International Journal of Science Education
Publication details for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713737283

The Positive and Negative Effects of Science Concept Tests on Student Conceptual Understanding
Chun-Yen Chang*a, Ting-Kuang Yehb, James P. Barufaldic
a Department of Earth Sciences, National Taiwan Normal University, Taiwan b Science Education Center, National Taiwan Normal University, Taiwan c Center for Science and Mathematics Education, University of Texas at Austin, Texas, USA

First published on: 21 May 2009

To cite this Article Chang, Chun-Yen , Yeh, Ting-Kuang and Barufaldi, James P.(2010) 'The Positive and Negative Effects of Science Concept Tests on Student Conceptual Understanding', International Journal of Science Education, 32: 2, 265 — 282, First published on: 21 May 2009 (iFirst)

To link to this Article DOI: 10.1080/09500690802650055
URL: http://dx.doi.org/10.1080/09500690802650055

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
RESEARCH REPORT

The Positive and Negative Effects of Science Concept Tests on Student Conceptual Understanding

Chun-Yen Chang\textsuperscript{a,b}, Ting-Kuang Yeh\textsuperscript{a} and James P. Barufaldi\textsuperscript{c}

\textsuperscript{a}Department of Earth Sciences, National Taiwan Normal University, Taiwan; \textsuperscript{b}Science Education Center, National Taiwan Normal University, Taiwan; \textsuperscript{c}Center for Science and Mathematics Education, University of Texas at Austin, Texas, USA

This study explored the phenomenon of testing effect during science concept assessments, including the mechanism behind it and its impact upon a learner’s conceptual understanding. The participants consisted of 208 high school students, in either the 11th or 12th grade. Three types of tests (traditional multiple-choice test, correct concept test, and incorrect concept test) related to the greenhouse effect and global warming were developed to explore the mechanisms underlining the test effect. Interview data analyzed by means of the flow-map method were used to examine the two-week post-test consequences of taking one of these three tests. The results indicated: (1) Traditional tests can affect participants’ long-term memory, both positively and negatively; in addition, when students ponder repeatedly and think harder about highly distracting choices during a test, they may gradually develop new conceptions; (2) Students develop more correct conceptions when more true descriptions are provided on the tests; on the other hand, students develop more misconceptions while completing tests in which more false descriptions of choices are provided. Finally, the results of this study revealed a noteworthy phenomenon that tests, if employed appropriately, may be also an effective instrument for assisting students’ conceptual understanding.

Keywords: Assessment; Conceptual development; Earth science education; Misconception; Secondary school

\*Corresponding author. Department of Earth Sciences/Science Education Center, Graduate Institute of Science Education, National Taiwan Normal University, 88, Section 4, Ting-Chou Road, Taipei 11677, Taiwan. Email: changcy@ntnu.edu.tw
Introduction

One important objective of science education is the enhancement of learners’ scientific literacy, including, but not exclusive to science conceptual understanding, science procedural skills, and problem-solving ability (American Association for the Advancement of Science, 1993). How effectively students solve problems depends upon their domain knowledge, processing skills, and attitudes. Therefore, learners’ conceptual understanding always has been regarded as one of the most important research issues, in terms of evaluating science learning (Eylon & Linn, 1988). Recently, the roles of tests have been regarded as not only an important means for assessment, but also as an influential method for developing students’ conceptual understanding. Roediger and Karpicke (2006) reported that testing, rather than just assessing students’ knowledge levels, is a powerful means by which to improve student learning. A notable phenomenon is that tests may further improve one’s performance on later exams, something which is referred to as the “testing effect” (Roediger & Marsh, 2005; Tulving, 1967).

From the perspective of constructivism, an individual learner’s cognitive structure regarding a specific topic must be actively constructed by himself/herself. Therefore, students’ conceptual understanding regarding a topic might be correctly or incorrectly reconstructed during the process of testing. Roediger and March (2005) determined that, in addition to obtaining positive testing effects, multiple-choice tests may inadvertently lead to the creation of false knowledge, which then appears on later cued-recall performances. In our opinion, the “testing effect” could provide science education researchers with a progressive perspective and practical strategy by which to design tests appropriate for enhancing students’ science learning, rather than merely assessing their current knowledge base. Previous research related to testing effect rarely has been investigated, in terms of its effects and applications in the area of pedagogy. This study attempted to fill this gap by conducting such an inquiry.

Testing Effect during Multiple-Choice Questions

It generally is accepted that multiple-choice questions have certain limitations and drawbacks, such as: (1) student guessing contributing to error variance and diminishing the reliability of the test (Zimmerman & Williams, 2003); (2) selected choices not providing deep insights into student ideas or conceptual understanding (Rollnick & Mahooona, 1999); for example, it is very difficult to develop test items that assess higher cognitive skills (Miller, 1988); and (3) students being forced to choose each answer from among a very limited list of options, thereby preventing them from constructing, organizing and presenting their own answers (Airasian, 2001). However, there also are advantages to administering multiple-choice tests. Multiple-choice tests can be marked easily and are generally not time-consuming to prepare and administer to large classes. In addition, they can produce valid and objective scores for the purpose of assessment. Therefore, although current practice
in science education encourages the use of multiple means to assess student-learning outcomes (Mintzes, Wandersee, & Novak, 2001)—like multiple-choice questions (Gay, 1980), matching items, and short-answer questions (Bean, Searles, Singer, & Cowen, 1990)—the multiple-choice question still plays the primary role in the evaluation of scientific learning among students.

In brief, multiple-choice tests warrant much more attention from pedagogic researchers, to understand whether this commonly used instrument possesses any characteristics (like the testing effect) that might impact student learning and future performance. Therefore, the major purpose of this study was to explore whether and how a multiple-choice test affects students’ conceptual understanding, during and after taking the test.

**Mechanisms behind the Testing Effect**

Generally, considering the mechanism of testing effect, two series of mental operation are closely related. One mental operation is the retention effect, which is more implicit. Reading the context/choices in a test or retrieving related information by cues might activate relative memory and modify the memory trace of target items and increase the probability of a successful retrieval later (Kang, McDermott, & Roediger, 2007). The other line of the testing effect mechanism is more explicit and requires the implication of consciousness. This usually occurs when students are unfamiliar with or confused about some or all of the response options provided for a given question in a multiple-choice test. In this situation when none of the offered choices are congruent with what the students have already known, they are, as a result, forced to improvise and apply other mental strategies beyond simple recall, such as reasoning, logical thinking, and elimination. To eliminate options, students will need to distinguish between those options that are seemingly ludicrous from those that are not. However, during this eliminating process, students’ perception on the best possible choice might turn out to be the incorrect answer. This entire process (if they choose a wrong option finally) allows students to repeatedly ponder and think deeply on each of the options before their final selection which thereby may increase their belief and familiarities with certain incorrect information. Ultimately, from this eliminating process, students may construct new knowledge, both correct and incorrect.

The aforementioned speculation reveals that the testing effect might come into play while students ponder over how to select one response among several. Furthermore, factors like distraction and the correctness of response descriptions (key factors vs. distracters) might influence their construction of new positive and negative outcomes when they are unfamiliar with a question. Therefore, we attempted not only to examine the existence of a testing effect, but also to design three different types of test, based upon the level of correctness of the descriptions, in order to explore if the correctness of descriptions in a question is one of the important factors influencing the correctness of subjects’ conceptual understanding.
Neuroscience and Informed Science Education Research

One of the most important endeavors in science is to seek for the truth and reality using scientific method. Therefore, researchers in the areas of science and science education should explore and explain the nature of explicit behavior from different perspectives as far as possible. However, limited by the development of technology and methodology, mental operations have been regarded as a black box which greatly perplexed researchers for many years. A new research domain, neuroscience, emerges to provide more progressive and concrete evidence for human molecular/behavioral mechanism. Nevertheless, the development of current neuroscientific technology is undeniably insufficient for researchers to understand learners’ actual mental activities. This may be one of the major reasons that educationists or psychologists hesitate to incorporate neuroscience-related theories into their research. Robins, Gosling, and Craik (1998) reported that the number of times neuroscience is cited in important journals of pedagogy and psychology is insignificant.

In our opinion, the integration of neuroscience with science education might provide researchers in different areas with windows of opportunity through which they can seek ‘the truth’ collaboratively. Extrinsic behavioral pedagogy-oriented attributes—academic achievement, creativity, and problem-solving ability—have been deemed the output of mental operations (neuroscience-oriented). Therefore, there soon might be a trend towards applying the research results of neuroscience to pedagogic areas, and vice versa.

Purpose of the Study

Depending on the target educational domains, this study attempted to carry out the integration of science education and neuroscience initially. The major purpose of this study was to evaluate the testing effect during science concept assessments. Three different types of tests were designed based upon the hypothesis that correctness of the descriptions may impact students’ conceptual understanding during assessments, with the aim to explore the mechanism of testing effect. Furthermore, empirical evidence and latest theories from the area of neuroscience were applied to support our educational finding. Although applying research results from neuroscience to explain difficult learning behaviors may not be sufficient, it is anticipated that the integration of these two fields can shed light on future directions for researchers in various fields.

Methods

Test Topic: Greenhouse effect and global warming

Contemporary science education highlights the importance of global change; while, at the same time, it envisions that students acquire key related knowledge and develop their interest in and attention to our environment from curricula (Chang, 2005; Chang, Lee, & Yeh, 2006). In Taiwan, global warming and greenhouse effect
are covered extensively in the subjects of physics, chemistry, geography, and earth sciences, as shown in the Science and Life Technology Curriculum Standards (SaLTS, Grades 1–9) and Tentative Earth Science Curriculum Guidelines (TESCG, Grades 10–12) as summarized in Table 1. Since the greenhouse effect/global warming is an integral component of global change, examining students’ cognitive structure and misconceptions relating to this topic has been regarded one of the most critical issues in earth sciences education. This type of inquiry aims not only to identify students’ learning difficulties, but also to further modify their misconceptions regarding this topic. Because the phenomena of global warming and greenhouse effect cannot be observed directly by students, misconceptions might be developed more easily during the process of student learning (Boyes & Stanisstreet, 1993; Francis, Boyes, Qualter, & Stanisstreet, 1993; Rye, Rubba, & Wiesenmayer, 1997). Furthermore, we hypothesized that the misconceptions related to greenhouse effect and global warming are very likely to form during the process of a “concept test”.

Probing Students’ Cognitive Structure using Flow-Map Method

Educators and cognitive scientists have attempted to represent pre-acquired knowledge in terms of “cognitive structure”, a hypothetical construct that demonstrates the extent of concepts and their relationships with long-term memory. Beyond this, cognitive structure could be a fundamental step towards a better understanding of how students construct (or reconstruct) knowledge (Tsai & Huang, 2002). Anderson and Demetrius (1993) have proposed the “flow-map method” to probe learners’ cognitive structures. As suggested by Tsai and Huang (2002), the flow-map method may be considered the most powerful method by which to represent learners’ cognitive structures, while requiring minimal intervention by the interviewer and the least inference for its construction.

This study, therefore, utilized the flow-map method as a baseline instrument to probe students’ cognitive structures. To elicit students to represent their cognitive structures regarding the global warming and the greenhouse effect as completely as possible, they were asked non-directive flow-map questions by well-trained

<table>
<thead>
<tr>
<th>SaLTS component</th>
<th>TESCG component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Themes</td>
<td>Themes</td>
</tr>
<tr>
<td>Interactions</td>
<td>Earth</td>
</tr>
<tr>
<td>within nature</td>
<td>environment</td>
</tr>
<tr>
<td>(9th grade)</td>
<td>change</td>
</tr>
<tr>
<td>Global change</td>
<td>Climate change</td>
</tr>
<tr>
<td>– Greenhouse effect</td>
<td>– Climate change in history</td>
</tr>
<tr>
<td>– Ozone</td>
<td>and its impacts</td>
</tr>
<tr>
<td>– Climate change</td>
<td>(10th grade)</td>
</tr>
<tr>
<td></td>
<td>– Short-term climate change</td>
</tr>
</tbody>
</table>

SaLTS: Science and Life Technology curriculum Standards  
TESCG: Tentative Earth Science Curriculum Guidelines
researchers as follows: “Please tell me what you know about global warming and the greenhouse effect”; “Could you elaborate more about the ideas you have just mentioned?”; and “Could you specify the relationships that exist among the ideas that you have just talked about?” These questions were presented in a relatively non-directive way in which fewer hints were provided about scientific concepts for learners (Tsai & Huang, 2002). The learners’ open-ended answers then were transcribed into flow maps. Figure 1 shows a student’s flow map about global warming and the greenhouse effect. In essence, the flow map is constructed from entering statements in a sequence expressed by the learner. The sequence of discourse is examined and linked by connecting arrows. For example, the student’s narrative mapped in Figure 1 shows a sequential pattern beginning with the cause of the greenhouse effect. The student also stated the ozone hole causes global warming, which was a misconception. Consequently, the subject in Figure 1 retrieved five concepts, among which, four concepts are correct and one is a misconception.

Types of Test for Investigating the Mechanisms that Influence the Testing Effect

To initially investigate the mechanisms behind the testing effect, three types of tests were developed. Based upon the aforementioned reasoning, we hypothesized that the distraction and correctness of descriptions in questions will play an influential role in students’ conceptual understanding, when they complete a test in which they are unfamiliar with several response options. Therefore, we not only designed a traditional **multiple-choice test** (MCT) for examining the existence of the testing effect, we also divided the options (distracters and key) of this multiple-choice test

1. **$CO_2$** causes the **greenhouse effect**
2. The **ozone hole** causes **global warming**
3. **Greenhouse gases** absorb **long-wave radiation**
4. **Earth’s surface** releases **long-wave radiation**
5. **Solar radiation** is **short-wave radiation**

Figure 1. A student’s flow map about the greenhouse effect and global warming

Note: In this example, the subject retrieved 5 concepts, of which 4 concepts are correct and 1 (number 2) is misconception. The sequence of discourse is examined and linked by connecting arrows.
into two new true/false type tests, namely a correct-concept test (CCT) and an incorrect-concept test (ICT):

(1) **Multiple-Choice Test** (MCT): Twenty items were developed for a MCT, testing concepts related to global warming and the greenhouse effect, in accordance with the *Tentative Earth Science Curriculum Guidelines* (TESCG) that students previously had been taught in a senior high school level course. On the MCT test, the number of options usually available, per question, ranges from two to three, yet only one (potentially correct) key option is provided. Table 2 illustrates examples of the items developed for the three types of test.

(2) **Correct-Concept Test** (CCT): The options (distracters and key) in each item of the MCT were further assigned to two tests with either the true/false type, named the CCT and the ICT. The CCT, presented using a listing of all correct scientific descriptions, combined and modified all key options from each item in the MCT to a 20-item true/false type test.

(3) **Incorrect-Concept Test** (ICT): Contrary to the CCT, the ICT, presented using a listing of all incorrect scientific descriptions, combined and modified all distracters from each items of the MCT to a 20-item true/false type test.

**Reliability and Validity of the Instrument**

The reliability of the flow-map method was determined by a second independent researcher to code conceptions from the students’ narratives. The Pearson correlation coefficient (r) for each student ranged from 0.86 to 0.96, and the average intercoder agreement was 0.91. In addition, all three different types of test—MCT, ICT, and CCT—were examined and content validated by four professors. Two of the professors were in the field of science education, while the other two were in meteorology.

**Participants**

The participants in this study were selected from a public senior high school in Taiwan. The total sample was comprised of four classes of 11th graders and four classes of 12th graders (n = 208). Students were divided into four different groups (A, B, C, and D). Each group included one Grade 11 and one Grade 12 class. All participants had been taught about global warming and the greenhouse effect during their 9th and 10th grade studies.

**Research Design and Procedures**

The data collection consisted of three phases: (1) exploring students’ pre-test cognitive structure using the flow-map method, (2) administering different tests (treatments), and (3) exploring students’ post-test cognitive structure using the flow-map method. During the first phase, all subjects were interviewed using the flow-map
method, for the purpose of exploring their initial cognitive structure before the test as a baseline group condition. In phase two, the classes were randomly assigned to four groups and asked to complete various kinds of test. The CCT was administered to Group A \( (n = 54) \), the ICT to Group B \( (n = 46) \), the MCT to Group C \( (n = 60) \), and no test (the N-T, control condition) to Group D \( (n = 48) \). Finally, all subjects were interviewed again, using the flow-map method, two weeks after all the tests had been administered. Figure 2 presents the procedures and overall study architecture.

**Data Analysis**

A number of variables, such as participants’ prior enrollments in earth sciences courses and the duration of research, were held constant. The major independent variable in this study was the format of the conception test (CCT, ICT, MCT, and N-T condition) and the dependent variables were related to student conceptual understanding, in terms of global warming and the greenhouse effect. To further understand how students’ concepts would be affected subconsciously by the intervention, univariate analysis of variance (ANOVA) was conducted, with post-test scores as the dependent variable, to identify any significance differences between the four interventions. The assumptions for ANOVA first were checked to ensure that they were in accordance with the actual data: the measures for each dependent variable were determined to be independent; and Levene’s tests for the homogeneity of variance indicated that variance for each dependent variable was equal across the groups, suggesting that it was appropriate to conduct ANOVA.

To meet contemporary calls for improvement in the interpretation and reporting of quantitative research in education (Rennie, 1998; Thompson, 1996), this study
testing effects on conceptual understanding

reports practical significance (effect magnitudes) along with each statistical significance test. The effect size index \( f \) was used, since it is more appropriate for the analysis of variance or covariance (Cohen, 1988). According to Cohen's rough characterization (1988, pp. 284–288), \( f = 0.1 \) is deemed to be a small effect size, \( f = 0.25 \) a medium effect size, and \( f = 0.4 \) a large effect size.

This data presentation method is quite important, in terms of interpreting research results. Researchers have cautioned against the use of statistical significance testing alone when making statistical inferences (Cohen, 1988; Daniel, 1998; McLean & Ernest, 1998), primarily because the computation of statistical significance is highly dependent upon sample size. Moreover, it is quite common to observe a statistical significance with a large sample size, even if there is little practical effect. Consequently, in addition to reporting the results of statistical significance testing, effect magnitudes also are reported here. All analyses were conducted using Statistical Package for Social Sciences (SPSS) version 13.0.

result

students' conceptions before and after the test

During the initial flow-map free response interview, participants in the CCT group held an average of 6.74 concepts, participants in the ICT group 6.50 concepts,
participants in the MCT group 6.32 concepts, and participants in the N-T (control) group 6.50 concepts. As shown in the first row of Table 3, the four groups of students were not significantly different in their baseline (pre-test) level of conceptual retrieving; \( F(3, 206) = 0.35, p = 0.79, \eta^2 = 0.005, f = 0.071 \). The effect size

Table 2. Some items of different types of concept tests on the topics of global warming and greenhouse effect

<table>
<thead>
<tr>
<th>Tests</th>
<th>Types of test</th>
<th>Content examples</th>
</tr>
</thead>
</table>
| MCT   | Multiple-choice | - The greenhouse effect is caused primarily by: (1) Water vapor. Water vapor is the most significant greenhouse gas, and is responsible for roughly 95% of the greenhouse effect. (2) Carbon Dioxide. The concentration of manmade CO\(_2\) has increased since the start of the industrial revolution and plays the primary role in the greenhouse effect. (3) Chlorofluorocarbons (CSFs). CSFs can destroy ozone molecules, resulting in the so-called “ozone hole” phenomenon, and contribute primarily to the greenhouse effect.  
   - The greenhouse effect is a phenomenon that is (1) Natural: The greenhouse effect has existed since the formation of the Earth, with its water vapors and original atmosphere, over four billion years ago. (2) Manmade: the greenhouse effect has contributed to a steadily elevating average temperature of the Earth’s near-surface air since the start of the industrial revolution.  
   - Is the greenhouse effect a phenomenon that is unique to the Earth? (1) Yes. The greenhouse effect does not exist on other planets, like Venus and Mars. (2) No. The greenhouse effect occurs not only on Earth, but also on other planets, like Venus and Mars.  
   - Is the Ozone hole the main reason for global warming? (a) Yes. The ozone hole over the South Pole has been the longest lasting on record in recent years. This depletion of the ozone layer allows for more sun radiation to reach the Earth’s surface, increasing the temperature on Earth and causing global warming. (b) No. The ozone hole over the South Pole has been the longest lasting on record in recent years. This depletion of the ozone layer allows for more ultraviolet radiation to reach the Earth’s surface. The ultraviolet (UV) radiation is harmful to organisms, but is not the main reason for global warming. |
| ICT   | True/False   | - The concentration of manmade CO\(_2\) has increased rapidly and has played the primary role in the greenhouse effect since the start of the industrial revolution.  
   - The greenhouse effect has contributed to a steady elevation in the average temperature of the Earth’s near-surface air since the start of the industrial revolution.  
   - The greenhouse effect is a unique phenomenon on Earth, and does not exist on other planets in our solar system, like Venus and Mars.  
   - The ozone hole over the South Pole has been the longest lasting on record in recent years. This depletion of the ozone layer has allowed more of the sun’s radiation to reach the Earth’s surface, increasing the temperature of the Earth and causing global warming. |
Testing Effects on Conceptual Understanding

The index was calculated using the eta square statistic ($\eta^2 = 0.005$) as $f = 0.07$, and approached a small effect size for subject group.

Two weeks after the test, the final flow-map interview showed that students, on average, held 10.02, 10.07, 9.52, and 6.85 concepts in the CCT, ICT, MCT, and N-T groups, respectively, differences that were statistically significant and associated with a large effect size; $F (3, 208) = 13.45$, $p < 0.01$, $\eta^2 = 0.161$, $f = 0.438$. As expected, the number of concepts was significantly higher for each of the three test groups (CCT, ICT, and MCT) relative to the control group ($p < 0.01$).
findings, therefore, revealed that tests themselves might be regarded as efficient stimuli in light of the increased numbers of students’ concepts, which can be retrieved from learners’ cognitive structure after their completion of the tests.

As shown in Table 2, results from the final flow-map free response interview also indicated that completing a test increased the number of both correct and incorrect concepts. Thus, this may indicate that there were either positive or negative effects on students while taking tests which consisted of either true or false or multiple-choice questions. The results can be further explored when we analyze the correct/incorrect concepts in a greater detail as follows.

**Impact of Correct/Incorrect Choices**

In this study, ANOVA was computed for pre-test data, and no significant differences were identified ($p > 0.05$), suggesting that the four groups of students start with similar background knowledge about global warming and the greenhouse effect. After completing the test, a significant difference was found in the number of correct concepts among the students, $F(3, 206) = 9.10$, $p < 0.01$, $\eta^2 = 0.114$, $f = 0.359$, and post-hoc Scheffe analysis revealed that the CCT group, as compared to the ICT, MCTs, and N-T group, experienced a much greater increase in the number of correct concepts. There also was a significant difference in the increased number of incorrect concepts between the four conditions after testing, $F(3, 206) = 7.20$, $p < 0.01$, $\eta^2 = 0.092$, $f = 0.318$, with the ICT group found to have a much greater number of incorrect concepts post-test.

**Discussion**

In this prospective trial comparing different types of test, versus each other and versus a control, no-test condition, it is noteworthy that the CCT yielded the greatest increase in correct concepts, versus the other three conditions; and that the ICT appeared to exert the opposite effect, increasing the number of incorrect concepts relative to other test conditions as illustrated in Figure 3.

Contemporary reforms in science education have highlighted the importance of assessing students’ cognitive abilities and level of learning achievement. In the assessment of science learning and science concepts, multiple-choice tests have been used most extensively. It was also noted that the pre- and post-test coefficients of effect size in three types of tests (1.13 for CCT, 1.39 for ICT, and 1.17 for MCT) were approximately equal in strength comparing with instructions learning in science education, such as 0.45–1.94 for inquiry-based instruction (Marx et al., 2004), and 0.40–1.02 for computer assisted instruction (Chang, 2003; Chang & Tsai, 2005). This result revealed that tests can be regarded as a powerful means of learning.

However, we also found both positive and negative effects were exposed from the tests employed, when more true statements are provided in a test (CCT), a more
positive than negative testing effect may emerge; whereas, when more false statements are included in a test (ICT), a more negative than positive testing effect results. The test itself can stimulate students to both learn and “misunderstand”.

For students’ conception acquisition, one of the most important reasons for a testing effect may be reinforcement. From a psychological perspective, cues in tests help learners to retrieve relevant knowledge they already have (Roediger & Marsh, 2005). Multiple-choice tests provide more cues than other types of tests, like open-ended tests, because the answers actually are provided. Therefore, it is reasonable to speculate that students might retrieve more relevant information from their memory while answering multiple-choice tests. Information retrieved by cues might provide an amplification effect similar to repeated practice, which reinforces relevant knowledge/memory stored in the brain, so that future recall becomes relatively effortless. In addition, multiple-choice tests consist of questions with several false statement choices and only one correct answer. In other words, the distracters in multiple-choice tests expose students to a considerable amount of incorrect information. Reading the choices might increase the student’s familiarity with the subject, but also cause them to misperceive erroneous information as correct, causing negative consequences later, especially among students who were less familiar with the subject area in the first place.

Mechanisms of Testing Effects from a Neuroscience Perspective

“Memory consolidation” has been studied for more than a century (McGaugh, 2000). It refers to the process whereby a memory becomes increasingly stable and resilient to disruption with the continued passage of time. In addition, it is widely accepted that memory consolidation is dependent upon changes in synaptic efficiency that permit strengthening of associations between neurons (Lynch, 2004). Therefore, explained from a neuroscience perspective, the storage of conceptions or memory may depend upon changes in synaptic strength, termed “synaptic plasticity”. The explanation for this that has received the most attention relates to the “long-term potentiation phase” (LTP-phase), a cellular model of activity-dependent enhancement of synapses. The LTP-phase has been used widely as a model for the investigation of long-term memory mechanisms (Bailey & Kandel, 1993; Lynch, 2004). As a result, it is reasonable to infer that reading the choices in a test or retrieving related information by cues might arouse weak synaptic connections and turn these connections into strengths. This may, therefore, make it easier for learners to remember something later, during a free response task.

Another dramatic phenomenon that might exist is that established memories, once recalled, become labile and sensitive to disruption, requiring “reconsolidation” to become permanent (Lee, Everitt, & Thomas, 2004; Nader, Schafe, & Le Doux, 2000). Recent findings supporting this reconsolidation hypothesis have been associated with procedural learning (skilled sensory and motor tasks) and declarative memory (Forcato et al., 2007). In other words, learners’ conceptions and memories might become unstable and labile during the retrieval process.
required during a test. If such learners then are exposed to numerous incorrect, but reasonable "facts", they may reconstruct their cognitive structure using false information. Eventually, this new information (especially among those who initially were more confused in the subject area) can negatively impact their future recall and reasoning.

Related research using functional magnetic resonance imaging (f-MRI) substantiates certain brain activities being based upon different types or amounts of cues. Kohler, McIntosh, Moscovitch, and Winocur (1998) reported that episodic retrieval requires at least two main components. The first is mediated by the prefrontal cortex, and represents resource-demanding processes that are needed to maintain and implement strategic aspects of retrieval. An example of this is the retrieval mode, which involves searching, monitoring, and coordinating competing task demands. The second component is mediated by the medial temporal lobes/hippocampus (MTL/H), and involves the relatively automatic re-activation of memory traces, resulting from their interaction with memory hints (Fernandes, Pacurar, Moscovitch, & Grady, 2006). It is consistent with our previous arguments that, under some conditions, learners not only extract knowledge from their brain (prefrontal cortex demanded), they also may have a new problem-solving experience (MTL/H demanded). This problem-solving process, enabling students repetitive pondering about the accuracy of greater distracters in a short period of time, will help students to construct new long-term memories. Frey, Huang, and Kandel (1993) stimulated the hippocampus’ CA1 area with a 3–5 times higher frequency within a short period of time, an LTP-phase was produced. Hippocampus has been demonstrated to play the most crucial role for human long-term memory formation. Students rapid pondering in some distracters and think deeply may stimulate long-term memory formation more effectively.

Implications and Limitations

In recent years, public teachers and researchers generally acknowledge that educational assessments play fundamental roles in educational practices at different levels, from learning diagnosis in the classrooms to the determination of quality of educational institutes; however, the result of this study revealed that the level of distraction and the correctness of statements on a test may affect students’ conceptual understanding after they take the test. Therefore, teachers and educational researchers should be more cautious in testing practices. As a result, how to design an appropriate instrument for assessment becomes an important issue for science education researchers. Instructors should exercise extra precautions while designing questions, so that not only the goal of the assessment can be achieved, but also conceptual understanding can be enhanced. A couple of ways to develop questions that better assess and enhance students’ learning include utilizing diverse assessment strategies rather than solely depending upon one particular strategy, avoiding overuse of false information on tests, and providing students with the correct answers immediately followed by the test.
The result of this study supports the claim of contemporary educational practices, calling for the adopting of diverse range strategies and systematic ways to evaluate students’ learning and to find out how close students are to educational goals (Mintzes, Wandersee, & Novak, 2001; Tsai, 2001). By employing multiple assessment practices, including open-ended, multiple-choice, and interview approaches, teachers can gauge students’ understandings more fairly and with less negative consequences. Good examples of such forms of assessments are the TIMSS and PISA which are the international assessment projects and programs that have contributed significant efforts to the development of innovative assessment for science education. Open-ended questions and interview strategies that elicit students’ constructed responses and give students a higher degree of freedom in reasoning may serve as a less negative testing effect for learners. Although multiple-choice tests, due to the existence of deterministic answers, can be marked easily and administered to large classes and can produce valid, objective scores for the purpose of assessment, they also provide more distractions and incorrect statements than other forms of assessment.

Neuroscience-related findings have supported that corrections made immediately followed by the test can reduce the formation of misconceptions, both in misconception disrupting and interrupting misconception formation. Based upon the reconsolidation hypothesis as aforementioned, learners’ memories will become labile and sensitive once recalled/retrieved. In other words, corrections made immediately follow by the test can activate related memory and provide an opportunity for students to replace the misconception (test made) with the correct concept (correcting).

For the interrupting misconception formation effect, consolidation of memory is determined by time-dependent processes in specific brain states, like certain stages of wakefulness and sleep (Graves, Pack, & Abel, 2001). It is well known that protein expression is required for the formation of long-term memory (Bliss & Collingridge, 1993; Steward & Schuman, 2001). One way that sleep might facilitate long-term memory is through its effects on protein signaling pathways during rapid-eye-movement (REM) sleep (Graves et al., 2001). Consequently, immediately correcting false impressions after a test likely would hinder the formation of misconceptions in long-term memory, and help to construct a correct cognitive structure.

In educational practice, corrections made immediately followed by the test are essential to improve and accelerate learning. Timely feedbacks after assessments not only help students to understand why an unacceptable alternative is “undesirable” and to grasp key concepts but also provide them an opportunity to strengthen their weaknesses. To meet the goal of adapting needed feedbacks to individual differences, there had been techniques developed to help students effectively gather more timely and individualized feedbacks/corrections during an assessment (Graesser, Chipman, Haynes, & Olney, 2005; Tseng & Tsai, 2007). However, one of the most challenging issues here is how to enable computers to identify students’ concepts accurately and give them feedbacks promptly (Wang, Chang, & Li, 2008). In order to enhance students’ learning, how to employ feedback acts as task engaging will be
a great follow-up study topic. Multiple-choice items, without unconstrained responses such as essay writing, may raise the efficiency in putting these techniques into practice rather than the other types of assessments.

In conclusion, integrating psychologically oriented and neuroscientific evidence, in this study we have been able to identify somewhat unexpected consequences of multiple-choice test taking, consequences that are both positive and negative. In particular, the correctness of students’ conceptual understanding was appeared to be influenced by the response options offered during the test. The results of this study certainly suggest that teachers and education researchers should be more cautious when designing tests, so as to enhance positive effects and avoid negative consequences related to conceptual understanding. However, the mechanisms behind these consequences remain relatively obscure and warrant further investigation.

There are a number of limitations in this study. First, in order to elicit learners’ cognitive structures as complete as possible, students were asked to elaborate their thoughts related to greenhouse effect and global warming with different times according to their response which is impacted by individual difference in personality. This may result in a different influence on students’ response. It is our opinion that learning is not merely memorizing knowledge but has to apply recall/retrieval knowledge into solving real-world problems. Therefore, to analyze students’ change of high level ability after testing will be a reasonable follow-up study (it is also our underway endeavor) and will more consist with envisions of science education. Furthermore, the small- to middle-sample size and the middle- to large-effect size in this study not only remind us of the need to generalize results more cautiously in a practical sense, but also to suggest further replicated studies conducted in this research area, which are presently being investigated in Taiwan.

Acknowledgement

This paper is based on work supported by the National Science Council of Taiwan under contract NSC 96-2511-S-003-022-MY2.

References


