

How to develop tiered tests: A developmental framework using statistical indexes and four tier types in secondary physics

Minkee Kim* · Jinsun Jung¹ · Sung-Jae Pak²

University of Helsinki, Finland · ¹Seoul National University · ²Science Culture Education Research Institute

Abstract: In the era of the outcome-based education, multiple-choice test has been widely employed owing to its efficiency that enables educators to evaluate a quantity of students with much objectiveness. However, the prevalent test has not been reconsidered enough to overcome its apparent shortcomings: examiners' effort for developing plausible and faultless distracters defending from every falsification, and students' random guessing on key choices. For alleviating such defects, *tiered test* as an experimental format of multiple-choice tests has been suggested in science education. Since there has not accumulated much study on the implementation of tiered tests, our research aim is set to construct a framework suggesting statistical indexes for rationally discerning tiered units that develop an effective tiered test. Graded both by our tiered-scoring and by the conventional partial-scoring, the preliminary tiered test in secondary physics attests the improvement in its discrimination and difficulty distribution. The findings reveal that the two indexes discern effective tiered items: discrimination increase (Ct-p) and difficulty decrease (Dp-t). Based on the index information, 4 heterogeneous tier types are recommended in the content of secondary physics: *directional manipulation, repeated calculation, diverse explanation, and plural variables.*

Key words: tiered test, summative assessment, item discrimination, item difficulty, secondary physics

I. Introduction

Ever since the outcome-based education reform emerged in the 1980s (Hargreaves *et al.*, 2001; Sahlberg, 2004), evaluating students' school achievement, regulated by performance standards, has been regarded as a reliable method to provide governments and schoolteachers with remedial information of every learner (Andersson, 2000; Atkin *et al.*, 2005; Barr, 1993; Bell, 2007; Lawrenz, 2007). On the requests of school evaluation, *multiple-choice* test has been employed for examining students' achievement, owing to its effective feature of dealing vast students with much objectiveness in grading. However, reliability issue of multiple-choice tests—whether to consistently evaluate students' competence—has arisen, because building distracters (incorrect choices) requires elaborate efforts, and examiners occasionally fail to deliberate all possible falsifications (Bae, 2007). For example, the National College Scholastic

Ability Test (CSAT, called “Suneung” in Korean) in 2006 was an issue for failing to reliably examine a physics concept of *ideal gas*. The examiners questioned 5 choices by combining 3 statements that faultily or correctly describe the given problem situation, instead of generating 5 statements. Such a prevalent and deficient strategy of multiple-choice tests, which inquires fewer statements than the number of choices by their combination, indicates that the conventional test needs to be reconsidered.

As an attempt to enhance conventional multiple-choice tests overcoming challenges of school practice, *tiered tests* or *sequential problems* have been suggested in physics education (Hudson & Hudson, 1981; Lee, 1998; Lee & Pak, 1996). The experimental test combines two or more items into a tier unit according to subject-specific rationales. In this study, the hierarchy of the terminologies is restructured: a tiered test consists of generally 20 tier units combining two or three items each; every item presents plural

*Corresponding author: Minkee Kim (physhero@gmail.com)

**Received on 2 January 2009, Accepted on 25 May 2009

choices including keys and distracters (i.e., tiered test > tier unit > item > choice)". This hierarchy of tiered tests originates from Linn and Gronlund's (1993) general description of test, item, and choice. In their definition, this study added *tier unit* as a combination of two or three of items.

Graded by the tiered-scoring and by the conventional partial-scoring, a preliminary tiered test in secondary physics is compared with conventional multiple-choice tests, which characterizes the advantageous features (alleviating students' guessing and enhancing test discrimination). The discussion is followed by a suggestion of the framework for implementing tiered tests.

II. Background

Conventional multiple-choice test

In the era of the student evaluation, multiple-choice tests have been prevalently employed in science education owing to their perceived efficiency. One of these known advantages is that multiple-choice tests can stimulate students' deep consideration in the numerous problems where key choices are not solely true or false but vary in degree of their appropriateness, e.g., best method, best reason, or best interpretation (Linn & Gronlund, 1993). Others adhere to the objective test in the perception that students' achievements, examined by open-ended tests and by multiple-choice tests, present little differences (Rebello & Zollman, 2004). Due to these potential advantages, many standardized tests targeting a large number of students such as Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Korean CSAT have employed multiple-choice tests to assess students' long-term achievement in school curriculum.

Others, however, have cautioned the multiple-choice tests with pieces of evidence contradicting the advantages, in evaluating student achievements. For instance, it has been argued that the prevalent test overlooks explicit influence

of students' guessing (Burton, 2001; Lee & Pak, 1995). The literature which examines influence of guessing in multiple-choice tests with science learners reveals that adopting an extra choice into a four-choice test, in order for constructing a five-choice test, decreases total scores by 5.46% among the overall sample students (Lee & Pak, 1995). A simple calculation of the random selections presents that students' guessing probability to a key choice vary between 20% in a five-choice item and 25% in a four-choice item. The arithmetical difference predicts 5% decrease when a choice is added into a four-choice item, which approximates to the finding by Lee and Pak. Likewise, Burton (2001) reassures that since the number of choices for item influences students' guessing, many multiple-choice tests are unreliable unless properly manipulated. These research findings statistically affirm the apparent shortcoming in multiple-choice tests.

Conventional strategies to alleviate the apparent influence of guessing on multiple-choice tests have been attempted by various methods. For instance, Dimes (1973) adds a "*do-not-know*" choice which allots zero point to encourage students to skip unsure items. In case students guess their answer and mark on distracters, total score of the test is deducted, which in turn is believed to hinder them in guessing key choices. Others attempt to produce more plausible distracters to decrease the probability of guessing (Dimes, 1973; Hudson & Hudson, 1981). However, such conventional strategies are examined later to arouse undesirable side effects. When deducting scores for wrong answers, passive students might not dare to represent their ideas that they vaguely know (Burton, 2001). In addition, producing more plausible distracters to confuse students is not effective in terms of teachers' practice in school science, e.g., examiners thus build five-choice items that combine fewer than 5 statements as the CSAT does. Likewise, the literature in science education suggests producing more plausible distracters by adopting choices from students' misconceptions reviewed in the precedent

cognitive studies or by directly examining students' erroneous open-ended answers on pilot tests (Rebello & Zollman, 2004; Tamir, 1998).

As discussed above, multiple-choice tests have been employed for the major assessments that evaluate students' long-term school achievement even with its shortcomings such as students' guessing and decrease in test discrimination. The remedial strategies have been devised against their empirical shortcomings, not succeeding to eradicate them.

Tiered test

Tiered test scores each item by weighting total number of items in a tier unit. That is, only when a student responses all the key choices in a tier unit, s/he obtains its unit points. In contrast, a student earns zero point in a tier unit, if any of their answers is wrong. This unique grading method is referred to as *tiered-scoring*, while the conventional grading method in multiple-choice tests is labeled as *partial-scoring* for better distinction in this study. Figure 1 presents how tiered tests and multiple-choice tests differ in scoring items. Student A and B undertake the conventional multiple-choice test. Although Student A does not complete any of the given tier units, s/he earns the even total point, as Student B who completes the tier unit 1 does. Because the conventional test partially scores each item, it raises the concern of fair discrimination. In the premise that students should fully understand a given problem situation (such as physics phenomena that ask plural, linked concepts) presented in a tier unit, the partial-scoring fails to clearly distinguish students who are competent with solving a given problem from those who simply memorize partial concepts in a sporadic way (Kwon & Lee, 1987). On the contrary, for Student C and D in the tiered test with a tiered-scoring, their problem-solving capability is properly discriminated by differences in the total point zero from two.

Another potential advantage in tiered tests has been known as alleviating students' guessing to

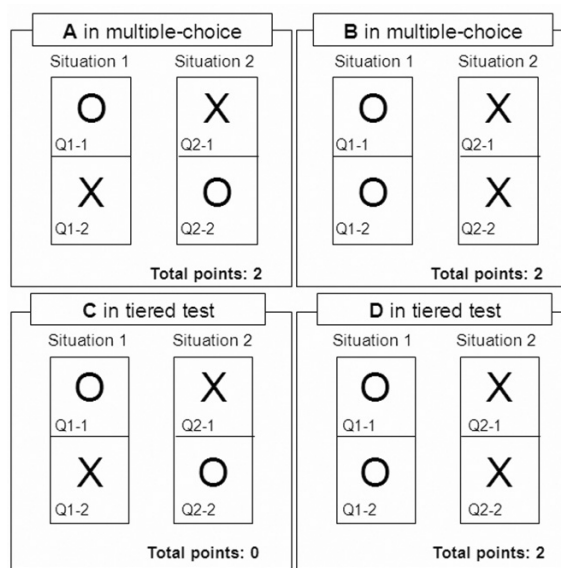


Fig. 1 Comparison between tiered tests (*tiered-scoring*) and conventional multiple-choice tests (*partial-scoring*)

key choices (Lee, 1998; Lee & Pak, 1996). The probability in which a student scores a tier unit through random guessing can be estimated by an uncomplicated calculation, e.g., the probability would be 4% in the case of a two-tier unit containing one key choice and 4 distracters the guessing probability (20%) of each item is squared. This is significantly lower than the probability in a case that any of two distinct items is guessed to key choices, e.g., the both key choices are guessed by 4%, the first by 16%, and the second by 16%; thus, the total probability of guessing reaches to 36%. Likewise, applying tiered-scoring to transform conventional multiple-choice tests into tiered tests is found to significantly alleviate students' random guessing (Lee, 1998). Lee's study examines the relationship between students' scores in tiered tests and their assurance on each item. His research findings indicate that high-assurance responses tend to be scored correct by the tiered-scoring, while alleviating low-assurance responses against obtaining any score. This is, students achieve scores within their competence in tiered tests, which empirically proves that tiered tests alleviate the probability of

students' guessing to key choices.

As attempting to coincide with students' everyday experience, various situations need to associate with evaluations of problem-solving capability in physics (Scott, 1985; Stark, 1999). That is based on the fact that physics instructions frequently analogize pairs of mutually-conflicting situation more than other science subject does (Bak & Kwon, 1990; Maloney, 1994; Mitchell & Baird, 1986). Others insist that students could achieve certain parts of testing purposes and they might deserve partial scores within the criteria of assessment purpose. In the same light, the general criteria of assessments is suggested to consist of knowledge, understanding, application, thinking skills, general skills, attitudes, interests, appreciations, and adjustment (Gronlund, 1971). However, two or three items that are derived from a situation would better compose a tier unit to properly measure whether students are capable of solving given situation as a unit, especially in physics. By combining situation-specific items into a tier featured by such proper tier types, tiered tests can discern fully-competent students who succeed to understand a given situation from incompetent students (see Student C in Figure 1). Tiered tests are thus expected to enhance test discrimination.

III. Research Questions

The two potential advantages of tiered tests discussed above, i.e., alleviation of students' guessing and enhancement of test discrimination, could be valid under the premises: items are consistently aiming to evaluate students' competence in a given problem situation by a reasonable and compulsory tier the way how to combine items into a tier unit should be theoretically clarified. In order to suggest types of effective tier units, the previous studies attempt to combine two items in a tier unit by setting a first item examining students' scientific concepts which is then followed by a second item questioning their reasons for the first choices

(Haslam & Treagust, 1987; Tan *et al.*, 2002). Even though tiered tests could improve the format of multiple-choice test, there have been few studies that suggest reliable types of tier units to implement the effective testing format. For example, an attempt of building reasonable tier types concerned a three-tiered test in physics by a series of three items questioning conclusions, processes, and relevant scientific concepts (Lee, 1998). However, as pointed out by Tamir (1998), the methodology of linking items into units was arbitrary and lacks of theoretical background. Therefore, as another step toward implementing tiered tests, this study resolves the two research questions as below:

- *How can a framework using statistic indexes be set for discerning effective tier units?* The previous studies on tiered tests report the improvement of test discrimination and the alleviation of students' guessing (Lee, 1998; Lee & Pak, 1996). Examining the effects of tiers item-by-item, this study compares our tiered-scoring with the conventional partial-scoring to identify indexes for discerning effective tier units in secondary physics. Accordingly, the first research product of this study will be an overall framework and its indexes for identifying effective tier units.
- *Which tier types in-between effective tier units can be identified in physics contents by the indexes?* Although how to combine items into a tier unit is the most critical for building a tiered test, little study has clarified how to combine items by proper rationales. As the literature suggests that sequenced items, the naive type of tier units, can alleviate time and effort for reliably evaluating physics learners (Hudson & Hudson, 1981), tier types in-between the effective tiered items need to be specified to promote tiered tests for practical and academic purposes.

IV. Methodology

The tiered test

In order to avoid the shortcomings of the tiered test reviewed from the literature, the 3 basic premises were identified to guide the development of tiered tests, as follows:

- Items in a tier unit should question a single concept (or an exclusive chunk) of relevant concepts.
- Plausible distracters could be reused to alleviate examiners' effort for interpreting the given items.
- A tier unit should be combined by meaningful tier types which stimulate students to consider its items as a whole.

On ground of the premises, a preliminary tiered test was developed, aiming to examine the yearly achievement for eleventh graders in secondary physics. It consists of 23 tier units (Q1 – Q23). Every tier unit contains two or three items,

totaling 53 five-choice items (see Table 1). Cronbach's Alpha of the test is measured by 0.89 through partial-scoring, and by 0.81 through tiered-scoring, which indicates that the item and the tier units reliably measure students' competence in secondary physics. For the format of the tiered test, it involves two-tier units and three-tier units following the precedent literature (Lee, 1998; Lee & Pak, 1996).

Sample and validity

190 eleventh graders in Seoul and Inchoen, Korea were associated with the tiered test. At the moment of the examination, the students had completed the national curriculum of *Physics I*, which provided them with enough instruction of the subject in the tiered test. Content validity—it involves how properly a test comprises the content domain (Haynes et al., 1995)—was examined by three secondary schoolteachers and two educational researchers. The judges examined whether every of the contents of tier units listed in Table 1 are valid to evaluate students physics

Table 1
Characteristics of the physics test for eleventh graders

Number of Tier Units	: 23	
Content of Tier Units	: Q1, Velocity I; Q2, Velocity II; Q3, Velocity and speed; Q4, Newton's first and second law I; Q5, Newton's first and second law II; Q6, Dynamics with gravity; Q7, Elastic force; Q8, Momentum; Q9, Conservation of mechanical energy I; Q10, Conservation of mechanical energy II; Q11, Dynamics with air resistance; Q12, Ohm's law I; Q13, Ohm's law II; Q14, Electric and Thermal energy; Q15, Electric transmission and energy loss; Q16, Connection of electric utensil and electric power; Q17, Magnetic field from current; Q18, Force in magnetic field; Q19, Electromagnetic induction; Q20, Traveling wave; Q21, Concave mirror and its image; Q22, Sound; Q23, Total reflection of light	
Number of Items	: 53	
Format	: Two-tier unit (grades: 0, 2), Three-tier unit (grades: 0, 3)	
Recommended Grade	: 11 th graders	
Time Allowance	: 60 minutes	
Reliability (Cronbach's Alpha)	: Partial-scoring 0.89 (Number of items = 53)	: Tiered-scoring 0.81 (Number of items = 23)
Difficulty (average)	: Partial-scoring 0.07– 0.96 (0.64)	: Tiered-scoring 0.02 – 0.94 (0.46)
Discrimination (average)	: Partial-scoring 0.03 – 0.63 (0.36)	: Tiered-scoring 0.03 – 0.75 (0.43)

competence regarding their school curriculum, and whether the items are understandable and faultless among the eleventh graders. The validity of our tiered test commenced with the amendments according to their comments.

As shown in Table 1, series of statistical procedure also validated whether the tiered test was properly developed for resolving our research questions. Its guessing correction was obtained by the conventional guessing correction formula (for n -choice): $\text{Guessing correction} = (\text{Number Wrong}) / (n - 1)$ (Dimes, 1973). Regarding that decrease in average difficulty implies removal of the student guessing, the guessing correction 0.14 in tier-scoring approximated to the decrease in average difficulty 0.18 which was obtained by shifting the grading method from the partial-scoring (average difficulty = 0.64) to the tiered-scoring (0.46). In addition, when compared with Lee and Pak's (1996) two-tier test (decrease in average difficulty = 0.17; guessing correction = 0.10) and with Lee's (1998) three-tier test (0.14; 0.25), the tiered test in this study was statistically validated to possess the characteristic of guessing correction.

The discrimination was determined by deducting average difficulty among lower third of students from the one among upper third of students, and then divided by a third of the total number of students. While the criterion of upper-lower group varies, e.g., quartile, 27%, and one-third, based on sample size, the literature had suggested adopting the one-third criterion with 210 examinees (Engelhart, 1965). This study of 190 sample students employed upper third ($n = 63$) and lower third ($n = 63$) groups for the discrimination. The tiered test revealed 0.07 increase in the average discrimination between 0.36 by partial-scoring and 0.43 by tiered-scoring.

V. Findings

Discerning indexes of effective tiered items

In order for identifying individual tier units that contribute to the efficiency of the overall tiered test, we scrutinized whether the shifting from

partial-scoring to tiered-scoring involves any advantages on every tier unit. The discrimination refers that each item evaluates students in the same tenor in which the overall test is supposed to do positive discrimination indexes are thus required for every item to suffice purpose of tiered tests (Gronlund, 1971). The tiered test is examined to enhance its discrimination by tiered-scoring. Accordingly, the framework for discerning effective tier units adopts the discrimination increase ($Ct-p$), calculated by shifting the grading method from partial-scoring to tiered-scoring. In order to analyze tier types, the framework in this study employs tier units that possess $Ct-p$ higher than 5%. This study suggests the index, as follows:

$$\text{Discrimination increase } (Ct-p) = (\text{Discrimination by tiered-scoring}) - (\text{Highest discrimination by partial scoring})$$

Another critical method of item analysis is to examine the difficulty. In general, an ideal test demands average difficulty of 0.50 and zero Skewness in which a histogram of item difficulty approximates to the normal distribution (Gronlund, 1971). By grading the multiple-choice items through tiered-scoring, the identical items in this study transform their distribution closer to the ideal average difficulty (0.46 by tiered-scoring 0.64 by partial-scoring). In addition, its Skewness also approaches to the ideal value of zero, i.e., the distribution develops into the normal distribution, when graded by tiered-scoring (0.003 by tiered-scoring; -0.507 by partial-scoring) (see Figure 2). On the ground of the two rationales that (a) each item should contribute to the average decrease of difficulty in the same tenor, and (b) the comparison by the partial-scoring and the tiered-scoring should help to identify advantages of the tiered test, the difficulty decrease ($Dp-t$) is nominated as the second index to discern effective tiered items. In order to analyze the tier types, the framework in this study employs tier units which possess $Dp-t$ higher than 5%. This study suggests

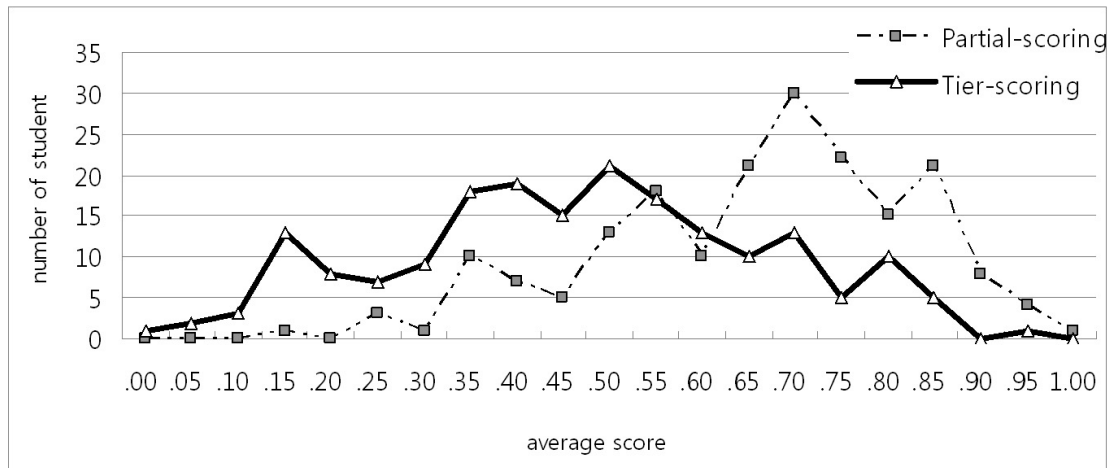


Fig. 2 Number of student for average score the histogram presents the near-to-zero Skewness, approaching to the ideal normal distribution, when graded by tier-scoring (0.003) more than by partial-scoring (-0.507).

the index, as follows:

$$\text{Difficulty decrease (Dp-t)} = (\text{Lowest difficulty by partial scoring}) - (\text{Difficulty by tiered-scoring})$$

According to the Ct-p and Dp-t indexes, the framework in this study identified 9 tier units as contributing most to the effective tiered test (see Table 2). For example, Q18 increased its discrimination index by tiered-scoring (Ct-p = 0,08). The Ct-p is obtained by deducting the highest discrimination by partial-scoring (0,48) from the one by the tiered-scoring (0,56). Likewise, Q18 also obtains its difficulty close to

the ideal average. Its Dp-t is calculated as 0,10, deducting the difficulty by tiered-scoring (0,49) from the lowest difficulty by partial scoring (0,59). On the ground of our premises, it is implied that 10% or more of the students who achieved scores in Q18 randomly guessed their choices and failed to continue their guessing. Consequently, this preliminary study on exploring tier types categorizes the effective tier units by the index information: Ct-p and Dp-t should be higher than 5% respectively; the tier unit should possess content validity. In the following section, the tier types from distinctive tier units are described in the content of secondary physics.

Table 2

Selected tier units with their discrimination increase (Ct-p) and difficulty decrease (Dp-t); the underlined maximum or minimum figures are chosen for calculation.

Item #	Discrimination			Difficulty			Tier Type
	PS*(a)	TS** (b)	Ct-p (b-a)	PS (c)	TS (d)	Dp-t (c-d)	
Q1-1	0.17	0.25	0.06	<u>0.83</u>	0.78	0.05	<i>directional manipulation</i>
Q1-2	<u>0.19</u>			0.85			
Q3-1	0.13	0.44	0.12	0.93	0.69	0.06	<i>directional manipulation</i>
Q3-2	0.25			0.84			
Q3-3	<u>0.32</u>			<u>0.75</u>			
Q7-1	0.49	0.68	0.09	0.74	0.47	0.11	<i>plural variables</i>
Q7-2	<u>0.59</u>			<u>0.58</u>			
Q7-3	0.51			0.63			

Q10-1	0.29	0.54	0.14	0.87	0.71	0.05	<i>plural variables</i>
Q10-2	<u>0.40</u>			<u>0.76</u>			
Q11-1	<u>0.38</u>	0.48	0.10	<u>0.28</u>	0.22	0.06	<i>diverse explanation</i>
Q11-2	0.32			0.69			
Q11-3	0.27			0.85			
Q12-1	0.25	0.40	0.13	0.87	0.52	0.05	<i>repeated calculation</i>
Q12-2	<u>0.27</u>			<u>0.57</u>			
Q16-1	0.57	0.75	0.16	0.72	0.46	0.11	<i>diverse explanation</i>
Q16-2	0.41			0.82			
Q16-3	<u>0.59</u>			<u>0.57</u>			
Q17-1	0.46	0.59	0.07	<u>0.52</u>	0.43	0.09	<i>diverse explanation</i>
Q17-2	<u>0.52</u>			0.76			
Q18-1	0.35	0.56	0.08	0.67	0.49	0.10	<i>directional manipulation</i>
Q18-2	<u>0.48</u>			0.61			
Q18-3	0.44			<u>0.59</u>			

*PS represents partial-scoring; **TS represents tiered-scoring.

Tier type 1: Directional manipulation

The characteristic of *directional manipulation* is frequently employed to explain gravitational field, electric field, or magnetic field in secondary physics. These physics fields are represented in equations where each variable is linked by curl, gradient, and divergence. Hence, the vector calculations require students to understand plural cases manipulated by different directions. As shown in Figure 3, the tier unit Q18 examines whether students can fully understand Faraday's law of electromagnetic induction with the reverse direction of electric and magnetic fields subsequently. Through the three-tiered unit repeatedly questioning the identical concepts, 10% of students failed to answer all scientific choices associated. Furthermore, the tier enhances the discrimination by 8%.

Tier type 2: Repeated calculation

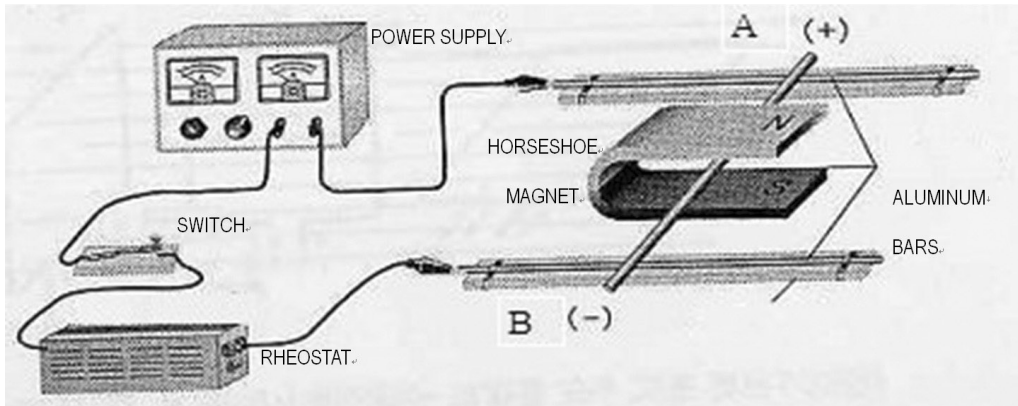
Mathematical calculations are required in many problem situations in order to understand a physical phenomenon. What is notable with the characteristic in physics education is that students tend to narrow their cognitive procedure into quantitative reasoning, failing to understand concepts that each calculation implies (McDermott, 1998; Park & Cho, 2005). The tier

type of *repeated calculation* in tiered tests encourages students to repeatedly employ a proceeding calculation for succeeding item in a tier unit, which alleviates their burden of mathematical tasks. The literature asserts that adopting a strategy to lessen students' computational burden in class instruction enhances their reconsideration on given contents (Sangster, 1992). For instance, Figure 4 presents the two-tier item Q12. This measures whether students understand the concept in electromagnetism: 'A resistance of metal wire varies in proportion to its resistivity and its length, while it varies in inverse proportion of its area of cross section.' Both items (Q12-1 and Q12-2) share the identical process of calculation and ask two interrelated concepts of resistance and resistivity. Hence, the two items are tiered according to the tier type of *repeated calculation*. The tiered-scoring discovered that 5% of students failed to answer scientific choices both on Q12-1 and Q12-2, and that the discrimination increased by 13%.

Tier type 3: Diverse explanation

The literature has revealed that students occasionally fail to understand new problem situations described in tests (Stark, 1999; Taber,

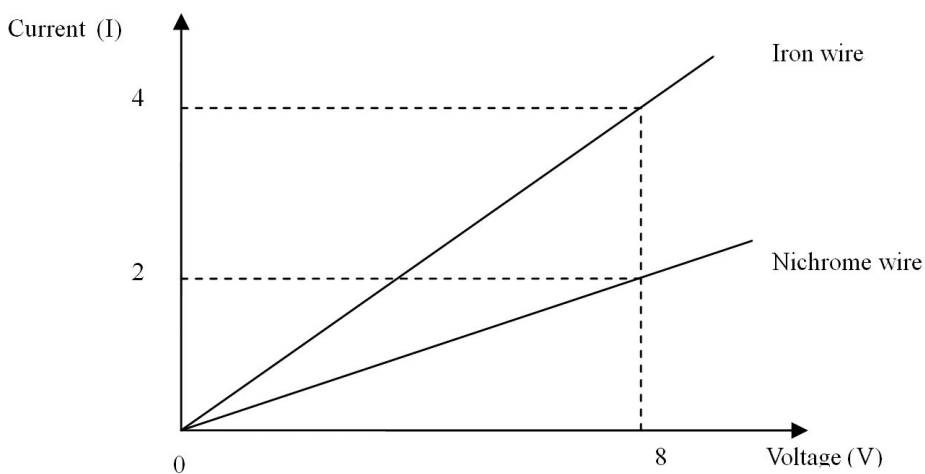
A horseshoe magnet, three aluminum bars, a power supply, a switch, and a rheostat are connected as follows. All resistance except the rheostat is neglected. *CHOICES: (a) The bar moves inward; (b) The bar moves outward; (c) Part A of the bar only moves inward; (d) Part A of the bar only moves outward; (e) Part B of the bar only moves inward*



- Q18-1. How the aluminum bar that is placed inside of a horseshoe magnet reacts, when you close the switch? (*key choice: b*)
- Q18-2. How the aluminum bar that is placed inside of a horseshoe magnet reacts, when you close the switch with the reverse current direction from Q18-1? (*key choice: a*)
- Q18-3. How the aluminum bar that is placed inside of a horseshoe magnet reacts, when you close the switch with the reverse direction of the magnetic field from Q18-1? (*key choice: a*)

Fig. 3 Three-tier unit Q18 with directional manipulation; $Ct-p=0,08$, $Dp-t=0,10$

Resistances of two same shaped iron and nichrome wires were examined with variation of voltage and plotted into the Voltage-Current diagram below. *CHOICES: (a) 1:1; (b) 1:2; (c) 2:1; (d) 1:4; (e) 4:1*



- Q12-1. What is the ratio of resistances between the nichrome wire and the iron wire? (*key choice: c*)
- Q12-2. What is the ratio of resistivity between the nichrome wire and the iron wire? (*key choice: c*)

Fig. 4 Two-tier unit Q12 with repeated calculation; $Ct-p=0,13$, $Dp-t=0,05$

2003). However, most secondary physics demands students to solve problems concerning ideal or everyday situation (McDermott, 1998). By the tier type of questioning given situations in diverse aspects—*diverse explanation*—students' competence of explaining everyday situation could be evaluated by higher discrimination. For instance, Q16 in Figure 5 demands students to explain the physics concepts of power consumption and parallel connection of resistances, illustrating a circuit with 5 home electric utensils. The aim of this tier unit is to measure whether students could understand that adding more home utensils will lower the total resistance in parallel and consequently raise the current, and that the resistance of each utensil only matters. Because poor understanding of this everyday situation might cause electric accidents from improper usages of home utensils (over-current), the tier unit Q16 should be understood as a whole, not as

a partial concept. However, the tiered-scoring discovered that, at least, 11% of students failed to answer all the 3 scientific choices simultaneously. In addition, the test discrimination increased ($Ct-p$) by 16%, when graded by tiered-scoring.

Tier type 4: Plural variables

Secondary physics such as dynamics contains diverse concepts with plural variables. For example, the conservation of mechanical energy has two independent variables of kinetic energy and potential energy, as presented in Figure 6. According to the law in physics, the sum of kinetic and potential energy of a flying ball is constant on the hypothesis that there exists no air resistance. Since each item in tier unit Q10 examines students on the two mutually dependent variables from a single physics concept of a flying ball, the tier unit deserves tiered-scoring for better test discrimination. Accordingly, this tier type is

An electric circuit below consists of home electric utensils. CHOICES: (a) TV (b) Refrigerator (c) CD Player (d) Laundry Machine (e) Electric Fan

Utensils	TV	Refrigerator	CD Player	Washing machine	Fan
Voltage (V)	220	220	220	220	220
Power consumption (W)	60	850	120	250	40

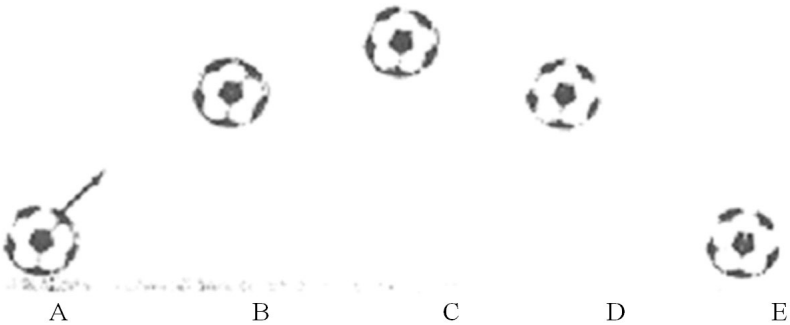
Q16-1. Which one of the five utensils has the highest resistance? (key choice: e)
 Q16-2. Which one of the five utensils has the highest current inside? (key choice: b)
 Q16-3. Find correct descriptions, when the switches are closed one by one. (key choice: d)

-----DESCRIPTIONS-----
 i. The total resistance increases.
 ii. The fuse has more current inside.
 iii. Every five utensils are provided with the same voltage.

 (a) i (b) ii (c) i, ii (d) ii, iii (e) i, ii, iii

Fig. 5 Three-tier unit Q16 with diverse explanation; $Ct-p=0.16$, $Dp-t=0.11$

A soccer ball was kicked with an angle of 45 degrees from ground. A, B, C, D, and E represent the position of a flying ball on the ground. The distances between two of the positions (AB, BC, CD, and DE) are all identical. The ball flew the distance AE of 80 m for 4 seconds. Assuming that there was no aero resistance, and gravitational acceleration is 10% g . CHOICES: (a) $A \rangle B \rangle C \rangle D \rangle E$, (b) $A \langle B \langle C \langle D \langle E$, (c) $A = E \langle B = D \langle C$, (d) $A = E \rangle B = D \rangle C$, (e) $A = B = C = D = E$



Q10-1. Compare the mechanical energy of the ball at each position. (key choice: e)
 Q10-2. Compare the kinetic energy of the ball at each position. (key choice: d)

Fig. 6 Two-tier unit Q10 with plural variables: $Ct-p = 0.14$, $Dp-t = 0.05$

labeled as *plural variables*. The tiered-scoring with the tier type produced the discrimination increase ($Ct-p$) by 14% and the difficulty decrease ($Dp-t$) by 5%.

VI. Conclusion and implication

A new framework for developing tiered tests

Introducing the 4 tier types in the content of secondary physics, our new framework for developing tiered tests proved to enhance conventional multiple-choice tests. As exemplified in the findings, following the framework will satisfy the 3 premises discussed in the literature: a tier unit should examine a singular concept (or an exclusive chunk) of relevant concepts, reuse distracters, and combine items by meaningful tier types. This will thus guide us into how to combine plural items into a tier unit with tier types in secondary physics and how to invent tier types in other subject contents. As a preliminary test, the 23 units that consist of 53 items were developed, and then graded both by tiered-scoring and partial-scoring. The grading comparison affirms that the tiered test in physics approaches to an ideal test that enhances discrimination and

difficulty distribution by tiered-scoring (see Figure 2). Consequently, the two statistical indexes (discrimination increase $Ct-p$ and difficulty decrease $Dp-t$) are identified for deciding effective tier units (see Table 2). According to the item analysis by these indexes, a tiered test featured by the 4 tier types (*directional manipulation, repeated calculation, diverse explanation, and plural variables*) in secondary physics is most likely to improve the shortcomings found in conventional multiple-choice tests (see Figure 3-6). In summary, for those who wish to implement tiered tests, our framework for developing tiered tests recommends them (1) to observe the 3 premises for overcoming partial-scoring issues, (2) to build content-specific tier units referring to the 4 tier types, (3) to conduct pilot tests for item analysis graded both by partial- and tiered-scoring, (4) to discern tier units by examining its discrimination increase ($Ct-p$) and difficulty decrease ($Dp-t$), and (5) to identify other new tier types in the testing content.

A reliable summative assessment of student achievement

Aims of evaluation have been distinguished as

formative or summative (Harlen, 2005). The formative aim is set to contribute to constructive education by providing teachers with remedial treatments of individual students. Among science teachers and educational researchers, recent requirements in school tests are identified to include that a testing system in physics should provide instructional data for designing teachers' lessons and recording phases of students' conceptual change (Kim et al., 2007). These specific requirements that fall into the formative assessment might not be solely implemented by a tiered test due to its tier-scoring. On the other hand, the summative assessment is considered to provide "a summary of achievement ... to those with an interest in students' achievement: mainly parents, other teachers, employers, further and higher education institutions, and the students themselves" (Harlen, 2005, p. 104). In the light, the contemporary outcome-based education still requests the summative assessment, when standardizing an evaluation among larger number of students. Furthermore, Herman and Golan (1993) addressed that the summative and standardized assessment would provide educators with important criteria to contemplate their instruction on students' skills set in a curriculum. For this specific purpose, tiered tests will be effective for measuring how a curriculum is implemented among large-scale students. Because a summative assessment concerns reliable measurements of student achievement (Knight, 2001), the enhancement in discrimination and difficulty distribution of tiered tests can apply to many multiple-choice, standardized tests (e.g., PISA, TIMSS, and CSAT).

Reference

- Andersson, B. (2000). National evaluation for the improvement of science teaching. In R. Millar, J. Leach, & J. Osborne (Eds.), *Improving science education: The contribution of research* (pp. 62–78). Birmingham: Open University Press.
- Atkin, J. M., Coffey, J. E., Moorthy, S., Thibeault, M., & Sato, M. (2005). *Designing everyday assessment in the science classroom*. New York: Teachers College Press.
- Bae, J. (2007, December 23). Massive lawsuits expected over college test. *The Korea Times*. Retrieved May 27, 2009, from http://www.koreatimes.co.kr/www/news/nation/2007/12/113_16003.html.
- Bak, H., & Kwon, J. (1990). A study on analysis of novice's protocol in solving physics problems. *Journal of the Korean Association for Research in Science Education*, 10(1), 57–64.
- Barr, B. B. (1993). Research on problem solving: Elementary school. In D. L. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 237–247). New York: Macmillan Publishing Company.
- Bell, B. (2007). Classroom assessment of science learning. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 965–1006). London: Lawrence Erlbaum Associates, Publishers.
- Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26(1), 41–50.
- Dimes, R. E. (1973). Objective tests and their construction. *Physics Education*, 8, 251–254.
- Engelhart, M. D. (1965). A comparison of several item discrimination indices. *Journal of Educational Measurement*, 2(1), 69–76.
- Gronlund, N. E. (1971). *Measurement and evaluation in teaching*. New York: The Macmillan Company.
- Hargreaves, A., Earl, L., Moore, S., & Manning, S. (2001). *Learning to change: Teaching beyond subjects and standards*. San Francisco: Jossey-Bass.
- Harlen, W. (2005). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (1st ed., pp. 103–118). London: Sage.
- Haslam, F., & Treagust, D. F. (1987). *Diagnosing secondary students' misconceptions of*

photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education*, 21(3), 203–210.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7(3), 238–247.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, 12(4), 20–25.

Hudson, H. T., & Hudson, C. K. (1981). Suggestions on the construction of multiple-choice tests. *American Journal of Physics*, 49(9), 838–841.

Kim, M., Choi, J., & Song, J. (2007). Developing a web-based system for testing students' physics misconceptions (WEBSYSTEM) and its implementation. *Journal of the Korean Association for Research in Science Education*, 27(2), 105–119.

Knight, P. (2001). A briefing on key concepts: Formative and summative, criterion and norm-referenced assessment. New York: Learning and Teaching Support Network.

Kwon, J., & Lee, S. (1987). A comparative analysis of expert's and novice's thinking processes in solving physics problems. *Journal of the Korean Association for Research in Science Education*, 8(1), 43–55.

Lawrenz, F. (2007). Review of science education program evaluation. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 943–963). London: Lawrence Erlbaum Associates, Publishers.

Lee, M. (1998). Development of the three-tier test items for the thinking skills of the scientific inquiry. *Journal of the Korean Association for Research in Science Education*, 18(4), 643–650.

Lee, M., & Pak, S. (1995). A comparative study on multiple choice items of 4 options and 5 options for the thinking skills of the scientific inquiry. *Journal of Science Education in Seoul National University*, 20(1), 151–160.

Lee, M., & Pak, S. (1996). Development of two-tier test items for the thinking skills of the scientific inquiry. *Journal of Science Education in Seoul National University*, 21(1), 19–33.

Linn, R. L., & Gronlund, N. E. (1993). *Measurement and evaluation in teaching* (10th ed.). New Jersey: The Macmillan Company.

Maloney, D. P. (1994). Research on problem solving: Physics. In D. L. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 327–354). New York: MacMillan Publishing Company.

McDermott, L. C. (1998). Students' conceptions and problem solving in mechanics. In A. Tiberghien, E. L. Jossem, & J. Barojas (Eds.), *Connecting research in physics education with teacher education* (pp. 1–6). Ann Arbor: International Commission on Physics Education.

Mitchell, I., & Baird, J. (1986). Teaching, learning and curriculum 1: The influence of content in science. *Research in Science Education*, 16, 141–149.

Park, Y., & Cho, Y. (2005). Analysis of physics problem solving processes of high school students to qualitative and quantitative problems. *Journal of the Korean Association for Research in Science Education*, 25(4), 526–532.

Rebello, N. S., & Zollman, D. A. (2004). The effect of distracters on student performance on the force concept inventory. *American Journal of Physics*, 72(1), 116–125.

Sahlberg, P. (2004). Teaching and globalization. *Managing Global Transitions*, 2(1), 65–83.

Sangster, A. (1992). Computer-based instruction in accounting education. *Accounting Education*, 1(1), 13–32.

Scott, B. L. (1985). A defense of multiple choice tests. *American Journal of Physics*, 53(11), 1035.

Stark, R. (1999). Measuring science standards in Scottish schools: The assessment of achievement programme. *Assessment in Education: Principles, Policy & Practice*, 6(1), 27–41.

Taber, K. S. (2003). Examining structure and

context—Questioning the nature and purpose of summative assessment. *School Science Review*, 85(311), 35–41.

Tamir, P. (1998). Assessment and evaluation in science education: Opportunities to learn and outcomes. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 761–790). London: Kluwer Academic Publishers.

Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching*, 39(4), 283–301.