Illustrations with Graphic Devices in Large-Scale Science Assessments:

An Exploratory Cross-Cultural Study of Students' Interpretations

Chao Wang
University of Colorado, Boulder
Guillermo Solano-Flores
University of Colorado, Boulder

Paper presented at the annual meeting of the
National Council on Measurement in Education, April 7-11, 2011

Abstract

In this exploratory, cross-cultural study, we examined students' interpretations of graphic devise-based illustrations used in science tests. Graphic devices are visual components (e.g., arrows, dotted lines) intended to ensure proper understanding of the scientific processes or phenomena represented by the illustrations. We address cultural differences in terms of the interaction of two factors, students' country of origin and items' country of origin. We hypothesized that interpretations made by students of device-based illustrations are more accurate for items generated in their own country than items generated in another country. Two matched samples of American college students who lived and studied in the U.S. (n=40) and Chinese college students who lived and studied in mainland China (n=40) were given illustrations from eight science items whose illustrations contained different sorts of graphic devices; four of those items were sampled from Chinese large-scale assessments and four from American large-scale assessments. For each illustration, students were asked: (1) to describe what they saw in the illustration, and (2) whether they thought the illustration represented a scientific concept and, if so, to describe which scientific concept was represented. The accuracy of the responses was scored based on scoring rubrics developed for each item. The results indicate that: (1) some illustrations were more difficult to interpret accurately than others, regardless of the students' or the items' country of origin; (2) Chinese students had more accurate interpretations than their American counterparts of the scientific concepts represented by the illustrations; and (3) students' interpretations of the scientific concepts illustrated were more accurate for items generated in the students' own culture than items generated in the other culture. We discuss lessons learned from this exploratory study and future directions for a full study.

Note: This study is part of a larger study titled, "Design and Use of Illustrations in Test Items as a Form of Accommodation for English Language Learners in Science Assessment," funded by the National Science Foundation (Award No. DRL 0822362). We are grateful to the funding agency, our colleagues in the project, and the members of our technical advisory board for their support. The opinions expressed are not necessarily those of our colleagues or the funding agency. We are especially grateful to Professor Xinying Li, Dr. Marilyn Blackmon, and Dr. Derek Briggs, for their generous support and constructive feedback.
Corresponding author: Chao Wang. chao@colorado.edu

Introduction

Despite their prevalent use in science assessments, little is known about the ways in which cultural factors shape students' perception and interpretation of illustrations. According to Trumbo (1999), both the ability to use visual representations to recall concepts and the ability to interpret visual representations are critical to understanding scientific images. In the context of science assessment, it is reasonable to assume that, in addition to creating meaning from the illustrations used in tests, test takers are expected to infer the scientific concepts underlying those illustrations. Boling, Smith, Frick, & Eccarius (2007, p. 4) distinguish between picture perception, which refers to "recognizing what is in a picture" and picture interpretation, which refers to "deciding what a given picture is supposed to mean"—which, in context of science, consists of inferring the scientific concepts represented by a given illustration.

Critical in the design of science illustrations is the role of graphic devices, which can be defined as visual components included with the intent to ensure proper understanding of illustrations (Boling, Eccarius, Smith, & Frick, 2004). Graphic devices consist of arrows, motion lines, speech ballowns,and many other conventional forms of representation. For example, Figure 1 shows a ball attached to the end of a string and spun in a circle by a hand holding the string. The dashed-line circle showing the path of the ball and the arrow showing that the ball is moving counterclockwise are graphic devices that have been added to the image, and which would not be visible to a viewer watching the scene.

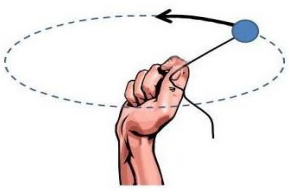

Figure 1. An example of an illustration with two graphic devices. Source: Arizona Department of Education. (2009). Arizona's instrument to measure standards AIMS Science, 2009 sample test for grade 8.

The factors that shape the effectiveness of graphic devices in enhancing access to content have been well documented (Boling, Eccarius, Smith, & Frick, 2004; Boling, Smith, Frick, & Eccarius, 2007; de Souza & Dyson, 2008; Xie, 2009). Especially important is the notion that, in order to properly convey meaning, graphic devices need to be designed bearing in mind that they are conventions—students need to be familiar with them if these visual resources are to support them to gain access to content as intended (Rankin & White, cited by Winn, 1987; Weidenmann, 1994). Findings from cross-cultural studies involving web icons and images used along with academic content suggest that culturally-bound conventions may affect how images are perceived and interpreted, and even lead viewers to interpret the illustrators in ways not intended by their creators (see Boling, Eccarius, Smith, & Frick, 2004; Boling, Smith, Frick, & Eccarius, 2007; Chua, Boland, & Nisbett, 2005; Nisbett & Miyamoto, 2005; Knight, Gunawardena, & Aydin, 2009).

In this paper, we report the results of an exploratory cross-cultural study that examines students' perceptions and interpretations of graphic device-based illustrations used in science tests. The study is part of a series of studies that we are conducting to examine how illustrations can be used to minimize language and culture as sources of measurement error by making the content of science items more accessible to test takers in both the context of English language learners testing (Solano-Flores, 2010, 2011) and the context of international test comparisons (Solano-Flores & Wang, 2011; Wang, 2009).

Unlike conventional cross-cultural studies that examine population differences based on stimulus materials generated in one culture, we address cultural differences in terms of the interaction of two factors, the students' country of origin and the items' country of origin. We examined whether college students from the U.S. and China differed significantly in the ways in

which they interpreted illustrations from large-scale science assessments originated in the U.S. and in China. We hypothesized that students make more accurate interpretations of graphic device-based illustrations generated in their own culture than graphic device-based illustrations generated in another culture.

We regard this study as exploratory because we are limiting our analysis to data from a subset of the total of thirty graphic device-based illustrations used in the investigation and because double-scoring procedures were used only for a subset of these responses. This approach allows us to refine our scoring procedures before we double-score the full set of student responses. Accordingly, we discuss interrater reliability as it relates to interpreting the responses from students from different cultural groups. This approach also allows us to become acquainted with the kinds of results that we should expect to obtain and the kind of evidence we should focus on in the full study. As shown below, despite its exploratory nature, the results from this study have important implications for test development.

For the purpose of our study, we use, *culture* and *country* interchangeably, although we recognize that these concepts are quite different (though related).

Methods

Participants

In this study participated 80 college students; 40 were from a research university in the U.S. Mid West and 40 were from a university in Beijing, China. As Table 1 shows, these two groups of students were comparable in terms of gender proportions and age range; the majority were in non-science majors.

Table 1
Demographic Information of the Sample of Participants.

		Gender		Age (years)		Major	
Country	Number	Male	Female	Mean	Range	Science	Non-science
U.S.	40	17	23	18.78	18-22	9	31
China	40	11	29	19.33	18-22	2	38

Instrument and Administration Procedure

The original instrument includes 30 device-based illustrations accompanying science test items. Of these 30 device-based illustrations, 14 were selected from 8th grade large-scale science assessments from the U.S. and 16 were selected from 9th grade large-scale science assessments from China. (Unlike the U.S., middle school large-scale assessments for the corresponding age range in China are administered in 9th grade, not 8th grade.) Both samples of items consisted predominantly of Physical Science.

Qualtrics, an online survey program, was utilized to administer the instrument on line. The computer projected on the screen the illustrations one by one *without* the text of the items. For each illustration, the computer asked participants to type, below the illustration, their responses to two questions intended to gauge their interpretations of the illustrations. Question 1, Image Perception, addressed "recognizing what is in a picture" (see Boling et al., 2007); it asked students to describe what they saw in the illustration. Question 2, Scientific Concept Interpretation, addressed "deciding what a given picture is supposed to mean" (see Boling et al., 2007); it asked students whether they thought the illustration represented any scientific concepts and, if so, to describe the most important scientific concept.

The instrument was administered individually in the native language of the participants at locations in the participants' universities in the U.S. and in China. On average, students took about 50 minutes to complete this online survey.

In order to balance for the effect of sequence of administration, the 30 items were assigned to two groups of 15 items each; then the students were randomly assigned to one of two sequences. Half of the students were given Illustrations 1-15 first, then Illustrations 16-30; the other half of the students were given the illustrations in the reverse sequence.

In this exploratory study, we used eight of the 30 graphic device-based illustrations; four from tests from the U.S. and four from tests from China. All of the eight items selected were Physical Science items.

Scoring

For each graphic device-based illustration, we developed a set of scoring rubrics intended to measure the accuracy with which students interpreted the illustrations. These rubrics were developed with the participation of science experts in the content areas of the items. The rubrics specified four levels of accuracy on a four-point (1-4) scale and a five-point (0-4) scale respectively for Questions 1 and 2, in which 4 corresponded to the highest level of accuracy. The zero in the five point scale for Question 2 was given when a student responded that the illustration did not represent any scientific concept.

The scoring rubrics were made available to the scorers in both English and Chinese. They were supplemented with example responses (not used in live scoring) for each scale point.

The responses were scored by bilingual (English and Chinese) individuals who were native Chinese speakers. One of the scorers was a doctoral student in a program on language and culture in education; the other was a doctoral student in a science program. There is only scant

research on the effect of bilingual raters' native language background on the reliability of their scoring of student responses in two languages. However, we have evidence that, provided that bilingual raters have formal professional training on language issues (e.g., they are certified as bilingual educators), bilingual raters can reliably score student responses in two languages, regardless of which of those languages is the bilingual raters' native language (Prosser & Solano-Flores, under review). In addition, there is evidence that, as long as rigorous coding procedures are used, similar reliability coefficients can be obtained for diverse cultural groups (Solano-Flores & Li, 2009).

Data Analyses

We performed a series of two-way ANOVAs to assess the main and interaction effect of student's country of origin (U.S. and China) and item's country of origin (U.S. and China) on: 1) the accuracy of the descriptions of the illustrations and (2) the accuracy of the interpretations of the scientific concepts represented by the illustrations. In our analyses, we assumed independence of both the graphic device-based illustrations and Questions 1 and 2 for each of the illustrations.

Results

Interrater reliability. For the purposes of this exploratory study, to examine interrater reliability, we examined the rank ordering of the scores given by two independent scorers to the students' responses to two illustrations, one from an item generated in the U.S. (Item A) and the other from an item generated in China (Item B).

As Tables 2 and 3 show, the correlations for Questions 1 are moderately high while the correlations for Question 2 are high. Clearly, Question 1 poses more scoring challenges than Question 2. Because the patterns of magnitude of the correlations are consistent across both

students' and item's country of origin, these scoring challenges appear to be related to the open, constructed nature of the responses scored rather than difficulty in interpreting responses from individuals from different cultures. These challenges do not appear to be insurmountable.

Table 2
Inter-rater Reliability (Pearson correlation) for Image Perception, by Student's and Item's Country of Origin: Items A and B.

Students' country of origin	Item's country of origin	
	Item A: U.S.	Item B: China
U.S. (n=40)	0.65	0.79
China (n=40)	0.78	0.75
Total (n=80)	0.76	0.78

Table 3
Inter-rater Reliability (Pearson correlation) for Interpretation of the Scientific Concept Represented, by Student's and Item's Country of Origin: Items A and B.

Students' country of origin	Item's country of origin	
	Item A: U.S.	Item B: China
U.S. (n=40)	0.97	0.93
China (n=40)	0.92	0.97
Total (n=80)	0.94	0.98

Image perception. Figure 2 shows the mean Image Perception scores obtained by the students from each group. A two-way ANOVA revealed statistically significant mean score differences due to the students' country of origin (p=.008) and the items' country of origin (p=.03), not their interaction. Only the effect size due to illustrations was large (η^2= 0.157), which indicates that some illustrations tended to be more difficult to understand than others, regardless of the items' or the students' country of origin. Needless to say, given the moderate interrater reliability obtained with the sample of two items double-scored (see Table 2), these results need to be interpreted with caution.

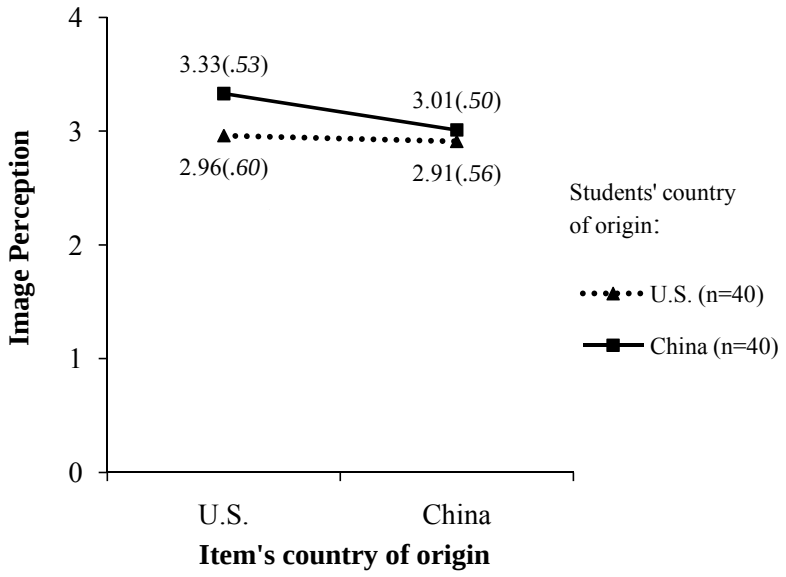

Figure 2. Mean Scores on Image Perception Across Country of Origin. Standard Deviations in Parentheses.

Scientific concept interpretation. Figure 3 shows the mean Scientific Concept Interpretation scores obtained by the students from each country. A two-way ANOVA revealed statistically significant mean score differences due to the interaction of students' country of origin and items' country of origin (p=.017) and to the main effect of student's country of origin (p=.000). Small effect sizes due to the interaction of specific illustrations and the students' country of origin (η^2=0.074) and specific illustrations (η^2=0.101) were observed. In contrast, a large size effect size (η^2=0.215) due to the students' country of origin was observed. This finding indicates two facts. First, students tended to have more accurate interpretations of the scientific concepts illustrated when the items were generated in their own culture than when the items were generated in another culture. Second, Chinese students' are more familiar than students from the

U.S. with diverse graphic representations of the same scientific concepts, regardless of the country of origin of the illustrations.

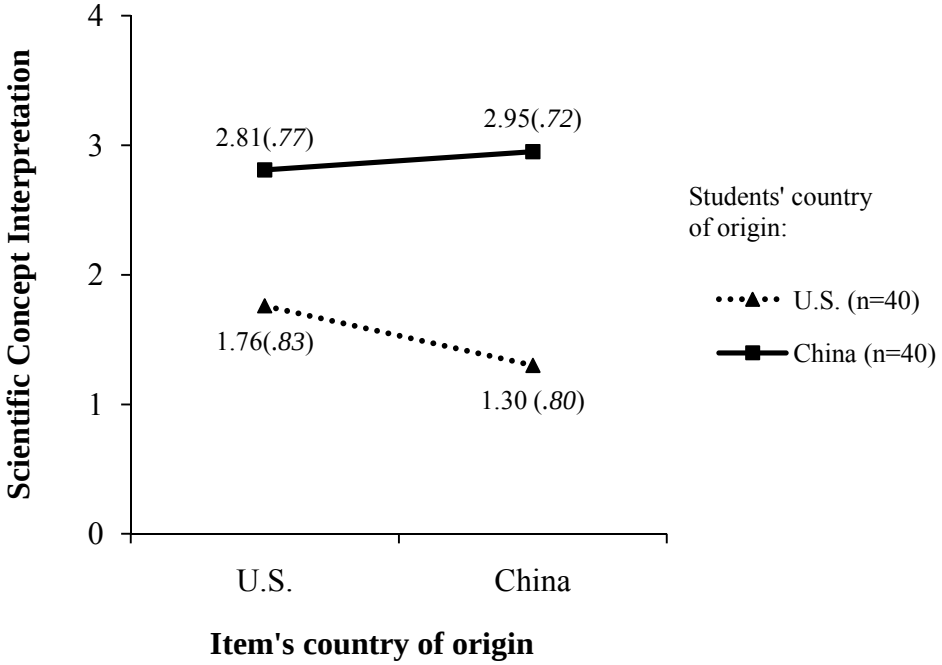

Figure 3. Mean Scores on Science Concept Across Country of Origin. Standard Deviations in Parentheses.

Summary and Conclusions

In this exploratory study, we examined students' interpretations of graphic devise-based illustrations used in science tests. We hypothesized that interpretations made by students of device-based illustrations are more accurate for items generated in their own country than items generated in another country.

In support to our hypothesis, we found that students from both China and the U.S. tended to have more accurate interpretations of the scientific concepts represented by illustrations generated in their own culture than illustrations generated in the other culture. However, this

relationship appears to be mediated by the characteristics of the illustrations and the familiarity students may have with visual representations of scientific concepts with graphic devices (e.g., through textbooks and tests)—which is reflected by the fact that Chinese students were more accurate than American students in their interpretations of the scientific concepts represented by the illustrations.

The results have important implications for improved test design in projects involving the testing of culturally and linguistically diverse populations. Against the implicit assumption that illustrations can be understood in the same ways by students from different cultures, our study shows that visual images generated in one culture are not interpreted with the same level of accuracy by students from different cultures.

Our study also shows that testing illustrations are not necessarily interpreted in the ways intended by test developers. Using cognitive interview procedures with samples of students that reflect the cultural and linguistic diversity of the populations tested should allow test developers to probe whether illustrations included in tests truly and equally provide the visual support intended.

Finally, our study contributes to the field of research involving cross-cultural testing with a more comprehensive approach to examining cultural differences. Unlike common cross-cultural research that focuses on cultural group as a factor, we address the interaction of cultural group and the culture of origin of the stimulus materials. We believe this kind of design makes it possible a more thorough examination of cultural issues in testing.

In sum, despite this is an exploratory study, the evidence obtained speaks to the fact that test developers must be extremely cautious in their assumptions about the properties of illustrations in science testing. While it may be true that a picture is worth a thousand words, the ways in

which students make sense of it depends on both the students' culture and the culture in which it originates.

References

Arizona Department of Education. (2009). Arizona's instrument to measure standards AIMS Science, 2009 sample test for grade 8. Retrieved October 28, 2009, from http://www.ade.state.az.us/standards/aims/sampletests/Gr8-AIMSsampletestscience.pdf

Boling, E., Eccarius, M., Smith, K., & Frick, T. (2004). Instructional illustrations: Intended meanings and learner interpretations. *Journal of Visual Literacy*, *24*(2), 185–204.

Boling, E., Smith, K., Frick, T. & Eccarius, M. (2007). *Graphic devices in instructional illustrations: Designers' intentions and viewers' interpretations*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America, 102(*35), 12629–12633

de Souza, J.M.B., & Dyson, M. (2008). Are animated demonstrations the clearest and most comfortable way to communicate on-screen instructions? *Information Design Journal*, *16*(2), 107–124.

Knight, E., Gunawardena, C. N., Aydin, C. H. (2009). Cultural interpretations of the visual meaning of icons and images used in North American web design. *Educational Media International, 46* (1), 17-35.

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: holistic versus analytic perception. *Trends in Cognitive Sciences, 9*(10), 467-473.

Prosser, R., & Solano-Flores, G. (Under review). Rater language background as a source of measurement error in the testing of English language learners. Manuscript submitted for publication.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28* (2), 9-18.

Solano-Flores, G. (2010). The use of pictorial supports as an accommodation for increasing access to test items for students with limited proficiency in the language of testing. Paper presented at the 7th Conference of the International Test Commission. Hong Kong, July 19-21, 2010.

Solano-Flores, G. (2011). *Development of illustrations as image supports for English language learners in large-scale testing: A report on the procedure for designing vignette illustrations*. Paper to be presented at the CADRE ELL-STEM Roundtable session, "Advancing English Language Learners in Science and Math: Realizing the Promise," at the annual meeting of the American Educational Research Association, April 7-11.

Solano-Flores, G., & Wang, C. (2011), *Development and use of a conceptual framework for analyzing and classifying illustrations used in science assessment.* Paper proposal submitted to the Annual Conference of the American Educational Research Association, New Orleans, LA, April 2011.

Trumbo, J. (1999). Visual literacy and science communication. *Science Communication, 20*(4), 409-425.

Wang, C. (2009). *Illustrations with graphical devices in large-scale science assessments: Cross-cultural perception and interpretation*. Unpublished manuscript, University of Colorado at Boulder.

Weidenmann, B. (1994). Codes of instructional pictures. In W. Schnotz & W. Kulhavy (Eds.), *Comprehension of Graphics*. Amsterdam: North-Holland, pp.29-42.

Winn, B. (1987). Charts, graphs, and diagrams in educational materials. In D. Willows & H. Houghton (Eds.), *The psychology of illustration: Volume 1, Basic research* (pp. 152-198). New York: Springer-Verlag.

Xie, J., & Bonn-Rhein-Sieg, H. (2009). Static visualization of dynamic processes. Retrieved September 16, 2009, from http://www2.inf.h-brs.de/mi/lv/smibi/ss09/ausarbeitung/pub/aa-xie.pdf