JOURNAL OF BALTIC SCIENCE EDUCATION
ISSN 1648-3898

EVALUATING THREE GRADING METHODS IN MIDDLE SCHOOL SCIENCE CLASSROOMS

Bo Zhang,
Jacob Misiak

Abstract. *This study aims to fill a gap in middle school science education that no direct comparison has been conducted on three popular grading methods - point based, rubric based, and rubric plus written feedback based - on science achievement and motivation. Participants in this quasi-experimental study are 136 7th and 8th grade students from Midwest USA. Results indicate that for both grades, the written feedback group performs significantly better than the point-based and the rubric-only groups on both achievement and motivation measures. However, science achievement under rubric-based grading but without written feedback is not different from that under the traditional points-based grading, underscoring the important extra power that written feedback delivers. This study recommends that written feedback be highly relevant (to the academic standard being assessed), limited in number (to the major misconceptions students have), and giving a second chance (for students to act on to improve learning).*

Key words: *feedback, formative assessment, rubric, motivation, science achievement.*

Bo Zhang
University of Wisconsin - Milwaukee, USA
Jacob Misiak
New Berlin School District, New Berlin, USA

Introduction

The grading practices in most science classrooms today still rely on a point system under which lack of understanding is communicated through point loss. Popular as it is, point-based feedback suffers from major limitations. First, it fails to clearly communicate to students what they are able to accomplish and how to reach the next level. Second, point-based feedback contains high level of subjectivity, thus generates unreliable scores (Marzano, 2002). Finally, point-based feedback may have a negative effect on student motivation. When achieving high points is the goal of learing activities, students are less likely to pick challenging assignments in order to avoid the increased risk of failure (Kage, 1991). Since engagement in learning is closely associated with the availability of challenging activities (Thomas et al., 1993), point-based feedback usually fails to motivate students to reach their highest possible level.

Rubric feedback has been used to partially address the above limitations. A rubric tells students what is assessed, where they are at, and how to attain a preset educational standard. Rubric feedback is formative. It allows for the adjustment of student knowledge over time (Black & Wiliam, 2004). With rubrics feedback, students can integrate new knowledge, make corrections, and continue the learning process towards the mastery of standards. In addition, as good rubrics clearly define how each student response should be read, the subjectivity in scoring can be greatly reduced (Marzano, 2002).

Written feedback under a standard-based system may hold the highest potential to improve student learning. Standard-based grading measures student achievement in terms of their progress on pre-determined standards. That is to say, instead of assign points to a test, for instance, 90 out of 100 points, a number of scores or descriptors (e.g., beginning, developing, mastery) that represent proficiency levels will be provided to assess standards the test is designed to measure. This way, grading will be more towards knowledge and skills than towards points. According to Hattie & Timperley (2007), effective feedback should give clear answer to the following three questions: 1) where am I going? 2) how am I going? and 3) where should I go next? Chin (2006) described four ways to give feedback specifically for science classrooms. Overall, point-based numerical feedback presents students with a very general but vague idea of how they are going. It fails to point out where to go next. A well-crafted rubric will be able to answer the above three questions in such a standardized way that all students will be evaluated against the same criteria. However, it still lacks the specific guide for each individual student as to where and how to get to the next level. Individualized written feedback coupled with standard-based rubric has the potential to address the above three questions in a tailor-made manner for each student.

Problem of Research

So far no direct comparison of the three grading practices (i.e., point-based, rubric-based, and rubric plus written feedback based) has been conducted in science teaching. While it seems reasonable to assume the extra benefits of rubric-based grading and written feedback over the traditional point-based grading, empirical evidence is scarce. This study aims to fill that gap by investigating the effect of three grading methods on both achievement and motivation in middle-school science teaching.

Research Focus

Several studies have identified the aspects of written feedback that will increase student knowledge. Elawar and Cormo (1985) determined that written feedback that focuses on specific errors and suggestions on how to improve problem solving strategies leads to learning increases. Clymer and Wiliam (2007) have found similar results that when feedback is centered on what the learner needs to improve and how the improvements can be made, learning increases. Feedback related to the process of a task also encourages persistence on challenging tasks and increases student motivation (Mueller & Dweck, 1998). In a study conducted by Butler and Nisan (1986), participants were exposed to three forms of feedback: no feedback, numerical grades, and comment feedback. Those who received no feedback demonstrated a decrease in their interest in learning. Students who received numerical feedback scored high on quantitative tasks but low on tasks that required divergent thinking. The comment feedback group demonstrated the highest interest and attained the highest achievement. Collectively, these studies illustrate that carefully prepared written feedback can effectively identify specific strengths and weaknesses, hence target the exact needs of each student. On the other hand, an alternative view of written feedback is that it may be too much for both students and teachers. Students sometimes feel written feedback focuses on the details not necessarily important for the completion of the task, the so-called "hyperspecific corrections" by Willingham (1990). Teachers at the same time feel they spend tremendous amount of time on giving written feedback, which not many students actually use to further their learning (Bailey & Garner, 2010; Glover & Brown, 2006).

Different grading practices may affect student motivation in learning science. Feedback can be a means for sustaining intrinsic motivation, enhancing students' psychological need for independence, and supporting the need for competency. Students want to make sense of the natural world; therefore, the need for understanding creates a high level of intrinsic motivation to learn science (Deckers, 2004). In that sense, there is little need to provide extra external reward (Deci, Ryan & Koestner, 1999), such as a high test score. On the contrary, research has also shown when a student determines learning occurs for a high grade, the reason for learning becomes more externally controlled, which actually undermines the intrinsic motivation (Kast & Connor, 1988)

When written feedback is phrased around progress toward specific standards, students can utilize the information to adapt their conceptualization of relevant knowledge or to correct major misconceptions. Written feedback can also be tailored to individual students, thus it is reasonable to assume written feedback will be more practical and meaningful than information communicated through grades or rubrics only. As a result of additional formative feedback, students will have the necessary guidance to work consistently towards achieving mastery of standards.

Methodology of Research

General Background of Research

This study used a between-group, pre and post design. Students came from three groups. In the rubric only condition, they were graded based on standard-based rubrics, which communicated learning goals, criteria to demonstrate standards. The rubric also included a system to track students' current level of development by a four-point scale: basic, approaches standard, meets standard, and exceeds standard. In the rubric plus written feedback condition, students received additional written feedback along with a rubric rating. The written feedback included a suggestion or question that would allow students to continue their standard development. Two examples of written feedback are given in Figure 1. In the first example, the student demonstrated a misconception about why areas differed in temperature, therefore the feedback attempted to direct the student toward their observations and data about the angle of sun's rays and how that related to difference in heat. In addition, the misconception was addressed by pointing out that in the model used in class, the distance between the Earth and sun did not change. The second response demonstrated the student had made a connection between the lab activity and heat; therefore the student was asked to take the next step and connect their understanding to the main concept being addressed in the activity – when the energy from the sun hits earth at an angle, the energy is distributed over a larger area.

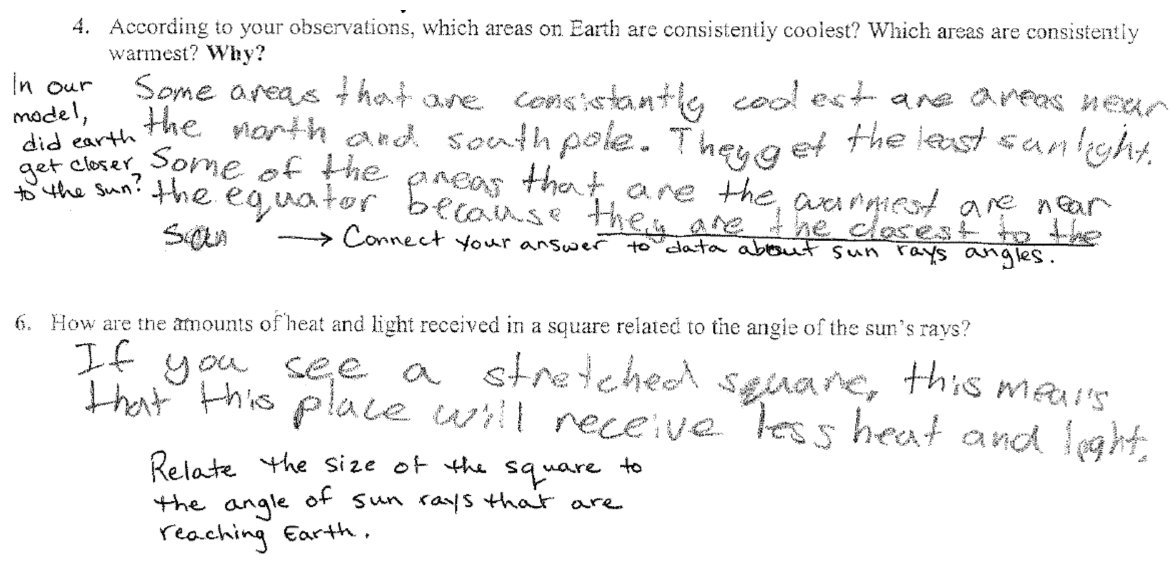


Figure 1: An example of feedback given to students.

All students in the standard-based system were taught by the same teacher in an inquiry-based educational setting and using the same instructional tasks. Furthermore, students were evaluated using the same forms of assessments. A standard-based grading system was implemented to assess the concept development of eighth graders in the area of magnetism and electricity and seventh grade in astronomy. Formative assessment occurred throughout each unit. Students developed content and inquiry standards through performing laboratory procedures, conducting experiments, analyzing data, and exploring research questions. Tasks were leveled to guide students through concept development. After each completed task, students in the two standard-based groups responded to open-ended summary questions to communicate their progress toward standards. The summary questions and work completed on tasks were evaluated, and students were updated on their progress. After each evaluation, these students had the opportunity to make corrections and update any work which did not exceed standard level. The instruction of each area took 4 weeks.

ISSN 1648–3898

Sample of Research

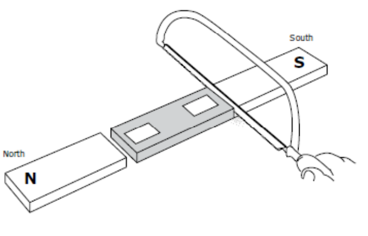
The participants for the study were 136 students from a suburban middle school in the Midwest area of the United States. The sample included 68 seventh graders and 68 eighth graders. Each grade level consisted of three separate natural classes, each serving as one condition of three grading methods. Two of the classes were taught by the same teacher under rubric-based grading. In addition, written feedback was provided to one class randomly selected from the above two. The third class was used for the total-point based grading comparison, which was taught by a different teacher. Class sizes range from 21 to 24. All classes contained students with special needs. Students in the sample were 93% Caucasian, 4% Asian, and 3% African American.

Instrument and Procedures

To determine students' achievement in the rubric and written feedback conditions, students were pre and post assessed on their development toward identified content standards. Students in all three groups were also measured using an assessment comprised of high-quality standardized test items. The assessment contained released items from large-scale assessments in science, including the National Assessment of Educational Progress (NAEP), Trends in International Math and Science Study (TIMSS), and state standardized tests. As shown in Table 1, great efforts were made to ensure test questions were relevant to the standards being addressed in each unit for both grades. Overall, these questions required students to explain their content knowledge in short constructed response format, to respond to questions using diagrams and tables, or to answer multiple-choice questions. The total point earned by each student was used as the common scale to compare the performance across three groups.

Table 1. Sample questions from the standardized tests.

1. (7th grade) Explain why daylight and darkness occur on Earth?
2. (7th grade) Which of the following is an important factor in explaining why seasons occur on Earth?
   a) Earth rotates on its axis.
   b) The Sun rotates on its axis.
   c) Earth's axis is tilted.
   d) The Sun's axis is tilted.
3. (8th grade) Which of the following is a true statement about the magnetic field between two magnets?
   a) The south pole of one magnet is attracted to the south pole of the other magnet.
   b) The south pole of one magnet is attracted to the north pole of the other magnet.
   c) The north pole of one magnet is attracted to the north pole of the other magnet.
   d) The south pole of one magnet is attracted to both poles of the other magnet.
4. (8th grade) The diagram shows a bar magnet which is cut into three pieces with a hacksaw



Write an "N" or an "S" in each box on the diagram to show the polarity of each end of the center piece.

Motivation was measured using the Students' Motivation Towards Science Learning Questionnaire (SMTSL) developed by Tuan, Chin, and Shieh (2005). The SMTSL consists of 35 questions on the following six subscales: self-efficacy, active learning strategies, science learning value, performance goal, achievement goal, and learning environment stimulation. The full SMTSL questionnaire has a Cronbach alpha of 0.89, high enough for research use (Nunnally, 1978). Criterion-related validity evidence for this scale was provided by Tuan, Chin and Shieh (2005).

The formative assessment tools used to assess student progress toward standards were developed according to the guidelines recommended by the National Research Council (NRC, 2001). Rubric given to students in this study measured four inquiry based criteria: 1) designing and conducting experiments, 2) evidence development and collecting data, 3) understanding and connecting concepts, and 4) communicating scientific evidence. In addition, rubrics also measured relevant content standards associated with each unit derived from the national science standards and state standards. An exemplar rubric to grade student responses on the knowledge integration standard is given in Table 2.

Table 2. Knowledge integration rubric.

Standard Level	Link Levels	Description
Exceed Standard	Complex Link	Elaborate two or more scientifically valid links among relevant ideas Link structure is complex when multiple full links are presented in conclusion.
Meet Standard	Full Link	Elaborates a scientifically valid link between two relevant ideas Link is scientifically valid. Link is elaborated fully.
Approaching Standard	Partial Link	Elicit relevant ideas but do not fully elaborate the link between the relevant ideas Ideas are connected but missing key element, link is not fully developed.
Basic	No Link	Identifies a concept relevant to the scientific phenomenon involved in task or question, no explanation is given to demonstrate understanding.

Data Analysis

The performance of the three groups were evaluated in the following way. First, to compare the academic achievement of the three groups, ANOVA (Analysis of Variance) analysis was conducted on the summative achievement test score. Post hoc comparsion was conducted to pinpoint group difference when significance was detected from the general test. Second, same ANOVA analysis was performed on science motivation using the SMTSL total score as the depedent variable. In addition, for the two standard-based grading groups, progress on standard development was evaluated by paired t-test on the difference between the number of standard mastered before and after the instruction. Alpha level was set at 0.05 for all the analyses.

Results of Research

First, results on the student achievement measured by standardized test items are reported. Table 3 presents the mean and the standard deviation for the three grading conditions.

Table 3. Standardized achievement assessment: descriptive statistics.

Grade	Group	Mean	Standard Deviation
Grade 8	Total Points	16.29	1.88
	Rubric Only	16.76	2.44
	Rubric+Feedback	18.43	1.53
Grade 7	Total Points	14.79	2.34
	Rubric Only	17.04	3.34
	Rubric+Feedback	18.59	2.87

A one-way ANOVA test shows that student achievement was significantly different for both grades. For the 8^{th} grade, $F(2, 65) = 7.57$, $p < .01$, and for the 7^{th} grade, $F(2, 65) = 10.06$, $p < .01$. Table 4 gives the post hoc groupwise comparison results. The table reveals two important findings. First, the rubric plus feedback group consistently scored higher than both the rubric only and the total point groups. Second, for both grades, the rubric-only group did not show a significant improvement over the total point group.

Table 4. Standardized achievement assessment: comparing groups.

Grade	Pairwise Comparison	Mean Difference	Standard Error	Level of Significance
Grade 8	Rubric Only vs. Total Points	.47	.58	p =.43
	Rubric+Feedback vs. Total Points	2.14	.57	p < .01*
	Rubric+Feedback vs. Rubric Only	1.67	.59	p = .01*
Grade 7	Rubric Only vs. Total Points	1.55	.87	p =.08
	Rubric+Feedback vs. Total Points	3.8	.85	p < .01*
	Rubric+Feedback vs. Rubric Only	2.2	.85	p =.01*

** Significant at .05 level.*

Next, standard development was evaluated for the two standard-based groups. As no standard-based grading was given to the total point group, that group was not included in this analysis. The research question being addressed is how many preset standards were mastered. Table 5 gives the descriptive statistics. First, the gain from pre-test to post-test for two group combined was evaluated by a paired T-tests. For both groups of grade 8, significant gain was detected (for the ruric only group, $t\,(20) = 13.51.08$, $p < .01$; for the feedback group, $t\,(21) = 29.38$, $p < .01$). Significant gain was also achieved for both groups of grade 7 (for the ruric only group, $t\,(21) = 24.08$, $p < .01$; for the feedback group, $t\,(21) = 26.76$, $p < .01$). To explore whether the the rubric only or the feedback group mastered more standards from the pre-assessment to the post-asessment, an independent t-test was conducted on the the change score but found no difference for either grade (for the 8th, $t\,(42) =.77$, $p = .45$ and for the 7th, $t\,(42) = 1.82$, $p = .08$).

Table 5. Standard development: descriptive statistics.

Grade	Group	Mean	SD
Grade 8	Pre-Assessment, Rubric Only	4.05	1.91
	Post-Assessment, Rubric Only	10.48	2.52
	Pre-Assessment, Rubric + Feedback	4.35	1.53
	Post-Assessment, Rubric + Feedback	11.17	1.87
Grade 7	Pre-Assessment, Rubric Only	2.82	1.40
	Post-Assessment, Rubric Only	8.68	2.25
	Pre-Assessment, Rubric + Feedback	3.14	1.17
	Post-Assessment, Rubric + Feedback	9.73	1.91

Finally, results on how different assessment methods affect student motivation to learn science are presented in Tables 6 and 7.

Table 6. Motivation in learning science: descriptive statistics.

Grade	Group	Mean	Standard Deviation
Grade 8	Total Points	117.91	16.68
	Rubric Only	126.38	8.94
	Rubric+Feedback	133.91	9.44

Grade	Group	Mean	Standard Deviation
Grade 7	Total Points	117.54	12.02
	Rubric Only	130.00	12.10
	Rubric+ Feedback	137.09	9.88

The ANOVA analysis indicated that motivation was significantly different among the three groups for both the 7th and 8th grades. For the 8th grade, $F\,(2, 65) = 9.82$, $p < .01$, and for the 7th grade $F\,(2, 65) = 17.21$, $p < .01$. The group comparsion analysis results in Table 7 indicated that all groups were different from each other and the rank order of the motivation level was rubric plus feedback > rubric only > total points.

Table 7. Motivation in learning science: comparing groups.

Grade	Group Comparison	Mean Difference	Standard Error	Level of Significance
Grade 8	Rubric Only vs. Total Points	8.46	3.70	p =.03*
	Rubric+Feedback vs. Total Points	16.00	3.61	p < .01*
	Rubric+Feedback vs. Rubric Only	7.53	3.70	p = .04*
Grade 7	Rubric Only vs. Total Points	12.46	3.39	p < .01*
	Rubric+Feedback vs. Total Points	19.5	3.39	p < .01*
	Rubric+Feedback vs. Rubric Only	7.09	3.46	p = .04*

** Significant at .05 level.*

Discussion

While most science teachers recognize the value of formative written feedback, support from the research community has been very limited. Consequently, many classroom teachers still have doubt about the effectiveness of written feedback. On one hand, giving written feedback is time consuming. On the other hand, not many students actually act on it (Brown & Glover, 2005; Lea & Street, 1998). This study provides strong fresh empirical evidences that written feedback, when used appropriately, can indeed help students learn science better.

The use of standard-based grading, coupled with written feedback, has much more to offer to the students than the sheer number of points, which remains to be the most prevalent form of feedback students receive, albeit all its drawbacks. The combination of standard-based grading and written feedback has potential to overcome many barriers in giving feedback. One such hinderance is the amont of time it may take. Glover and Brown (2006) vividly described the amount of time "tutors" in their study spent as well the disappointmet they had when students did not act on their feedback. Feedback in the current study has three distinctive features. First, it is always standard-based. In other words, the teacher did not comment on all errors in student work. In reality, many errors were ignored, such as spelling or grammatical errors in a science assignment as long as they don't affect understanding. Only those closely related to the interested standards were addressed. Second, the number of written feedback was controlled so that students would not be overwhelemed. As shown in Figure 1, only major misconceptions were pointed out. Finally, a second chance was given. The real power of feedback lies in that students can actually use it to turn an incorrect response into an additional learning opportunity. As pointed out by Willingham (1990), merely reading the suggestions for corrections is not good enough. What really helps is students can use feedback to continue working on problems, which is in perfect agreement with standad-based grading and formative assessment.

This study finds no significant difference on the number of standards mastered by students from the rubric only and rubric plus written feedback groups. However, it was observed in classroom teaching that students in the rubric only condition struggled more on how to improve. They also needed more attempts to reach standards. In contrast, the targeted written feedback informs students directly of what to work on next. This no difference may be also due to the lack of perfect control in research design. Like in many classroom-based studies, highly motivated students from the more controlled condition will seek assistance anyway when they need help. In this study, it was observed that the rubric-only group noticeably asked more questions and sought more help in class. In other words, these students merely used another form of feedback, such as verbal feedback, to get necessary information. In this sense, not providing formative feedback will be most detrimental to students who don't seek help.

This study has a few limitations. First, sample size is not large, which may limit the generalization of its findings. Second, the duration of the experiment was relatively brief due to the fear of exposing students to an inferior method for a lengthy period. While the effect of written feedback was to be examined by the current study, the teacher was still concerned about limiting the learning experience for some students. For the same reason, no point-based group was intentionally formed. Instead, a natural class already practicing point-based grading was used. One side effect of that arrangement is the possible teacher effect as that class was taught by a different teacher. While both teachers were clearly competent in teaching the relevant content, students may have responded to them differently. In the future, this possible teacher effect should be controlled.

While motivation to learn science was measured, the change of motivation was not, which can be a topic for future study. Another related topic is middle school students' attitude towards feedback. Under standard-based grading, students may be less discouraged by errors in their work but more motivated by informative yet manageable feedback and the second chance to exceed standards. The change of attitude towards feedback may be related to both achievement and motivation in science learning.

Conclusions

This study clearly demonstrates that among the three common grading methods currently used in middle school science classrooms, the best practice is standard-based grading plus written feedback. Standard-based grading informs students where they stand on important educational standards, which is hard to deliver by the traditional point-based grading. Meanwhile, written feedback can provide tailored-made suggestions for each student on how to meet or exceed standards. In practice, this study recommends that written feedback be highly relevant (to the academic standard being assessed), limited in number (to the major misconceptions students have), and giving a second chance (for students to act on to improve learning). For future study, it will be interesting to explore how these conclusions and suggestions apply to other populations, such as students in elementary school or high school science classrooms.

References

Bailey, R., & Garner, M. (2010). Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices. *Teaching in Higher Education*, *15* (2), 187- 198.

Black, P., & Wiliam, D. (2004). The formative purpose: assessment must first promote learning in *towards coherence between classroom assessment and accountability* - 103rd yearbook of the National Society for the Study of Education (ed. M. Wilson), pp.20-50 Chicago: University of Chicago Press for the NSSE.

Brown, E., & Glover, C. (2005, September). Refocusing written feedback. In *13th Improving Student Learning Symposium*, Imperial College, London, UK (pp. 5-7).

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology*, *78*, 210-16. doi: 10.1037/0022-0663.78.3.210.

Chin, C. (2006). Classroom interaction in science: teacher questioning and feedback to students' responses. *International Journal of Science Education, 28 (11),* 1315-1346. doi:10.1080/09500690600621100.

Clymer, J. B., & Wiliam, D. (2007). Improving the way we grade science. *Educational Leadership, 64* (4), 36–42.

Deci, E. L., Ryan, R. M., & Koestner, R. (1999). A Meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125* (6), 627-668. doi: 10.1037/0033-2909.125.6.627.

Deckers, L. (2004). *Motivation: Biological, psychological, and environmental.* Boston: Pearson/Allyn & Bacon.

Elawar, M. C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: changing teacher behavior a little rather than a lot. *Journal of Educational Psychology*, *77,* 162-173. doi: 10.1037/0022-0663.77.2.162.

Glover, C., & Brown, E. (2006). Written feedback for students: Too much, too detailed or too incomprehensible to be effective? *Bioscience Education E-journal, 7-3.* doi:10.3108/beej.2006.07000004

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77* (1), 81-112. doi: 10.3102/003465430298487.

Kage, M. (1991). *The effects of evaluation on intrinsic motivation*. Paper presented at the meeting of the Japan Association of Educational Psychology, Joetsu, Japan.

Kast, A., & Connor, K. (1988). Sex and age differences in response to informational and controlling feedback. *Personality and Social Psychology Bulletin*, *14* (3), 514-523. doi: 10.1177/0146167288143010.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in higher education, 23* (2), 157-172. doi:10.1080/03075079812331380364.
Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15* (3), 249-267. doi:10.1207/S15324818AME1503_2.
Mueller, C. M., & Dweck, C. S. (1998). Intelligence praise can undermine motivation and performance. *Journal of Personality and Social Psychology, 75,* 33-52. doi:10.1037/0022-3514.75.1.33.
National Research Council (2001). *Classroom assessment and the National Science Education Standards*. National Committee on Science Education Standards and Assessment. Washington, DC: National Academy Press.
Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
Thomas, J. W., Bol, L., Warkentin, R. W., Wilson, M., Strage, A., & Rohwer, W. D., (1993). Interrelationships among students' study activities, self-concept of academic ability, and achievement as a function of characteristics of high-school biology courses. *Applied Cognitive Psychology*, *7*, 499-532. doi: 10.1002/acp.2350070605.
Tuan, H. L., Chin, C. C., & Shieh, S. H. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education*, *27* (6), 639-654. doi:10.1080/0950069042000323737.
Willingham, D. B. (1990). Effective feedback on written assignments. *Teaching of Psychology*, *17* (1), 10-13. doi: 10.1207/s15328023top1701_2.

Received: *December 15, 2014* Accepted: *March 31, 2015*

Bo Zhang	Ph.D., Associate Professor, Department of Educational Psychology, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA. E-mail: boz@uwm.edu
Jacob Misiak	MS., New Berlin School District, New Berlin, Wisconsin, USA. E-mail: Jacob.Misiak@nbexcellence.org