Test (ESKT*)	English
Problem Solving Test (PST*)	English
Physics Achievement Test (PAT*)	English
Motivation Questionnaire (MQ*)	German

Described and depicted in this section are the development of the paper-pencil instruments and how the quality of instruments was tested.

(*) *ESKT Experimental Strategy Knowledge Test, PST Problem Solving Test, PAT Physics Achievement Test, MQ the Questionnaire of Motivation Towards Science*

Experimental Strategy Knowledge and Problem Solving Tests

Experimental strategy knowledge and problem solving were assessed within one combined test administered together. Whether these two constructs can be empirically differentiated will be presented in the results. The aim of this combined test was to assess students' knowledge and abilities concerning experimenting and problem solving. The test was developed according to the procedures of metacognitive strategies (Marschner, 2011; Thillmann, Künsting, Wirth and Leutner, 2009; Wirth & Leutner, 2006), and the physics content of the topic of density and buoyancy. In addition, the topic provides a broad basis for problem solving tasks with many already well-established experiments. To solve an item (see sample of tasks in Figures 2) and 3), students had to rate different types of action (items) according to a given situation (tasks). In the experimental strategy knowledge part, the tasks are related to identifying new information (e.g. finding out relationships between variables by using the control-of-variables strategy) and to integrating new information into one's own knowledge base (e.g. storing relationships between variables in one's long-term memory by elaborating upon them) (see Figure 2). The tasks for the problem solving part of the test include identifying and formulating a problem, exploring various avenues of solving a problem, looking back at the idea (hypotheses) and evaluating the problem solving process (see Figure 3). All tasks related to experimental strategy knowledge and problem solving are based on the topic of density and buoyancy. After that, students have to rate a list of possible action alternatives on a scale from strongly agree to strongly disagree (see Figures 2 & 3). The five items for experimental strategy knowledge and the four items for problem solving in this multiple-choice test were constructed based on the model of problem solving outlined in Figure 1. The following items are examples from the student test of experimental strategy knowledge:

(1) *"You want to find out on which variables the property of buoyancy (whether a body sinks, floats or swims in water) is dependent."*

You consider the following procedures in order to answer the question. Review the procedures with the scale from "strongly agree" to "strongly disagree": (Please tick only one box in each row)					
	SA	A	U	D	SD
a) I make a drawing with all possible variables.	□	□	□	□	□
b) I check the individual variables one after the other.	□	□	□	□	□
c) I try to explain the effect of each variable.	□	□	□	□	□

Figure 2. Three sample items from the experimental strategy knowledge test

(2) *"Remember that the force of buoyancy is the upward force caused by fluid pressure that keeps things afloat. Something that is positively buoyant floats in water. Something that is negatively buoyant sinks in water. Things that are neutrally buoyant neither float nor sink in water. With regard to buoyancy, two important aspects of objects are mass and volume. In an experiment, it is shown that a heavy steel model of a ship floats, but a heavy solid block of steel sinks in freshwater. Which procedure can explain the observation?"*

You consider the following ideas in order to answer the question. Review the procedures with the scale from "strongly agree" to "strongly disagree":					
	SA	A	U	D	SD
a) I measure the mass of the steel ship model and the mass of the solid block and compare them.	□	□	□	□	□
b) I measure the volumes of the solid block and the ship model by using a scale.	□	□	□	□	□
c) I think about the force of buoyancy and apply it to compare the weights of the steel ship model and the solid block.	□	□	□	□	□

Figure 3. Three sample items of problem solving

Physics Test to Measure Students' Achievement

This instrument is a multiple-choice test with 30 items on the topic of density and buoyancy based on the six categories of Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001), which include remembering, understanding, applying, analyzing, evaluating, and creating as cognitive activities. The items' difficulties should increase from understanding to creating. Some items were self-developed and others were adapted based on the PISA and TIMSS tests. The distribution of the 30 items over Bloom's revised categories is illustrated in Table 2.

TABLE 2

Physics test items according to Bloom's Revised Taxonomy

Category	Remembering	Understanding	Applying	Analysing	Evaluating
Items (30)	11	6	4	5	4

Items in the "applying" category are similar to the example shown in Figure 4

What happens to liquid with a density of 2 g/cm^3 when added to water?	
Floats or swims	◻
Sinks	◻
Sinks at first, then floats slowly	◻
Neither floats nor sinks	◻

Figure 4. One item in the "applying" category of the physics test to measure students' achievement

Questionnaire Assessing Students' Motivation

The fourth instrument is a questionnaire assessing students' motivation towards physics instruction. This instrument consists of 60 items in German and was adapted from PISA 2006 to the conditions in Egypt (e.g., integrated science courses, gender, middle school, period time, age) without changing the meaning of the items. Students have to rate different statements (tasks) according to their occurrence within various issues related to science, careers, and teaching and learning science. The example items in Figure 5 give an idea of the structure of this instrument.

How much do you agree with the statements below? (Please tick only one box in each row)				
	SA	A	D	SD
a) It is fun for me to deal with scientific topics.	◻	◻	◻	◻
b) I like reading about science topics.	◻	◻	◻	◻
c) I am pleased investigating scientific problems	◻	◻	◻	◻

Figure 5. Three sample items of motivation questionnaire

Translation and Editing Processes

After constructing the paper and pencil instruments, this study faced the challenge of translating the items into Arabic. Based on the unique characteristics of the educational systems of each country, language and cultural differences must be carefully taken into consideration to ensure the comparability of the data (Arora, Ramírez & Howie, 2006). In order to ensure a high quality of translation (see Brislin, 1970;1980), two steps were taken:

- The three English-language instruments (experimental strategy knowledge test, problem solving test and physics achievement test) were translated from English into Arabic by an Arabic native speaker fluent in English – a professional translator. One instrument (motivation questionnaire) was translated from German into Arabic again by an Arabic native speaker fluent in German—a professional translator.
- The validity of the Arabic translations was evaluated by translating the Arabic version back into English and German with other professional translators' support. The quality and accuracy of all translations into the original language were verified with the help of a native Arabic speaker who specializes in English at university, a native English speaker and native German speakers.

After we finished the translation process, we found that sometimes various Arabic words could be used for one English word. To solve this problem, three steps of comparison, with native speakers in Arabic, English and German, were done, with an acceptable inter-rater agreement of Cohen's Kappa between 0.41-0.62 (Wirtz & Caspar, 2002), and these steps as follow:

- The original English (E1) and German instruments (experimental strategy knowledge test, problem solving test, physics achievement test and motivation questionnaire) were compared with the translated English (E2) and German instruments by researchers and with the help of a native English speaker and two native German speakers.

- The Arabic translated instruments (experimental strategy knowledge test, problem solving test, physics achievement test and motivation questionnaire) were compared with the original English instruments by researchers and with the help of a native Arabic speaker.
- ΔE_{12} (comparison between the original English and translated English words) and ΔEA (comparison between the English and Arabic words) were measured. ΔE_{12} was positive (+) when the two English words had the same meaning and negative when they had different meanings. In this case, we needed to look at ΔEA and if we also found it to be negative, the Arabic word would need to be modified. If ΔEA was positive the original English and Arabic words had the same meaning and the Arabic word was accepted.

VALIDATION OF THE MODEL AND INSTRUMENTS

One important aspect of a test or scale is its validity (Peers, 2006). Criterion validity as a special aspect of construct validity reflects the evidence that a test reproduces the statistical relationship with the trait being measured by expert rating, tournament standing and/or predetermined criteria refered from peers. “A more critical test of validity is called criterion validity, which concerns whether the measure (a) can accurately forecast some future behavior or (b) is meaningfully related to some other measure of behavior” (Goodwin, 2010, p.132). In this study, the reason for gathering criterion validity was that the test or measure is to serve as a “stand-in” for each of the tests of problem solving, strategy knowledge, physics knowledge and motivation (Kaplan & Saccuzzo, 2009).

To ensure the criterion validity of the three developed/adapted tests and questionnaire, an expert rating of all instruments was conducted, including 14 Ph.D. students, three post-doctoral students and five teachers, who are specialized in science education in Germany or Egypt. Each participant received a sample of instruments and reviewed the instruments’ items according to the six categories of Bloom's Revised Taxonomy, PISA 2006 scales, problem solving, and experimental strategy knowledge definitions, and commented on a checklist related the correctness of the items. To estimate the criterion validity of the model for *problem solving,* three post-doctoral students as experts in science education in Germany reviewed the proposed model according to the steps of Dunbar and Klahr (1988), and Oser and Baeriswyl (2001) and commented on a checklist related to the quality of the model. The result of the inter-rater agreement of the experts rating was acceptable with Cohen's kappa between .58-.87 (Elliott & Woodward, 2007). For the problem solving and experimental strategy knowledge test items (Figures 2 & 3) the agreement is between $.63<\kappa<.85$ and $.72<\kappa<.86$ respectively, whereas, for the physics achievement test items (Table 2) the agreement is between $.58<\kappa<.81$. For the motivation towards science instruction questionnaire items (Figure 5), the agreement is between $.83<\kappa<.86$. Finally, for the model of problem solving (Figure 1), the agreement is between $.79<\kappa<.87$. Some minor modifications were made to instruments based on the comments provided. Taking into account that the expert rating required high level of interpretation, (high-inferent) we can register that, in a first attempt, criterion validity can be seen as given in our study.

Methodology For Evaluating The Quality Of The Instruments

Sample and Data Collection As already described, the assessment methodology used for these instruments was criterion-referenced (Peers, 2006). The resulting scores will be interpreted based on students' performance on the dependent variables of experimental strategy knowledge, problem solving abilities, achievement, and motivation towards science. Due to the Egyptian circumstances, we had a permission to implement this study with one class in one school type. Thus, this study was conducted with an ad-hoc and typical sample of 7^{th} grade students (mean age of 12.2 years (SD .41)) that included 44 students (girls) from one class in a general middle school in the city of Aswan, Egypt during the school year 2010-

2011. The sample size was different for every instrument because of varying numbers of absent students at the time of administration.

Data Analysis To investigate whether or not the predicted trait was defined by the test items, Rasch analysis (Boone & Scantlebury, 2006) was conducted on the data set because it offers better results with sufficient stability and even smaller sample size than do classical analyses. "Rasch analysis can handsomely handle a sample size of 25-30 to generate a sound 95% CL statistics and 50-60 for a 99% CL". (Linacre, 1994, p. 328). Nevertheless also classical values have been calculated to give a reference for interpretation (Wendler & Walker, 2006). For classical test analysis (Meyer, 2010) SPSS® (PASW statics; version 18) was used for processing. The item difficulty in this study was measured to be the percentage of students who answer an item correctly. Accordingly, the correct solution frequency was estimated. Accepted average of the correct solution rate of an item ranged from about 20-80% (see Bühner, 2004). To ensure the quality of items, the reliability was estimated; in this study, Cronbach's Alpha should be >.7 (Carver & Nash, 2012; Cortina, 1993). The required quality criteria for test and questionnaire items are discussed at the end of following section. The acceptable range of the INFIT mean square statistic for each item was taken to be (MNSQ) from .5 to 1.5 for (T) from –2.0 to +2.0, (T) is only useful for salvaging non-significant MNSQ > 1.5 when the sample size is smaller or the test length is short (Linacre, 2010). In order to check for the distribution of item difficulty and person ability (Linacre, 2010), the Wright map was investigated.

In order to analyse the test items of experimental strategy knowledge and problem solving, an expert rating was conducted with a group of five experts, all Ph.D. students in science education, according to the definitions described in the theoretical background. With five expert ratings for each task, the median for each task was estimated and later on compared pair-wise with students' answers. The medians of the experts' rating were then calculated (Field, 2009) as 0.38 and 1.34 for the experimental strategy knowledge test and the problem solving test, respectively. The example items in Figure 6 give an idea of the experts rating.

"Your class has conducted different experiments in small groups on the same topic. You shall now explain to the rest of the class what you have done in your group."

You consider the following procedures in order to answer the question. Review the procedures with the scale from "strongly agree" to "strongly disagree":
(Please tick only one box in each row)

	SA	A	U	D	SD
a) I describe in detail the results that are most important.	X	□	□	□	□
b) I describe who did perfectly in our group.	□	□	X	□	□

Figure 6. Sample of expert rating

Measurement of students' scores was "0" (no agreement between student and experts) or "2" (full agreement between student and experts). This degree of agreement is based on pair-wise comparisons (Aczbl & Saaty, 1983). For example, if the experts stated that in Task 1, alternative (a) is better than alternative (b), then the student gets two points for the same order. If the student says that alternative (a) is equally good as alternative (b), one-point is awarded, and if the student says alternative (b) is better than (a) zero points are awarded. This procedure was carried out for all tasks. At the end, all points were totalled and divided by the number of comparisons. Finally, there were 35 pair-wise comparison items for the experimental strategy knowledge test and 62 pair-wise comparison items for the problem-solving test. The quantitative analysis was used to identify students' achievement, experimental strategy knowledge, problem solving abilities and motivation towards science. Quality criteria were then estimated on the basis of the pair-wise comparisons. To underpin the results in this study, the instruments were additionally analysed using the Rasch model applying WINSTEPS® version 3.70.0.5.

RESULTS OF THE QUALITY OF THE INSTRUMENTS

In this section, we describe how the data from the four instruments' were subjected to classical test theory and how we checked the quality of the instruments based on this small sample size using the Rasch model as additional and even better evidence. First, the quality criteria described above are reported for each instrument and, in closing, the compilation of the final version of instruments is discussed.

Classical Statistical Analysis

Referring to Table 3, the internal consistency was estimated according to the classical test theory for the three tests related to experimental strategy knowledge, problem solving abilities, physics achievement, and the motivation questionnaire. Eight items showed unsatisfying discriminatory power and were therefore excluded in the physics achievement test (PAT). The internal consistency of the remaining items could be confirmed with a Chronbach's (α) = .81 (for 22 items; discriminatory power >.3). For the problem solving test (PST), the internal consistency of items could be confirmed with an Alpha value of .74, (for 62 items; discriminatory power >.3). For, experimental strategy knowledge test (ESKT), one item showed unsatisfactory discriminatory power and was therefore excluded. The internal consistency of the remaining items could be confirmed with an Alpha value of. 66 (for 34 items; discriminatory power >.3). For the motivation questionnaire (MQ), five items showed unsatisfactory discriminatory power and were therefore excluded. The internal consistency of the remaining items could be confirmed with an Alpha value of .96 (for 55 items; discriminatory power >.3). This low reliability result for ESKT was acceptable due to the study small sample size.

TABLE 3

Descriptive statistics and Cronbach's Alpha coefficient of instruments

Instrument	No. of respondents	No. of acceptable items*	No. of excluded items	M	SD	α
ESKT	44	34	1	26.02	7.10	.66

EPST	44	62	-	45.70	12.34	.74
PAT	40	22	8	9.03	3.84	.81
MQ	36	55	5	31.22	9.73	.96

* Discriminatory power > .3

As described in the previous section, whether the two constructs of experimental strategy knowledge and problem solving can be empirically differentiated needs to be analysed. Therefore, a bivariate correlation as shown in Table 4 has been estimated, revealing a moderate correlation (.43) only between the problem solving and physics achievement tests. This means that generally there is a relationship of a given direction and strength between these tests in the sample (Urdan, 2001). This significant relation is discussed in the next section.

TABLE 4

Correlations between the four instruments used in this study

Instrument		ESKT	EPST	PAT	MQ
ESKT	Pearson Correlation		.058	-.107	.071
	N		44	39	34
EPST	Pearson Correlation			.429**	-.038
	N			39	34
PAT	Pearson Correlation				-.007
	N				31
MQ	Pearson Correlation				
	N				

** $p < 0.01$ level (2-tailed).

Rasch Analysis

The data were also analysed according to the Rasch model. Overall, the analyses suggested 1, 8 and 5 misfitting items for the experimental strategy knowledge test (ESKT), the physics achievement test (PAT) and the motivation questionnaire (MQ) respectively. Hence, these items were excluded. Accordingly, the analyses supported using the 34, 62, 22 and 55 items of ESKT, EPST, PAT and MQ to define a single overall strategy knowledge, problem solving, achievement knowledge and motivation variables. Table 5 gives overview about the quality criteria of the instruments' items.

TABLE 5

Summary of persons, items separation, reliabilities and infit value of instruments

Instrument	No. of items	Item			Person			Infit	
		Reliability	Separation	TRUE SD	Reliability	Separation	TRUE SD	MNSQ	T
ESKT	34/35	.88	2.69	.77	.66	1.26	.37	.50-1.4	-3.0-2.5
EPST	62	.75	1.72	.37	.72	1.57	.29	.70-1.4	-2.7-2.9
PAT	22/30	.90	2.97	1.03	.79	1.92	1.26	.50-1.5	-2.0-2.0
MQ	55/60	.82	1.45	.35	.68	2.14	.39	.50-1.5	-2.1-2.47

CONCLUSION

This study has demonstrated the criterion validity of the model for problem solving used in this study, as estimated by experts. Furthermore, the study showed how instruments for experimental strategy knowledge, problem solving, physics achievement, and motivation towards science can be developed and adapted reliably for use by students in Egyptian middle schools.

In summary, the instruments were developed and adapted according to the proposed model of problem solving, newly performed studies on strategy knowledge and problem solving in Germany, PISA and TIMSS tests, and the school conditions in Egypt. This study revealed a moderate correlation between problem solving abilities and physics achievement for those students. This might be expected because the underlying cognitive activities in Bloom's Taxonomy, like analysing and evaluating, are also used for solving problems. The result showed no significant correlations among the other instruments which supported the discriminant validity of the instruments.

Overall, the results showed that even with this small sample size, instruments with acceptable reliability and validity could be effectively designed. As a result, a valid, independent and multidimensional model for problem solving was constructed and successfully applied in an Egyptian classroom. Moreover, this study developed a valid experimental strategy knowledge tests with 34 out of 35 pair comparisons, problem solving with 62 pair comparisons, physics achievement test instrument with 22 out of 30 items and a motivation questionnaire with 55 out of 60 items.

It has to be kept in mind that this study focused on a very limited area of one topic of density and buoyancy for 7th grade. Moreover, the sample was selected only from one location – the city of Aswan, Egypt. Results need to be discussed carefully due to the relatively small sample size in this study. It would be beneficial in the future if we could replicate and validate the study on a larger and representative sample in order to generalize the results to the Egyptian educational system.

FUTURE PERSPECTIVES

In the planned future intervention study, the aim is not to differentiate between unskilled or low achieving or unmotivated students but to measure the increase of experimental strategy knowledge, problem solving abilities, achievement, and motivation in a pre-post design, assuming that the students are more skilled, have higher achievement and are more motivated after the teaching program and therefore that the items will fit better in the future intervention study.

As a next step, the developed model will be used to construct a lesson plan for teaching the physics unit of density and buoyancy over a period of six weeks in a middle school in Egypt. Scaffolding for students will proceed over the aforementioned eight stages (see Figure1) by identifying and formulating a problem, activating pre-knowledge related to the problem, defining and representing the problem, formulating an expected result (hypotheses), exploring a possible way of solving the problem (variable discrimination), performing the solving process, organising data and calculating, and finally looking back to the idea (hypotheses) and evaluating. The unit of study developed by this model will be used in an intervention study and the developed instruments described in this study will be used to assess the students before and after learning the unit on density and buoyancy.

In this future study, a four-group, pre/post-test, quasi-experimental design (Shadish, Memphis, Dood & Evanston, 2002) will be used to determine whether or not students who are taught how to use the model of problem solving for experiments in the topic of density and buoyancy show improvement in experimental strategies knowledge, problem-solving processes, achievement and motivation towards physics. The intervention group will be compared with students who are taught according to the status quo to determine possible differences in learning outcomes or abilities. The groups will be similar, based on measures of age, gender, academic pre-knowledge, background and cognitive abilities. We will reveal how a physics program based on the validated model of problem solving described in this article can be constructed and evaluated. Moreover, the Egyptian school system and pre-service science teacher education programs in general will also be presented alongside the development of science teaching. This future study may also help and guide teachers and teacher educators on how to implement experimental learning situations in classrooms, which has high expectations in Egypt. The same, therefore, can be said of studies which provide guidance for Egyptian science teachers.

REFERENCES

Aczbl, J. & Saaty, T.L. (1983). Procedures for synthesizing ratio judgments. *Journal of Mathematical Psychology, 27*, 93-102.

Anderson, L.W. & Krathwohl, D.R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy*. New York: Longman.

Arora, A., Ramirez, M.J. & Howie, S.J. (2006). Using indicators of educational contexts in TIMSS. In S.J. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS* (pp.31-49). London: Routledge.

Berg, C.A.R., Bergendahl, C.B., Lundberg, B.K.S., & Tibell, L.A.E. (2003). Benefiting from an open-ended experiment? A comparison of attitudes to, and outcomes of, an expository versus an open-inquiry version of the same experiment. *International Journal of Science Education*, *25* (3), 351-372.

Bernholt, S., Neumann, K., & Nentwig, P. (Eds.). (2012). *Making it tangible: Learning outcomes in science education*. Muenster/New York/Muenchen/Berlin: Waxman.

Bloom, J.W. (Ed.). (2006). *Creating a classroom community of young scientists* (2[ed] ed). New York: Routledge.

Brislin, R. W. (1970). Back translation for cross-culture research. *Journal of Cross-Culture Psychology, 1*, 185-216.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-culture psychology: Methodology* (pp.389-444). Boston MA: Allyn & Bacon, Inc.

Bühner, M. (2004). *Einführung in die Test- und Fragebogenkonstruktion.* [Introduction to test- and questionnaire development]. München: Pearson.

Carver, R.H. & Nash J.G. (2012). *Doing data analysis with SPSS®Version 18*. Boston MA: Brooks/Cole Cengage Learning.

Chen, Z. & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098-1120.

Cortina, J.M. (1993). What is coefficient Alpha? Examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

Darling-Hammond, L. (2000), Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8* (1), Retrieved from (5.3.2012): http://epaa.asu.edu/ojs/article/view/392

Dogru, M. (2008). The application of problem solving method on science teacher trainees on the solution of the environmental problems. *Journal of Environmental & Science Education, 3* (1), 9-18.

Duschl, R. & Hamilton, R. (2011). Learning science. In R.E. Mayer & P.A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 78-107). New York: Routledge.

Emden, M. & Sumfleth, E. (2012). Assessing experimental procedures through different formats – A comparative study. In C. Bruguière, A. Tiberghien & P. Clément (Eds.), *E-Book Proceedings of the ESERA 2011 Conference, Lyon France. Science learning and Citizenship.* Part 10: co-ed. R. Millar (pp. 23-29). Lyon, France: European Science Education Research Association.

Engelhard, G. (2011). Historical views of invariance: Evidence from the measurement theories of Thorndike, Thurstone, & Rasch. In N.J. Salkind (Ed.), *Sage directions in educational psychology* (Vol. V, pp. 76-97), Los Anglos: SAGE Publications, Inc.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications.

Fischer, H.E., Klemm, K., Leutner, D., Sumfleth, E, Tiemann, R. & Wirth, J. (2005). Framework for empirical research on science teaching and learning. *Journal of Science Teacher Education, 16* (4), 309-349.

Good, T.L. & Brophy, J.E. (Eds.). (2008). *Learning in classroom* (10th ed.). Boston MA: Person Education, Inc.

Goodwin, C. J. (2010). *Research in psychology: Methods and design* (6th ed.). San Francisco CA: John Wiley & Sons, Inc.

Hashimoto, K., Pillay, H.K. & Hudson, P.B. (2008). Evaluating teacher education reform projects in developing countries: A case study of teacher educational reform in Egypt. *The International Journal of Learning, 15*(1), 123-131.

Hassan, F. (1997). Science education in Egypt and other Arab countries in Africa and west Asia. *The Interdisciplinary of Study Abroad, 3*, Retrieved from (22.10.2011): http://www.frontiersjournal.com/issues/vol3/vol3-11_Hassan.htm

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern* [Measuring, rating and improving the quality of instruction]. Seelze: Kallmeyer.

Hessen, D.J. (2010). Likelihood ratio tests for special Rasch models. *Journal of Educational and Behavioral Statistics, 35*(6), 611-628.

Hofstein, A. & Kind, PM. (2012). Learning in and from laboratories. In B.J. Fraser, K.G. Tobin & C. McRobbie (Eds.), *Second international handbook of science education* (Vol. 24, pp.189-207). New York: Springer.

Jeong, H. & Songer, N. (2008). Understanding scientific evidence and the data collection process: Explorations of why, who, when, what, and how. In C.L. Petroselli (Ed.), *Science education issues and developments* (pp. 169-200). New York: Nova Science Publishers.

Kaplan, R.M. & Saccuzzo, D.P. (2009). *Psychological testing: Principles, applications, and issues* (7^{th} ed.). Belmont, CA: Wadsworth, Cengage Learning.

Kim, M.C. & Hannafin, M.J. (2011). Scaffolding 6^{th} graders' problem solving in technology-enhanced science classrooms: A qualitative case study. *Instructional Science, 39*, 255-282.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48.

Klos, S., Henke, C., Kieren, C., Walpuski, M., & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen- zwei verschiedene Kompetenzen [Natural science experimentation and content knowledge - two different components]. *Zeitschrift für Pädagogik, 54* (3), 304-321.

Kneeland, S. (1999). *Effective problem-solving: How to understand the process and practise it successfully*. Oxford: How to Books, Ltd.

Lederman, N.G. & Lederman, J.S., (2012). Nature of scientific knowledge and scientific inquiry: Building instructional and capacity through professional development. In B.J. Fraser, K.G. Tobin & C. McRobbie (Eds.), *Second international handbook of science education* (Vol. 24, pp. 335-359). New York: Springer.

Linacre, J. M. (1994). Sample size and item calibrations stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J.M. (2010). *A User's guide to winstepsministep: Rasch-model computer programs*. Chicago: Winsteps.com.

Marschner, J. (2011). *Adaptives Feedback zur Unterstützung des selbstregulierten Lernens durch Experimentieren*. [Adaptive feedback to support self-regulated learning through experimentation]. Doctoral dissertation, Universität Duisburg-Essen.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J. & Chrostowski, S.J. (2004). *TIMSS 2003 international science report. Findings from IEA's trends in international mathematics and science study at the fourth and eighth Grades.* TIMSS & PIRLS International Study Center, Lynch School of Education. Boston College: Boston, MA.

Meyer, J.P. (2010). *Reliability*. Oxford: Oxford University Press, Inc.

Ministry of Education-Egypt (MOE) (2008). *The development of education in Egypt (2004-2008)*. A national Report, Cairo, Retrieved from (22.10.2011): http://www.ibe.unesco.org/National_Reports/ICE_2008/egypt_NR08.pdf

Ministry of Education-Egypt (MOE) (2010). *The future vision of pre-university education,* (in Arabic). Retrieved from (26.08.2012): http://knowledge.moe.gov.eg/Arabic/about/politic/vision/

Mullis, I.V.S., Martin M.O., Robitaille D.F. & Foy P. (2009). *TIMSS 2007 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Newburgh1, R. (2008). Why do we lose physics students? In Petroselli C.L. (Ed.), *Science education issues and developments* (2nd ed., pp. 1-4). New York: Nova Science Publishers, Inc.

Organization for Economic Co-operation and Development (OECD) & IBRD/the World Bank (2010). *Reviews of national policies for education: Higher education in Egypt* (2nd). Paris, France: OECD Publishing.

Oser, F.K. & Baeriswyl, F.J. (2001). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Ed.), *AERA´s handbook of research on teaching* (4th ed., pp. 1031-1065). Washington, DC: American Educational Research Association.

Peers, I.S. (2006). *Statistical analysis for education and psychology researchers*. London: The Falmer Press.

Posamentier, A.S. & Krulik, S. (2009). *Problem solving in mathematics, Grades 3–6": Powerful strategies to deepen understanding*. Thousand Oaks CA: Corwin.

Quintana, C., Reiser, B. J., Davis, E.A., Krajcik, J., Fretz, E., Duncan, R.G., et al. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*, 337-386.

Robertson, S.I. (2001). *Problem solving*. Philadelphia PA: Psychology Press/Taylor & Francis, Inc.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis, 23*, 197-227.

Ronald, H.H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration, 47*(2), 227- 249.

Schauble, L., Glaser, R., Ranghavan, K. & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Science, 1*(2), 201-238.

Schauble, L., Glaser, R., Ranghavan, K. & Reiner, M. (1992). The integration of knowledge and experimentation strategies in understanding a physical system. *Applied Cognitive Psychology, 6*, 321-343.

Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen?! [Measuring experimental competence?!]. *Physik und Didaktik in Schule und Hochschule, 8*(3), 92-101

Shadish, W.R., Memphis, T., Cook, T.D. & Evanston, I. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Sutman, F.X. & Schmuckler, J.S. & Woodfield, J.D. (2008). *The science quest using inquiry/discovery to enhance student learning, grades 7–12*. San Francisco CA: John Wiley & Sons, Inc.

Tang, X, Coffey, J.E., Elby, A. & Levin, D.M. (2009). The scientific method and scientific inquiry: Tensions in teaching and learning. *Science Education, 94*, 29-47.

Thillmann, H., Künsting, J., Wirth, J. & Leutner, D. (2009). Is it merely a question of ‘what’ to prompt or also ‘when’ to prompt? The role of presentation time of prompts in self-regulated learning. *Zeitschrift für Pädagogische Psychologie, 23,* 105-115.

Toth, E. E., Klahr, D. & Chen, Z. (2000). Bridging research and practice: A cognitively based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction, 18*, 423-459.

UNESCO (2008). *Egypt's educational context and the government's education priorities and strategies*. UNESCO Cairo Office-Egypt.

Urdan, T.C. (2001). *Statistics in plain English*. Mahway NJ: Lawrence Erlbaum Associates, Inc.

VanGundy, A. B. (2005). *101 Activities for teaching creativity and problem solving*. San Francisco, CA: John Wiley & Sons, Inc.

Weinstein, C.E., & Mayer, R.E. (1986). The teaching of learning strategies. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 315-327). New York: Macmillan.

Wendler, C.L.W. & Walker, M.E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.445-467). Mahway NJ: Lawrence Erlbaum.

Wirth, J. & Leutner, D. (2006). Selbstregulation beim Lernen in interaktiven Lernumgebungen. [Self-regulated learning in interactive learning environments]. In H. Mandl & H.F. Friedrich (Eds.) *Handbuch Lernstrategien* (pp. 172-184). Göttingen: Hogrefe.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität* [Interrater agreement and interrater reliability].Göttingen: Hogrefe.

Wragg, E.C. (2005). *The art and science of teaching and learning: The selected works of ted Wragg*. New York: Routledge.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99-149.