The Geological Society of America
Special Paper 474
2011

Designing a mixed-methods research instrument and scoring rubric to investigate individuals' conceptions of plate tectonics

Scott K. Clark*
Julie C. Libarkin
Department of Geological Sciences and Center for Research on College Science Teaching and Learning, Michigan State University, 206 Natural Science Building, East Lansing, Michigan 48824, USA

ABSTRACT

Research methods and underlying theories for research designs that integrate quantitative and qualitative approaches (i.e., mixed methods) are well documented in the field of education research. What is missing in the literature is a nuts-and-bolts description of the actual practice that goes into creating a good mixed-methods survey instrument for research in the science education domain. This paper will detail the steps involved in designing, implementing, and scoring a valid and reliable mixed-methods survey instrument. This survey instrument was designed to investigate experts' and novices' conceptual understanding of plate tectonics as inferred by their answers to a series of questions related to a modified version of a commonly used cross-section schematic published by the U.S. Geological Survey. Development of the instrument involved numerous revisions with iterative inputs from local and community-based experts. After integration of expert comments, the survey instrument was piloted to a physical science for nonscience majors course. This led to further revisions in the survey instrument to improve communication validity prior to widespread distribution. Development of scoring rubrics similarly required iterative modifications based on a thematic analysis of collected data. By outlining the steps involved in designing, validating, and analyzing this mixed-methods instrument, we believe that this paper can serve as a template for future survey instrument development. In particular, we hope to illustrate the iterative and time intensive nature of mixed-methods inquiry, both in terms of pre-investigation design and postinvestigation analysis, and to offer our empirically based insights into the instrument and rubric development process.

INTRODUCTION

The investigation of alternative conceptions held by students is a domain of research that has driven science education discourse for nearly a century (e.g., Driver, 1989; Duschl et al., 2007; Posner et al., 1982, and references therein). The approaches used to reveal student ideas range from multiple-choice survey instruments with constrained response options, to a researcher passively observing discussions in a classroom, to intensive, one-on-one interviews, to broad, open-ended survey

*Current address: Department of Geology, University of Wisconsin–Eau Claire, Phillips 157, Eau Claire, Wisconsin 54702-4004, USA; clarksco@uwec.edu.

Clark, S.K., and Libarkin, J.C., 2011, Designing a mixed-methods research instrument and scoring rubric to investigate individuals' conceptions of plate tectonics, *in* Feig, A.D., and Stokes, A., eds., Qualitative Inquiry in Geoscience Education Research: Geological Society of America Special Paper 474, p. XXX–XXX, doi:10.1130/2011.2474(07).

instruments designed to illicit freeform thought. Each of these approaches provides valuable information about the range and depth of student thinking, generating tangible evidence for the missteps students can take on the pathway to scientific literacy. While interviews provide opportunities to gain a detailed understanding of thinking and reasoning for a small handful of students, multiple-choice survey instruments offer an opportunity to evaluate the prevalence of ideas across many students. Among the possible research techniques, open-ended survey instruments are highly valued as tools that offer opportunities to both collect data from many students and probe, however lightly, for explanations of ideas.

Survey instruments of all types are attractive to researchers because of their: (1) perceived relative ease of development; (2) ability to acquire data from multiple population samples; and (3) possibility for simple content and semiquantitative analyses. Newly available mechanisms for web-based dissemination (e.g., www.surveymonkey.com) provide ready access to wide and diverse populations. In their broadest sense, surveys can be quantitative, such as multiple-choice concept inventories (e.g., Libarkin, 2008), semiquantitative, as in instruments that utilize a Likert scale (e.g., Adams et al., 2006), qualitative, such as open-ended questionnaires (e.g., Lederman et al., 2002), or open-ended surveys that combine components of both quantitative and qualitative methodologies, i.e., mixed-methods surveys (Creswell, 2003; Hossler and Vesper, 1993).

Quantitative surveys are particularly useful in science education research for large-scale assessments in comparison studies both nationally and internationally (e.g., Britton and Schneider, 2007). Qualitative surveys utilizing an open-ended question design are also well used. While many texts on qualitative research provide general guidelines for instrument development and analysis (e.g., Crotty, 1998; Denzin and Lincoln, 1998; Lincoln and Guba, 1985), very few works discussing the actual experience of developing a survey instrument have been published. Some very well-known qualitative surveys are supported by a literature that describes their conception, development, and use; one of the best examples is the Views of Nature of Science Questionnaire (VNOS; Lederman et al., 2002). Although the body of work documenting the development of the VNOS provides some insight into the actual process of instrument development, the actual, nuts-and-bolts process through which the VNOS was written, reviewed, revised, and piloted is never completely discussed. While tools such as the VNOS clearly required significant thought and effort for their production, the true time-consuming nature of survey instrument development and analysis is only suggested.

Mixed-methods research "combines quantitative and qualitative research techniques, methods, approaches, concepts, or language into a single study" (Johnson and Onwuegbuzie, 2004, p. 17). The mixed-methods approach has, at times, been shunned by both quantitative and qualitative research purists (see Johnson and Onwuegbuzie, 2004; Rossman and Wilson, 1985 for discussions), but has been accepted by many researchers who see the integration of qualitative and quantitative approaches as useful, (Greene et al., 1989; Johnson and Onwuegbuzie, 2004; Kidder and Fine, 1987; Rossman and Wilson, 1985; Tashakkori and Teddlie, 1998, 2003). A significant advantage to the mixed-methods approach is the ability to triangulate or corroborate findings obtained using both qualitative and quantitative techniques (Greene et al., 1989). Mixed-methods surveys are an attractive research method, but the time-intensiveness of survey instrument use in science education research needs to be explicit because scholars new to their use may be surprised as they engage in the process of survey development. This paper will detail the steps involved in designing, implementing, and scoring a valid and reliable mixed-methods survey instrument.

SURVEY INSTRUMENT AND RUBRIC DEVELOPMENT

As with any scientific endeavor, locating one's research includes a discussion of the research question, rationale for conducting the study, and a discussion of why a particular research approach (qualitative, quantitative, or mixed-methods) was chosen. However, unlike typical scientific research, one must also locate the researcher within the context of the research (e.g., Feig, this volume). This contextualization is grounded in the idea that data interpretation will be affected by the interpreter's incoming perspective (Maxwell, 2005; Patton, 2002). For example, researchers must ask: What is the researcher's position relative to the participants (Marshall and Rossman, 2006)? In a classroom setting, is the researcher also the instructor or an outside observer (Patton, 2002)? What is the researcher's perspective; that is, does the researcher view the data through the lens of a post-positivist, an interpretivist, or a naturalist (e.g., Crotty, 1998; Lincoln and Guba, 1985; Phillips and Burbules, 2000)? Addressing these questions provides insight to both the researcher and users of the research about study quality and potential limitations.

Survey Instrument Design

The overall design of a survey, as well as the design of individual questions, can have significant impacts on the quality of research (Creswell, 2003). A survey instrument that is designed without forethought of intention, or without considering the perspective of the target population, will likely yield results that are at odds with researcher expectations. Appropriate use of language and visuals, attention to page layout (e.g., Sanchez, 1992), and the limiting of distracting elements (Harp and Mayer, 1998) can all improve survey results. While the question of survey design has been discussed most extensively within the sociology or public opinion literature (e.g., Presser et al., 2004, and similar), all fields that utilize surveys in research practice adhere to similar approaches in design.

Regardless of domain, survey design follows a number of reasonable tenets (e.g., Siragusa and Dixon, 2006); these principles apply whether a survey is qualitative, mixed-methods, or

quantitative in structure. These principles have been laid out in any number of good works on survey and analysis design (e.g., Creswell, 2003; Fink, 2003; Thomas, 2004) or general research methods (Trochim and Donnelly, 2007). For our purpose of creating a survey instrument that contains open-ended, guided open-ended, and fixed-response questions, three principles are most important. First, questions need to be understandable to the target population. Efforts should be made to avoid language that is outside the common knowledge of the study population, to word questions as unambiguously as possible, and to use visuals in nondistracting ways. Ultimately, we need to ensure that users are interpreting questions as intended by the developers (Lopez, 1996). Second, pilot testing and subsequent revisions should precede dissemination of the survey instrument in a larger study; this piloting should occur with both experts and a small sample of the targeted population (e.g., Presser et al., 2004). Pilot results may inform the overall study, but cannot, in and of themselves, constitute a study. Finally, the use of cognitive interviews (i.e., think-alouds and probing) to inform survey design serves to validate the researcher's postulated inferences about test-taker intent and thinking (Collins, 2003; Beatty and Willis, 2007; Presser et al., 2004). Similarly, the significant effort invested into the application of these principles to survey design needs to be duplicated during development of rubrics for survey analysis (Ambrose et al., 2004; Bresciani et al., 2009).

Rubric Design

Rubrics for scoring or analyzing qualitative survey data can be used to categorize survey responses or to rank order responses along a relevant continuum (e.g., least to most scientific). In many ways, the development of scoring rubrics mirrors the development of survey instruments, with well-established mechanisms for ensuring that scoring is as unbiased and based in reality as possible. Thematic content analysis, a form of constant comparative analysis, is a common approach used in developing scoring rubrics for qualitative data. During thematic content analysis, the researcher uncovers common themes within the data through an inductive analysis of the data itself (Denzin and Lincoln, 1998; Patton, 2002). Thematic codes are continually changing as data analysis proceeds, although most questionnaire studies in science education are simple enough for major themes to emerge very early in the analytical process. Ultimately, the most important aspect of rubric design is attention to researcher bias; a rubric must reveal, as closely as possible, the perspective of the research subjects rather than the biases of the researchers themselves (Denzin and Lincoln, 1998).

Validity, Reliability, and Trustworthiness

As with any research study, assessments of the research design, data collection, and analytical methods are important in determining research quality. We may create a survey instrument that is intended to measure a specific phenomenon, but in reality may inadvertently measure something different, fail to measure anything meaningful, or may bias results toward our intended outcomes. The challenge of designing a good mixed-methods research project is to ensure that the fallibility of the data collection and analyses is limited to the extent possible. Just as mixed-methods research incorporates aspects of qualitative and quantitative research methods during design, implementation, and analysis, multiple approaches should be used in assessing the quality of mixed-methods research.

Qualitative research is commonly evaluated based on the concept of trustworthiness (Lincoln and Guba, 1985). In quantitative research, evaluation considers the rigor of the study (i.e., validity and reliability; Litwin, 1995; Morse et al., 2002). Similarly, Johnson and Onwuegbuzie (2004) and Onwuegbuzie and Johnson (2006) have presented an approach to evaluating mixed-methods research that they term legitimization. Each of these evaluation frameworks provides a means for judging the quality of research. Utilizing multiple evaluation frameworks provides flexibility in assessing those attributes of the instrument that are pertinent for the specific goals of a given research project (see Morse et al., 2002; Lewis, 2009; Trochim and Donnelly, 2007). We have chosen to assess the quality of this project using a combination of rigor and trustworthiness. We chose this approach over that of legitimization because (1) we are familiar with qualitative and quantitative approaches to trustworthiness and rigor; and (2) rigor and trustworthiness are well established within the science education community, while the legitimization approach is relatively new and unused. A blend of components of trustworthiness with specific metrics for validity and reliability seems to us to be a reasonable approach when evaluating a mixed-methods study. Indeed, a number of researchers have recently argued for various ways to apply validity and reliability to qualitative research projects (Creswell and Miller, 2000; Golafshani, 2003; Lewis, 2009; Morse et al., 2002).

Validity generally refers to how well a measurement represents the true value of the trait being measured (e.g., Trochim and Donnelly, 2007). For example, we might consider how well a test score represents the level of understanding of an individual student being tested. In the case of a conceptual rubric designed as the filter for analyzing survey data, we need to ensure that categories of qualitative data represent, as closely as possible, the underlying conceptions of the study population. Reliability, on the other hand, is concerned with the reproducibility or repeatability of a measure or study (e.g., Trochim and Donnelly, 2007). Although very difficult to actually test, a reliable measure would generate identical test scores if taken repeatedly by a single person, and assuming no change in understanding across test implementation. For qualitative questions in surveys, we can similarly ask if different researchers looking at a single data set would reach similar conclusions, a process referred to as peer review in qualitative research (Merriam, 2002). Similar to validity and reliability, trustworthiness is the application of the concept of rigor in ways that are tailored to the qualitative research setting (Lincoln and Guba, 1985). In particular, trustworthiness

considers the relationships between the researcher, the population under study, and the ways in which data are analyzed. Most importantly, sources of bias, agreement of the participants with the findings, and application of findings or the research process in other settings, for example, need to be considered (Lincoln and Guba, 1985).

We think it is useful here to provide a brief background of those forms of rigor and trustworthiness that are most important for mixed-methods instrument design and analysis (Table 1). Table 1 is not intended to be exhaustive, but rather to touch on those areas of validity and reliability that should be considered when designing a survey and scoring rubric, and that we attempted to address in the design of the survey instrument as discussed here. Finally, the forms of validity and reliability documented herein represent both standard measures and measures that we feel should be considered more routinely in instrument development and analysis.

DEVELOPMENT OF A SURVEY INSTRUMENT AND SCORING RUBRIC FOR ASSESSING CONCEPTS RELATED TO PLATE TECTONICS

We will now describe the steps we took to design a plate-tectonic conceptions survey instrument and the associated rubrics used to analyze collected data. We include details of our iterative approach, and provide a discussion of our insights and reflections on the entire process. Our survey instrument was designed with three research objectives in mind: (1) investigating people's conceptions (both scientific and alternative) of plate tectonics; (2) documenting how these conceptions might vary across the expert-to-novice continuum; and (3) investigating the role of images in communicating, and possibly miscommunicating, plate-tectonic concepts. For this study, novices are considered to be individuals with only an introductory exposure to the theory of plate tectonics, whereas geoscience faculty are considered to be experts. Other participants, such as geoscience graduate students, are positioned at intermediate levels along the expert-novice continuum. The survey instrument we created consists of questions about aspects and terminology related to plate-tectonic processes (Fig. 1). Some of these questions required respondents to view a schematic plate-tectonic cross section. Respondents were also asked to report their confidence in their answers as a measure of the role of an individual's perceived ability on performance (Bandura, 1984). In addition to broad utility, we wanted an instrument that could be widely distributed and serve as the basis for semi-structured, one-on-one interviews. We feel the resultant survey instrument has met our expectations: Novices are able to describe plate-tectonic concepts presented in the survey instrument, an, the image has enough layered knowledge—especially when used in interviews—to probe the deeper conceptual understandings of both novices and experts. The time required for the iterative development of the survey instrument to move from initial conception in early October 2007 to its current form, which was attained in April 2009, was 1.5 yr.

Locating the Research

The context, including setting, in which data are collected can influence study findings (Feig, this volume). For surveys that were administered to college-level students enrolled in introductory-level earth science courses (i.e., novices), the lead author distributed all surveys with the exception of surveys administered to students at a community college in the NE United States, where the course instructor distributed the surveys. For those courses in which the lead author administered the survey instruments, he had no other connection to the students. Surveys completed at an exhibitor booth at the 2008 Geological Society of America (GSA) Annual Meeting were distributed by both authors and by colleagues. Interviews were completed in private rooms at the GSA meeting and at four institutions of higher education. All interviews were conducted by the first author, who had no relationship to research participants. All participants were given a consent form and instructions that had received approval from an Institutional Review Board.

The location of the researcher within the context of the research is possibly more important than the setting for data collection (Feig, this volume; Marshall and Rossman, 2006; Maxwell, 2005; Patton, 2002). The lead author is a geoscientist with a research background in isotope geochemistry and geocognition, which is the study of how people perceive and understand Earth and Earth processes. The second author is also a geoscientist with a research background in geodynamics and geocognition. Both authors have a postpositivist perspective, meaning we perceive knowledge not as a fixed entity, but rather as being supported by the strongest warrants, or grounds, currently available, and subject to change as new evidence becomes available (Phillips and Burbules, 2000).

Instrument Design

In designing the survey, we initially sketched a cross-section image to be developed into a colored image, but then abandoned this approach in favor of modifying a preexisting image that was commonly used in entry-level geoscience instruction. We chose to modify an existing, open-access image instead of designing an original image because we assumed that experts (i.e., geoscience faculty) would accept this modified image as a reasonable model for plate tectonics, and we would then be able to then investigate the extent to which novices perceive the image relative to how experts perceive the image. This assumption was based on the nearly ubiquitous use of the image we chose. The image that we modified is in the public domain (http://commons.wikimedia.org/wiki/File:Tectonic_plate_boundaries.png) and is part of a wall map titled *This Dynamic Planet* (Simkin et al., 1994). The wall map, which was first published in 1989 (Simkin et al., 1989), is the best-selling map in the history of the U.S. Geological Survey (USGS Education webpage).We simplified the image using the drawing software Canvas v. 9.0.4 (ACD Systems). In modifying the image for our purposes, we removed all text, the continental

TABLE 1. VALIDITY AND RELIABILITY CRITERIA IMPORTANT FOR MIXED-METHODS SURVEY AND RUBRIC DESIGN

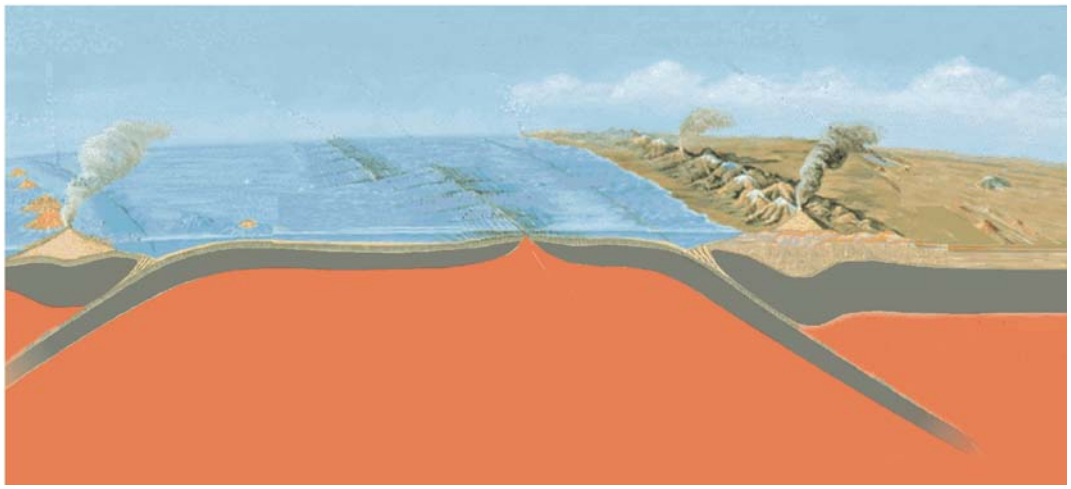
Criteria	Description and approaches	Plate-tectonics survey instrument
Content validity	A measure of whether or not items actually measure the latent trait that they are intended to measure. This is often evaluated through expert review of items and revision in response to expert opinion. Note: *Face validity* is a similar, but more casual assessment of instrument validity; we did not measure face validity per se.	Comments from five geoscientists from the Geocognition Research Laboratory and Geoeducation Research Interest Group listserv and two science educators from the Center for Research on College Science Teaching and Learning group on the pilot version of the survey instrument led to revisions. Analysis of novice responses in pilot data collection resulted in as many changes to the instrument as did expert feedback.
Conclusion validity, internal validity, credibility (see Lewis, 2009)	Conclusion validity is the measure of one's ability to determine the relationship, or lack thereof, between the variables being studied. This is a more general form of internal validity, which is most often considered when an attempt is made to determine a causal relationship between variables. In general, a researcher needs to ensure that they are not biasing study findings through personal expectations, their own actions, or failure to consider study limitations. For qualitative work, *credibility* also addresses researcher bias, and in particular the degree to which study participants agree with findings and the broader implications of the work.	We found this to be the most difficult metric of rigor and trustworthiness to evaluate. Experts exposed to our research findings during presentations at professional meetings generally agreed with the study findings and the implications for image redesign. Ultimately, we view credibility as the final step in the study validation process. As results become available for publication, we anticipate contacting interviewed experts to gauge their agreement with our general findings. In ongoing work, we are also investigating relationships, both causal and noncausal, among gender, confidence, and conceptual understanding.
Construct validity	A measure of whether or not strong support for the content of items exists. This can be estimated through both convergence and divergence of theory and reality. We expect concepts that should be related, such as expertise in plate tectonics and overall understanding of plate tectonics, to actually relate when measured by the instrument and scoring rubric. Similarly, concepts that need not be related, such as plate-tectonics understanding and attitude toward laboratory work, should not show significant correlation.	In general, participants with more expertise in geoscience received better scores on the survey instrument and provided more detailed responses. Interestingly, some misconceptions are retained until extreme levels of expertise are reached (Clark, 2009).
Criterion validity	The degree to which a measure correlates with other measures of the same latent trait (also called "concurrent" validity). Generally, qualitative measures are used to establish criterion validity for quantitative instruments, although quantitative or alternative qualitative measures (i.e., interviews) can be used to validate survey instruments.	Interviews with 61 subjects spanning the expert-novice continuum provided detailed confirmation of both the prevalence of ideas across multiple populations and our interpretations of survey results. For example, novice responses to the question of "How many tectonic plates are in the image?" are in strong agreement with novice responses from other instruments (Kortz et al., this volume).
Communication validity	Researchers develop surveys in order to generate an understanding of a study population. While researchers often assume that participants will interpret questions as intended, explicitly considering this aspect of instrument validity can generate important insights (e.g., Lopez, 1996).	Analyses of the survey instrument were enriched through comparison of researcher intentions with participant interpretations as recorded in think-alouds. The first 10 interviewees, undergraduate majors through experts, completed the survey instrument at the beginning of the interview; upon completion, they discussed their work and responded to interviewer probes about their thinking. Overall, we found that the geoscience major through expert population interpreted the survey instrument as we had intended. Communication validity for novices (nonmajors) was addressed in questions 1 and 2 as we modified the wording until nearly everyone who answered the questions was providing meaningful responses.
Cultural validity	A measure of the extent to which culture impacts participant interpretation of survey questions (Solano-Flores and Nelson-Barber, 2001). We consider this important in any effort to adopt or adapt established tools for new populations.	We do not know how culturally valid the survey instrument will be for subjects outside of the specific study population described here. Certainly, the survey instrument appears to be valid for undergraduates, graduate students and faculty affiliated with U.S. community colleges, and four-year institutions in the northeastern United States. Researchers interested in applying the survey instrument to other populations should consider whether or not cultural differences will require modification of the instrument.
Transferability	A measure of the extent to which results can be generalized to populations outside of the study. This validation is difficult to achieve, although the power of survey research lies in its ability to sample many populations, and hence generate measures of external validity.	Survey instruments were collected from 353 subjects (novices) enrolled in 5 different courses at two institutions (in Michigan and Rhode Island) and from 180 intermediate to expert subjects from an unknown number of institutions who were attendees at the 2008 GSA Annual Meeting. Interviews were conducted with 60 individuals across the expert-novice continuum from a range of universities and nations.

(*Continued*)

TABLE 1. VALIDITY AND RELIABILITY CRITERIA IMPORTANT FOR MIXED-METHODS SURVEY AND RUBRIC DESIGN (*Continued)*

Criteria	Description and approaches	Plate-tectonics survey instrument
Dependability	A measure of the extent to which other researchers would be able to replicate the study findings.	This manuscript is itself an audit trail of the survey instrument and rubric development, and it provides enough information for others to both evaluate the instrument's design and our analytical findings.
Internal consistency reliability	Although most often considered for quantitative instruments, internal consistency can provide a sense of the reliability of a mixed-methods survey. The stability of test results across samples of similar populations, consistency in test results over time, and generation of similar results using slightly different forms all provide evidence that a survey is generating reproducible findings.	Results from the piloted version through to the current version, separated by 14 mo, were similar, overall. Different forms (e.g., one-color versus two-color asthenopshere) produced the same range of responses outside of specific differences. As data analysis progresses, we will compare survey results from different universities; we would expect results to be consistent across populations once demographic or educational backgrounds are accounted for.
Inter-rater reliability	In qualitative design, inter-rater reliability can ensure that findings are reproducible. Often, this is established through an iterative process whereby multiple researchers code identical data and establish consistency in analytical results.	For the survey instrument, we utilized the inter-rater technique multiple times. Inter-rater reliability came into play at a number of analysis stages. Ultimately, we achieved 100% agreement in coding between two researchers; see text for details.

Notes: Except where noted, concepts of validity, reliability, and trustworthiness were adapted from Lincoln and Guba (1985), Litwin (1995), and Trochim and Donnelly (2007).



For each response, please mark the location on the scale that corresponds to your level of confidence

ON THE FIGURE ABOVE, PLEASE:

not confident---------- very confident
at all

1) Label all features related to plate tectonics

2) Show where you think melting could be occurring

3) Indicate relative direction plates are moving

4) What do the colors below the surface represent?

Figure 1. Original version of survey instrument (V1 in Fig. 2), with the one-colored asthenosphere.

rift, the hotspot, magma bodies, and the white area at the bottom of the original image. With these modifications, we then created two images: one with only an orange asthenosphere and another with the orange layer underlain by a yellow layer as seen in the original image.

The four questions initially written for the initial version (V1; Fig. 2) of the survey instrument (Fig. 1) were thoroughly discussed by the authors prior to dissemination of the survey for expert review. This version of the instrument (V1) was presented to our Geocognition Research Laboratory group in late November 2007, and to multidisciplinary (e.g., biology, chemistry, science education) members of the Center for Research on College Science Teaching and Learning at Michigan State University on 7 December 2007 for expert comments. Those comments led to changing the continuous confidence scale to a more easily quantifiable, numeric Likert scale, and making the image smaller so that the instrument could fit in a portrait alignment. This created more room under the image for questions, allowing a fifth question to be added: "Explain why melting occurs in the places you indicated in the figure" (V2).

This second version (V2) of the survey instrument was disseminated to the Geoeducation Research Interest Group listserv (geoed-research@list.msu.edu) on 5 February 2008. Feedback provided further expert validation of the instrument as well as initial ideas for the scoring rubric. On 19 February 2008, just prior to piloting the survey instrument in a nonscience majors class, the mantle lithosphere was thinned beneath the arcs so as to be more scientifically accurate (Strahler, 1998). This aspect change also aligns with the newest version of the web-based USGS image (Vigil and Tilling in Simkin et al., 2006; http://mineralsciences.si.edu/tdpmap/fom/xsection.htm). This version (V3) was pilot tested in a physical science for nonscience majors course (20 February 2008; n = 49) and in our initial, interview with a geoscience graduate student (26 February 2008).

The pilot testing of V3 provided a good example of how novices can notice aspects of an image that experts may not, and it illustrated how novices and experts can interpret questions differently. During the first interview, the interviewee saw and commented on an island and guyot that had not been masked in the image, and one of the students in the pilot course labeled the island as a hotspot (Fig. 1). We had previously removed the obvious hotspot feature in the image, and now recognized the need to remove the island and guyot from the survey instrument (V4). In reviewing the student responses to question 1, we realized we needed to modify how the question was worded. With the original version of: "Label anything related to plate tectonics," some respondents wrote "PT" over areas of the map that they felt were related to plate tectonics. The question was intended to probe a participant's ability to name specific features, and a response of "PT" was too generic for interpretation. Such a response could mean that (1) the respondent knows the name of the feature, but thinks that a generic label is an appropriate answer; (2) the respondent cannot remember the name of the feature; or (3) the respondent thinks the feature is related to plate-tectonic processes but is unsure. In an effort to minimize generic responses, we modified question 1 to read: "Identify anything related to plate tectonics."

Responses to question 2 in the pilot class resulted in rephrasing, as well. The original version read: "Show where you think melting could be occurring." Some respondents circled areas to indicate where they thought melting could occur; others wrote the word, "melting." When respondents used a circle, it tended to encircle an area such as a volcano, a trench, the subducting slab, or the "tip" of the subducting slab. However, when respondents wrote, "melting", it was sometimes [written] near a volcano, but not necessarily over the peaks of the volcanoes or below the volcano. "Melting" was also written near a subducting slab, or in the mantle next to the "tip" of the slab but not directly over it. Because the "tips" of the slabs and the volcanoes were very commonly circled responses, we felt it was likely that those who wrote "melting" near to, but not on top of, these features were likely indicating those features. However, our uncertainty in participant intentions prohibited precise coding of these "melting" data. As a consequence, question 2 was rephrased to read: "Circle areas below the surface where you think melting is occurring." This modification improved our ability to accurately code responses.

In addition to the pilot testing, we were continuously open to modifying the survey instrument in response to participant data. Throughout the study, interviews with participants whose

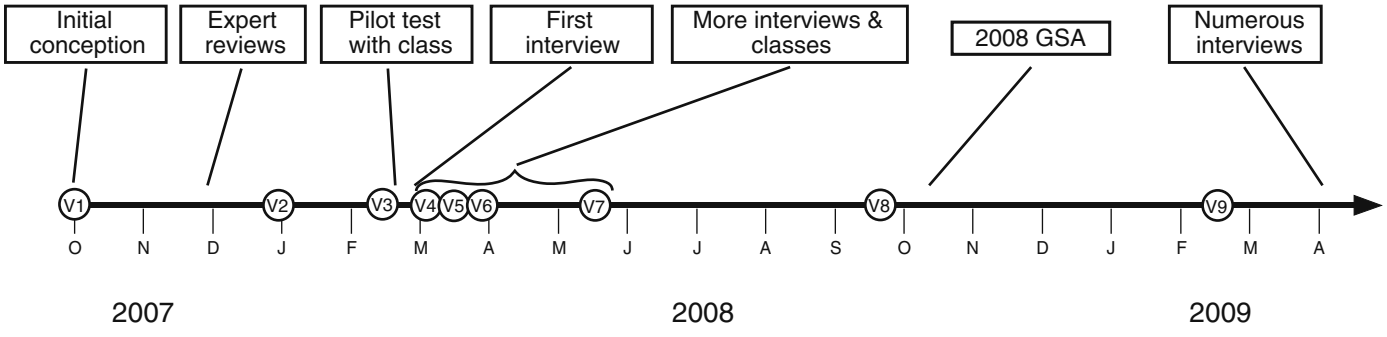


Figure 2. Time line showing the evolution of the survey instrument, indicated by version (V1–V9), and data collection events. After version 6, no changes were made to the image or to the wording of the original five questions. Versions 7 and 8 incorporate novel questions.

geoscience background ranged from novice to expert were conducted concurrently with self-administered survey instruments completed by both college-level students and attendees to a national geoscience conference (Fig. 2). Data from these interviews and completed survey instruments yielded results that both supported survey design and suggested necessary modifications. A modified version (V5) of the survey was disseminated to three courses in mid- to late March 2008. A few of the responses continued to not explicitly identify the geologically relevant features, so we again rephrased question 1 to read: "Identify by name any features related to plate tectonics." This modification further reduced the number of generic responses, and highlighted the need to be open to modifying questions to accommodate differences between our and the participants' reading (see communication validity in following). This version (V6) of the survey instrument was used through the end of April 2008, and no further changes were made to the image or to this initial set of questions.

As our data collection progressed, we obtained responses that led us to add more questions to the survey. For example, many respondents stated that the orange color represented magma. To gain additional information, we added the question, "Estimate the percentage of the mantle that is liquid (magma)." Given our focus on investigating peoples' fundamental understanding of plate tectonics, we also added the question, "Explain what causes tectonic plates to move." This seventh version (V7) was completed in late April and May 2008, while the final version (V8) was completed in September 2008 with the addition of one final question: "How many tectonic plates are in the image? (Number the plates on the figure.)" This question, which was based on a discussion between the first author and Mark Reagan, an igneous petrologist at a public Midwestern university, is in line with our research objectives, and has provided a wealth of information (e.g., Kortz et al., this volume). Both versions 7 and 8 were used during a data collection effort at the GSA annual meeting in October 2008. Version 7 was used in a booth where meeting attendees were invited to complete a survey; 182 attendees filled out the survey at the meeting. Version 8 was used during the 11 interviews that took place at the meeting. The current wording of the last two questions were finalized in February 2009, while the first author was working with Karen Kortz and one of her students to design a slightly modified survey instrument (see Kortz et al., this volume). These nine questions comprise the current version (V9) of the survey instrument (Fig. 3), which was subsequently used in 42 interviews between March and April 2009. Responses to the four questions that were added after the pilot testing of the survey instrument were continuously monitored for any communication validity issues. We did not detect any misunderstanding arising between the targeted concept of the questions and study participants' responses.

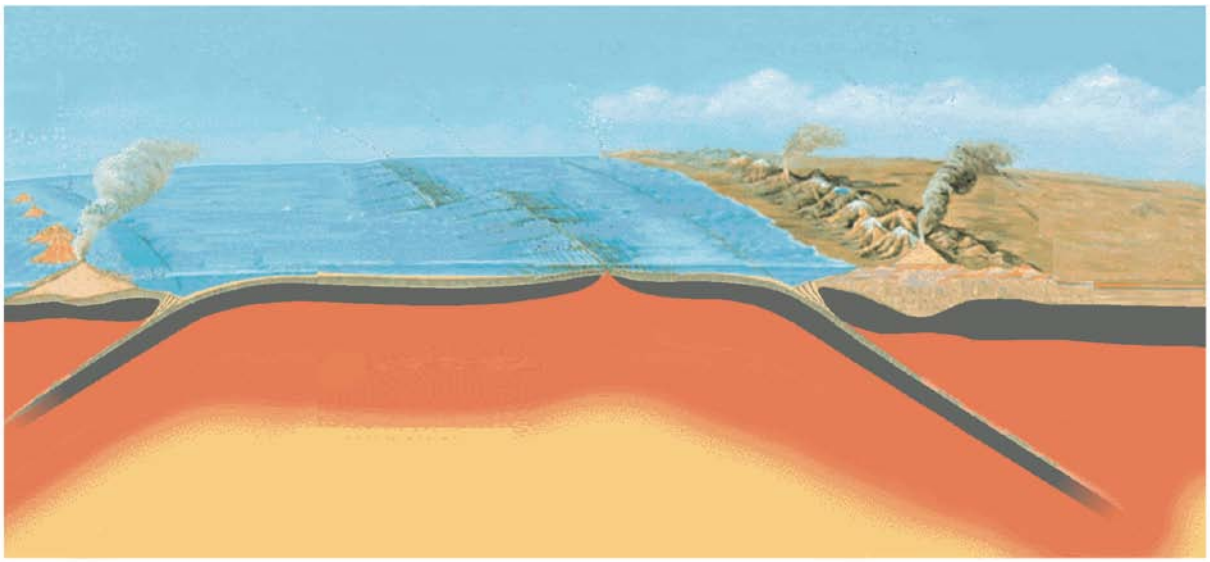
The preceding discussion illustrates how the instrument evolved concurrently with data collection. Although this does not preclude us from interpreting both early and later data, we do acknowledge that changes to questions can have an effect on subject responses. For example, responses of "melting" in answer to the question, "Show where you think melting could be occurring" cannot be interpreted as rigorously as the responses to the rephrased version, "Circle areas below the surface where you think melting is occurring." Differences in coding for these two versions of question 2 reflect this modification in wording, rather than differences in student conceptual understanding.

Rubric Design

The scoring rubrics used to analyze survey results were developed via iterative thematic content analysis of collected data (see Patton, 2002; Denzin and Lincoln, 1998). Rubric design required about 3 mo of discussion, application, and revision, with further improvements occurring as our contextual analysis continued. Initial versions of the scoring rubrics utilized codes developed during analyses of pilot data. For each question analyzed, the authors independently conducted thematic content analysis on a subset of surveys and discussed their observations. The first author then developed a preliminary scoring rubric for each question based on the analyses and points raised during discussion. Subsequent discussion between both authors yielded a scoring rubric that was grounded in the data and that could be easily explained to an undergraduate coder.

A discussion of the development of the rubrics for questions 1 and 2 can provide insights into the effort required to fully develop these rubrics. For question 1, respondents' terms were originally categorized into number of correct terms, number of incorrect terms, and total number of terms used. This approach was abandoned in favor of scoring each, individual term used by each respondent as correct, incorrect, or partially correct/incomplete. This coding scheme is effective at providing insight into participant understanding and use of language, and led to the construction of a list of commonly used terms against which each newly scored response can be compared. The list of commonly used terms grew as the study population expanded from mostly novices to include more intermediate to expert participants. For example, as one might expect, most novices do not identify features such as a forearc basin in the image. However, enough attendees at the GSA meeting in October 2008 did use the term to warrant its addition to the list. Other changes to the list dealt with how nuanced differences in term usage are handled. One example is whether or not to have a separate code for use of the word "crust" when it is used without a modifier to label the oceanic crust versus the continental crust. If a respondent writes only "crust" and uses an arrow or line to indicate that they are labeling the continental crust, then one might presume that the respondent intended the term to be understood as "continental crust." However, without a follow-up interview, ambiguity remains as to whether the term was intended to specifically label the continental crust or crust, in general. This issue was most clearly seen when a respondent labeled only one surface feature as "crust."

Responses to question 2 were the most difficult to analyze, and as with question 1, our initial scoring rubric was discarded



For each response, please circle the number that most closely corresponds to your confidence level.

ON THE FIGURE ABOVE, PLEASE:

	not confident at all				very confident
1) Identify by name any features related to plate tectonics.	1	2	3	4	5
2) Circle areas below the surface where you think melting is occurring.	1	2	3	4	5
3) Use arrows to indicate the relative direction tectonic plates are moving.	1	2	3	4	5
4) Draw a line along each plate boundary and identify the type of each of the boundaries.	1	2	3	4	5
5) How many tectonic plates are in the image? Number of tectonic plates: _____ Number the plates on the image.	1	2	3	4	5

IN THE SPACE BELOW (AND ON THE BACK, IF NECESSARY), PLEASE:

6) Explain what the colors below the surface represent.	1	2	3	4	5
7) Explain why melting occurs in the places you indicated in the figure.	1	2	3	4	5
8) Estimate the percentage of the mantle that is liquid (magma).	1	2	3	4	5
9) Explain what causes tectonic plates to move.	1	2	3	4	5

Figure 3. Current version of survey instrument (V9 in Fig. 2), with the two-colored asthenosphere. A one-colored asthenosphere version is also used.

as we endeavored to accurately represent the intent of the responses. During the summer of 2008, we designed and revised a rubric (Figs. A1 and A2) to the point where we attained an initial inter-rater agreement of 80% between the two authors, and a postdiscussion, inter-rater agreement of 100% on a set of 20 randomly selected survey instruments. The majority of the nonagreement was due to missed coding of terms. This inter-rater process occurred over a number of weeks, and although this rubric did allow us to code those areas what were most frequently indicated by respondents, it was overly complicated and did not necessarily align with the circles given by the subjects' responses. In looking over our initial approach, we had not truly allowed the data to speak for itself. We were literally trying to fit round pegs, the data, into rectangular holes, our rubric (Fig. A1). We abandoned this initial rubric and created a new rubric (Fig. A3) that more closely aligned scores with how subjects marked the image. Most of this revised rubric (i.e., the first 11 groupings) was developed over 2 wk in August 2008. As analyses proceeded, three more groupings were added to account for new themes as observed in responses. We found that the protocol needed to be very explicit in order to maintain a high inter-rater agreement and for temporal consistency for individual raters (see also Ambrose et al., 2004; Bresciani et al., 2009, and references therein). For example, the diagonal lines perpendicular to the subducting slabs were added as a guide for determining whether a specific circle was to be coded as a "4" or a "5." If the center of a subject's circle was above the line, then it was coded as a "4"; if the center of the circle was below the line, it was coded as a "5."

Both authors were involved in the development of all coding rubrics. The first author and one trained, undergraduate geoscience major coded questions 1–3 for a randomly selected set of 60 completed surveys. Training consisted of a discussion of the objectives of the study, the design of the survey instrument, and intended approaches for use of the scoring rubrics. Prior to coding the data, the student rater practiced applying the rubrics. During this phase of training, both authors worked with the student rater to clarify how to apply the rubrics to the data set.

Agreement of independently obtained codes between the first author and the student rater was initially 81.5%, 83.5%, and 90%, for questions 1–3, respectively. After initial scoring, the researchers discussed their scores, and attained a consensus agreement. The majority of the nonagreement was due to missed terms. After this establishment of inter-rater reliability, the undergraduate rater scored a further 184 surveys, independently. As a further step in our validity, she flagged ambiguous responses for later inter-rater discussion. To date, we have developed reliable rubrics for the first three questions. Rubrics for questions 4–9 have not been developed with the same rigor as with the first three questions because current scoring of these responses is not sensitive to nuances in answers. That said, as we continue to analyze our data, we will continue to assess the coding for all of the questions, and will revise and even construct new rubrics if and when that becomes necessary.

SUMMARY

The often circuitous and iterative development pathways described herein provided measures of a number of forms of validity and reliability for both the survey instrument and the rubrics used to score the instrument. Although we did not necessarily set out to establish all of these measures, retrospective evaluation of our research design was made possible through careful record keeping, which allowed us to document an audit trail. The development, validation, and scoring of a mixed-methods survey instrument is difficult and nonlinear; the right-hand column in Table 1 is derived from the culmination of the piloting, revision, and analytical blind alleys described here. At this stage in the research project, we can easily articulate the forms of validity and reliability that have been addressed, intentionally or unintentionally. We also note that we have not addressed all types of validity and reliability that may be considered important for survey instrument development. Table 1 provides explicit details of how each form of validity and reliability was, or was not, addressed.

Intentional Forms of Validity, Reliability, and Trustworthiness

Several forms of validity and reliability were intentionally targeted in our research design. In particular, we knowingly established content and conclusion validity, inter-rater reliability, credibility, dependability, and transferability of our work (Table 1). Content validity was established early in our work through collection of expert feedback on the survey instrument, including both design and content. In addition to expert opinion, we utilized novice responses to early versions of the survey instrument to inform revisions (see also communication validity).

Conclusion validity and credibility are both inherently difficult to measure and should be reviewed well after a study is considered completed. Bias in our interpretations was limited through careful discussion of findings and implications within our research group. In addition, oral and poster presentation of this research at professional meetings and in seminars exposed a variety of experts to our study conclusions; in general, experts agreed with our interpretations of the data in terms of expert-novice trends and implications for knowledge representation in images. Finally, and as documented herein, we carefully considered inter-rater reliability in designing assessment rubrics. Each of these forms of rigor and trustworthiness, coupled with the detailed description of our survey and rubric design as documented in this manuscript, lends dependability to our study (see Libarkin and Kurdziel, 2002) and provides a mechanism for other researchers to evaluate their agreement with our overall conclusions.

A limitation of this study is that while we did not particularly request participation from individuals, we did target specific entry-level courses and specific levels of expertise. As a result our sample is not entirely random; this is an inherent limitation to any survey research. Therefore, although one can never

completely address transferability of study findings, we sampled as broad and diverse a population as was feasible. While we cannot assume that our findings are applicable to all members of the expert-novice population, we have sampled broadly in terms of numbers and geographic distribution (Table 1) in an attempt to provide some far-reaching, and hence transferable, significance to our work. Finally, we acknowledge the importance of cultural validity to establishing transferability. Although we did not explicitly address cultural validity in our work, we encourage those interested in adapting this instrument to other cultures to consider the appropriateness of the survey design to their targeted demographic.

Unintentional Forms of Validity and Reliability

Although our intention was to construct an instrument that would provide insights into the conceptions held by individuals across the expert-novice continuum, we did not recognize the potential for documenting construct validity until we began analyzing our data and documenting the detailed responses of experts (Table 1). In particular, the most experienced experts provided more thorough and accurate responses than novices. Similarly, criterion validity was recognized through poststudy comparison of interview with survey results, as well as on a smaller scale through comparison with data collected in an unrelated study (Kortz et al., this volume). Although we did not intentionally target communication validity early in the study, some student responses to question 1 were initially so generic that they prompted us to revise the question until nearly everyone who answered the question provided feature-specific labels. Finally, the duration of our data collection and use of multiple forms provided us with a way to address internal consistency reliability. In particular, we find that the results from the survey instrument, separated by 14 mo and representing several different versions, are consistent across implementations.

REFLECTIONS ON THE PROCESS

In many ways, our research proceeded in ways that are similar to a stereotypical natural science research project. This project began with a question: the first author looked at a textbook image of plate-tectonic processes and asked himself, "Is this image confusing to students?" This led to a hypothesis: "The differences between how novices and experts view plate-tectonic representations can create barriers to learning." We felt we could investigate people's perceptions of plate tectonics, and study the role played by an image in affecting people's perceptions of plate tectonics in a well-designed survey instrument. We designed our instrument and then performed an initial check of the rigor and trustworthiness of the instrument through expert review and pilot testing. Next, we collected the bulk of our data while concurrently beginning our data analysis. Currently, we are continuing our analysis and documenting our findings for dissemination in publications. Our time line from initial conceptualization in October 2007, Institutional Review Board (IRB) approval for research with human subjects in January 2008, first implementation of the instrument in February 2008, presentation of initial findings on novices in October 2008 (Clark and Libarkin, 2008), receipt of National Science Foundation (NSF) funding in January 2009, to submitting research findings for publication in 2010, follows that of a typical research project.

One difference is that we had to create the instrument needed for measuring the traits we were interested in studying. Although instrument development is done in the natural sciences, it is not typical for most projects. Just as in the natural sciences, where an instrument's accuracy and precision must be determined, we needed to determine the rigor and trustworthiness of our instrument. For this project, rigor and trustworthiness steps required about the same amount of effort as was needed for designing the instrument. Indeed, rigor and trustworthiness testing is an ongoing process. We have asked ourselves, "When do we stop modifying a rubric?" Although we achieved 100% postinstruction inter-rater agreement on our first rubric for question 2, we felt our approach was not aligning well enough with how respondents answered the question. The revised rubric has been further tweaked at least three times, but any future potential benefits of refining our interpretations that might be gained through additional changes must be weighed against the need to be able to compare earlier scored surveys against more recently scored surveys, possibly requiring rescoring of all surveys. We feel our current rubrics are effective, while accepting that they are not perfect. At some point we have to say, "It's good enough."

In the normal course of doing research, we expected to repeatedly modify the instrument, recruit for and schedule interviews, recruit professors who would allow us access to their students, and obtain IRB approval of the instrument and study methods. During interviews, the first author encouraged subjects to provide as much detail as they wished in their explanations while trying to avoid leading questions, without coming across as didactic, and without making value judgments on responses. When subjects provided what was deemed to be an interesting explanation of a plate-tectonic process, whether scientifically valid or not, the goal was to probe deeply so as to obtain as much insight into subject's thoughts on the topic as possible (Kvale and Brinkmann, 2009).

An unexpected aspect of the research has been the unique challenge posed by interviewing experts. Whether asking experts questions that they perceived as too simple or pressing them to explain their reasoning on a topic for which they held an alternative conception, one has to be careful to not inadvertently offend the participant. Although neither author claims to be an expert in all facets of plate-tectonics research, as interviewers we needed to be well informed on the topic. Other facets that were not necessarily foreseen in the planning stages included how to handle the amount of data that quickly became quite substantial. Part of this data accumulation was due to addition of new questions to the instrument as the study progressed. This could be considered a problem of riches because those additional questions provided

important insights into the ways in which many plate-tectonic concepts are perceived along the expert-novice continuum. We also learned that one needs to be willing to scrap weeks of work invested in a rubric, and to design rubric protocols that are clear and explicit. The fewer interpretations in data analysis that are left to the discretion of a coder, the more likely that coder is to score the same survey the same way each time, and the more likely two coders are to score a survey similarly.

We feel that this instrument and associated rubrics are providing a wealth of data, and we feel that we did need to create this survey instrument. However, we would encourage researchers to adopt preexisting valid and reliable research instruments, whenever possible. When it is not possible, be prepared to invest a significant amount of time and effort in creating, validating, and revising your instrument and scoring rubric.

APPENDIX

Coding rubric protocols for question 2. The original protocol was implemented in August 2008, but it was replaced by the currently used protocol starting in September 2008.

Original Protocol:

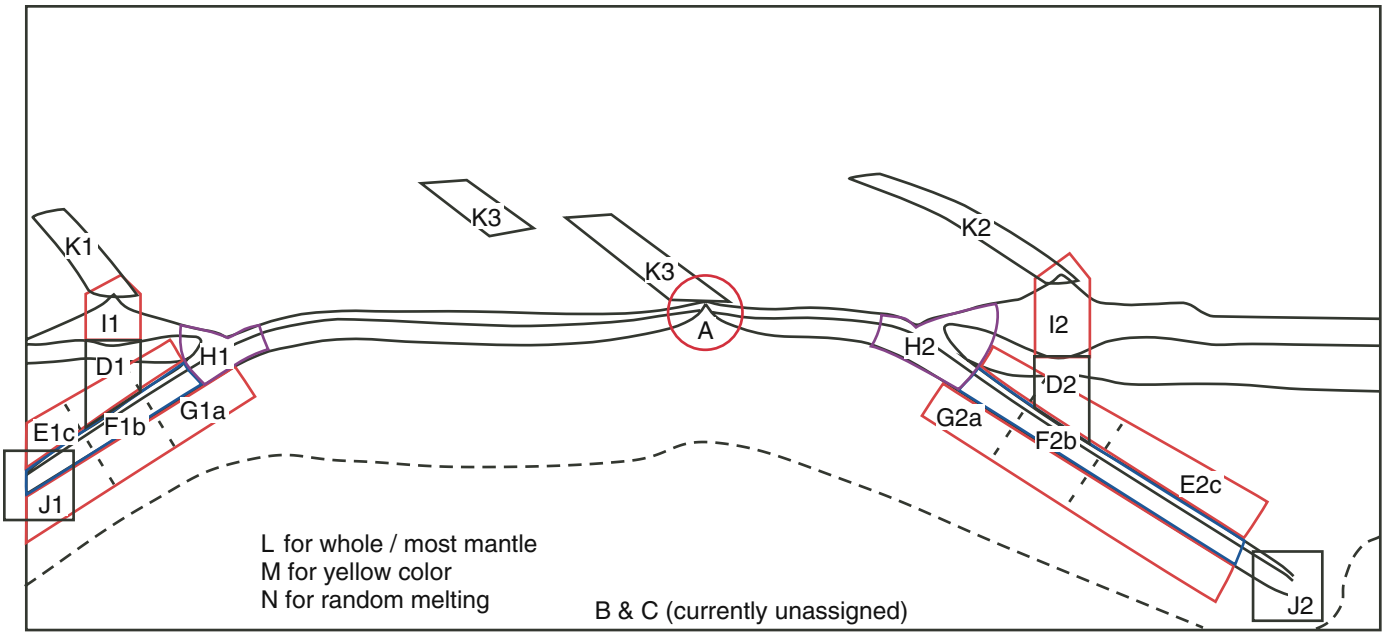


Figure A1. Early version of coding template for question 2. The template was printed on a transparency that was laid over a subject's responses.

If a circle encompasses ≈≥50% of a labeled zone, count that zone.

If ≈≥50% of a circle is within a zone, count that zone.

The subdivisions, **a, b,** and **c** of **E, F,** and **G** are designed to capture circled areas within those zones. E, F, and G are designed to capture ellipses parallel to the subducting slab. An ellipse of **E1a** and **E1b** would be **E1**, but a circle of **E1a, E1b, F1a, F1b, G1a,** and **G1b** would be listed as all of those.

D: Use for that specific area or circles in that area—do NOT include **D** in ellipses along slab.

K: Include any circled areas over volcanic peaks or mid-ocean ridges—except those responses that are centered on **A** or **I.**

J: If a circle is interpreted to represent the "end" of the slab, it should be adjudged as **J** regardless of its size.

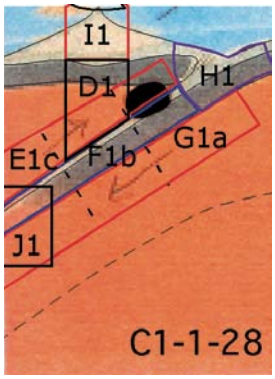
L: Use if respondent indicates all or most of the mantle.

M: Use if respondent's circle(s) or "melting indicators" are random, arbitrary, or not included within defined zones.

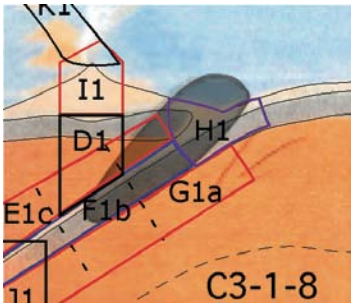
N: Use if respondent circled the orange color (of two-color mantle images).

O: Use if respondent circled the yellow color (of two-color mantle images).

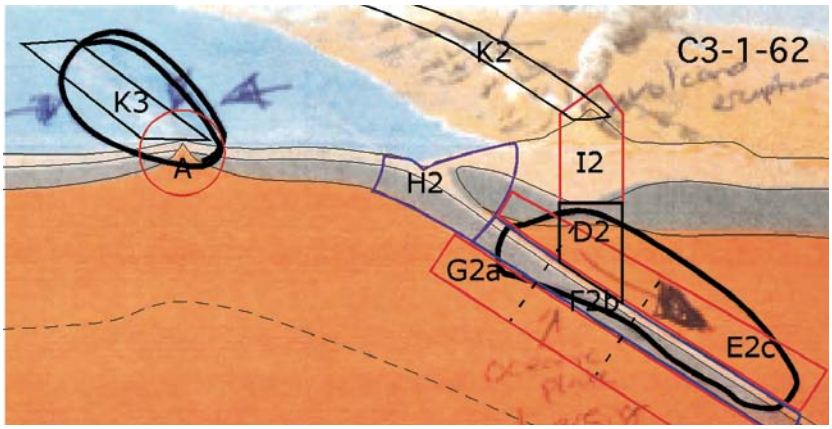
P: Use if melting is indicated by something other than circles (e.g., text or arrows).



code as: E1a



code as:
E1a, F1a, F1b, H1



code as: K3, and as E2, F2

code as: I1, K1, and as I2, K2

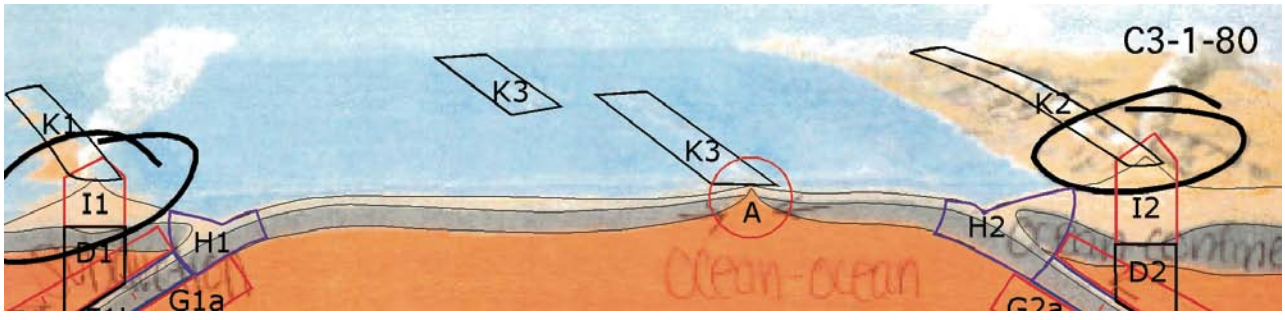
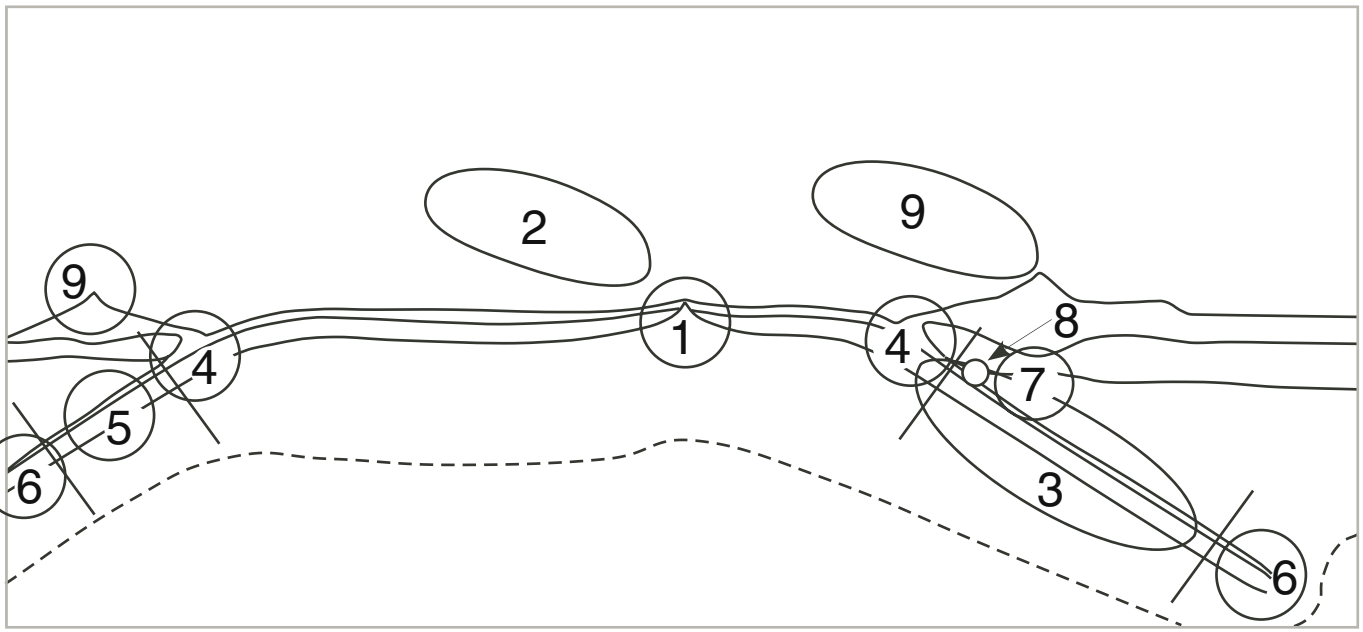


Figure A2. Images used to clarify how to apply the scoring of the template in Figure A1. Subject responses have been accentuated with a dark line or with a darkened area.



PAY ATTENTION TO ANY LABELS ON CIRCLES (Not all circles indicate melting. See code 12)

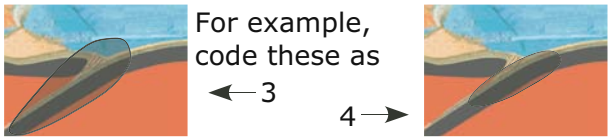
Use notes column when necessary to clarify a code, esp. useful for codes 10 and 11

Circles along the descending slabs whose center is within the diagonal lines are coded as 5

Codes apply to either or both sides of figure. For example, if a respondent circles the 'tips' of *either* or *both* plates, code this as a 6

'circle' and 'ellipse' are used in a relative, not exact sense

Every melting area should receive only ONE CODE



For a circle that covers area 4 but looks to also include 8, code that as only a 4; 8 is specifically for small circles in the corner of the mantle wedge.

CODES

1 circle at divergent boundary
2 circles over ocean ridges
3 ellipse along a significant part of subducting plate (inc. directly above &/or below plates)
4 circle over trench(es)
5 circle along middle of subducting plate(s)
6 circle at bottom 'tip' of subducting plate(s)
7 circle in mantle wedge directly below volcanoes
7b circle centered below a volcano but above asthenosphere
8 small circle in corner of mantle wedge
9 circle over volcanoes
10 circle over area outside of codes 1 - 9
11 something other than circles indicating melting (e.g., text, arrows)
12 circles that indicate something other than melting
13 no indication of melting by circles, text, arrows, etc.
14 circle includes mantle wedge ± crust ± trench ± upper section of descending slab. This circle must be too large to be classified as 5, 7, or 8, and is not a 9. Circle may include parts of 4, 5, 7, 8, & /or 9.

Example of a #14 code

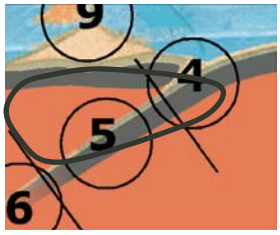


Figure A3. Current rubric for scoring question 2. Note that codes are symmetric for the subducting slabs. That is, both slabs have are coded for "3," "4," and "5" circles. The rubric is printed on a transparency that is laid over a subject's responses. Subject responses have been accentuated with a dark line or with a darkened area.

ACKNOWLEDGMENTS

We thank everyone who participated in this research. Discussions with and assistance from the co–principal investigators of National Science Foundation (NSF) project 0837185 and members of the Geocognition Research Laboratory at Michigan State University, especially Sarah Jordan, are appreciated. This manuscript benefited greatly from comments by Alison Stokes and two anonymous reviewers. Partial support for this work was provided by the National Science Foundation's DUE Course, Curriculum, and Laboratory Improvement program under awards 0717790 and 0837185.

REFERENCES CITED

Adams, W.K., Perkins, K.K., Podelefsky, N.S., Dubson, M., Finkelstein, N.D., and Wieman, C.E., 2006, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey: Physical Review Special Topics–Physics Education Research, v. 2, p. 010101:1–14.

Ambrose, R., Clement, L., Philipp, R., and Chauvot, J., 2004, Assessing prospective elementary school teachers' beliefs about mathematics and mathematics learning: Rationale and development of a constructed-response-format beliefs survey: School Science and Mathematics, v. 104, no. 2, p. 56–69, doi:10.1111/j.1949-8594.2004.tb17983.x.

Bandura, A., 1984, Recycling misconceptions of perceived self-efficacy: Cognitive Therapy and Research, v. 8, no. 3, p. 231–255, doi:10.1007/BF01172995.

Beatty, P.C., and Willis, G.B., 2007, Research synthesis: The practice of cognitive interviewing: Public Opinion Quarterly, v. 71, no. 2, p. 287–311, doi:10.1093/poq/nfm006.

Bresciani, M.J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., and Hickmott, J., 2009, Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines: Practical Assessment, Research, & Evaluation, v. 14, no. 12, p. 1–7. (Available online: http://pareonline.net/getvn.asp?v=14&n=12.)

Britton, E.D., and Schneider, S.A., 2007, Large-scale assessments in science education, *in* Abell, S.K., and Lederman, N.G., eds., Handbook of Research on Science Education: Mahwah, New Jersey, Lawrence Erlbaum Associates, p. 1007–1040.

Clark, S.K., 2009, Plate tectonics as viewed by novices and experts: Geological Society of America Abstracts with Programs, v. 41, no. 7, p. 251.

Clark, S.K., and Libarkin, J.C., 2008, Postinstruction alternative conceptions about plate tectonics held by nonscience majors: Geological Society of America Abstracts with Programs, v. 40, no. 6, p. 364.

Collins, D., 2003, Pretesting survey instruments: An overview of cognitive methods: Quality of Life Research, v. 12, p. 229–238, doi:10.1023/A:1023254226592.

Creswell, J.W., 2003, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (2nd ed.): Thousand Oaks, California, Sage Publications, 246 p.

Creswell, J.W., and Miller, D.L., 2000, Determining validity in qualitative inquiry: Theory into Practice, v. 39, no. 3, p. 124–130, doi:10.1207/s15430421tip3903_2.

Crotty, M., 1998, The Foundations of Social Research: Thousand Oaks, California, Sage Publications, 248 p.

Denzin, N.K., and Lincoln, Y.S., eds., 1998, Strategies of Qualitative Inquiry: Thousand Oaks, California, Sage Publications, 346 p.

Driver, R., 1989, Students' conceptions and the learning of science: International Journal of Science Education, v. 11, p. 481–490, doi:10.1080/0950069890110501.

Duschl, R.A., Schweingruber, H.A., and Shouse, A.W., eds., 2007, Taking Science to School: Learning and Teaching Science in Grades K–8: Washington, D.C., National Academies Press, 387 p.

Feig, A., 2010, this volume, Methodology and location in the context of qualitative data and theoretical frameworks in geoscience education research, *in* Feig, A.D., and Stokes, A., eds., Qualitative Inquiry in Geoscience Education Research: Geological Society of America Special Paper 474, doi:10.1130/2011.2474(01).

Fink, A.G., 2003, The Survey Handbook (2nd ed.): Thousand Oaks, California, Sage Publications, 184 p.

Golafshani, N., 2003, Understanding reliability and validity in qualitative research: Qualitative Report, v. 8, no. 4, p. 597–607.

Greene, J.C., Caracelli, V.J., and Graham, W.F., 1989, Toward a conceptual framework for mixed-method evaluation designs: Educational Evaluation and Policy Analysis, v. 11, no. 3, p. 255–274.

Harp, S.F., and Mayer, R.E., 1998, How seductive details do their damage: A theory of cognitive interest in science learning: Journal of Educational Psychology, v. 90, no. 3, p. 414–434, doi:10.1037/0022-0663.90.3.414.

Hossler, D., and Vesper, N., 1993, An exploratory study of the factors associated with parental saving for postsecondary education: The Journal of Higher Education, v. 64, no. 2, p. 140–165, doi:10.2307/2960027.

Johnson, R.B., and Onwuegbuzie, A.J., 2004, Mixed methods research: A research paradigm whose time has come: Educational Researcher, v. 33, no. 7, p. 14–26, doi:10.3102/0013189X033007014.

Kidder, L.H., and Fine, M., 1987, Qualitative and quantitative methods: When stories converge: *in* Mark, M.M., and Shotland, R.L., eds., Multiple Methods in Program Evaluation: San Francisco, California, Jossey-Bass, p. 57–75.

Kortz, K.M., Clark, S.K., Gray, K., Smay, J.J., Viveiros, B., and Steer, D., 2011, this volume, Counting tectonic plates: A mixed-methods study of college students' conceptions of plates and boundaries, *in* Feig, A.D., and Stokes, A., eds., Qualitative Inquiry in Geoscience Education Research: Geological Society of America Special Paper 474, doi:10.1130/2011.2474(12).

Kvale, S., and Brinkmann, S., 2009, Interviews: Learning the Craft of Qualitative Research Interviewing: Thousand Oaks, California, Sage Publications, 354 p.

Lederman, N.G., Abd-El-Khalick, F., Bell, R.L., and Schwartz, R., 2002, Views of nature of science questionnaire: Toward valid and meaningful assessment of learner's conceptions of nature of science: Journal of Research in Science Teaching, v. 39, p. 497–521, doi:10.1002/tea.10034.

Lewis, J., 2009, Redefining qualitative methods: Believability in the fifth moment: International Journal of Qualitative Methods, v. 8, no. 2, p. 1–14.

Libarkin, J.C., 2008, Concept inventories in higher education science, *in* National Research Council, Promising Practices in Undergraduate STEM Education Workshop 2 (Washington, D.C., 13–14 October 2008): http://www7.nationalacademies.org/bose/PP_Commissioned_Papers.html (accessed 30 October 2009).

Libarkin, J.C., and Kurdziel, J.P., 2002, Research methodologies in science education: Qualitative data: Journal of Geoscience Education, v. 50, no. 2, p. 195–200.

Lincoln, Y.S., and Guba, E.G., 1985, Naturalistic Inquiry: Beverly Hills, Sage Publishing, 416 p.

Litwin, M.S., 1995, How to Measure Survey Reliability and Validity: Thousand Oaks, California, Sage Publications, 87 p.

Lopez, W., 1996, Communication validity and rating scales: Rasch Measurement Transactions, v. 10, no. 1, p. 482–483.

Marshall, C., and Rossman, G.B., 2006, Designing Qualitative Research: Thousand Oaks, California, Sage Publications, 262 p.

Maxwell, J.A., 2005, Qualitative Research Design: An Interactive Approach (2nd ed.): Thousand Oaks, California, Sage Publications, 192 p.

Merriam, S.B., 2002, Introduction to qualitative research, *in* Merriam, S.B., ed., Qualitative Research in Practice: San Francisco, California, Jossey-Bass, p. 3–17.

Morse, J.M., Barrett, M., Mayan, M., Olson, K., and Spiers, J., 2002, Verification strategies for establishing reliability and validity in qualitative research: International Journal of Qualitative Methods, v. 1, no. 2, p. 13–22.

Onwuegbuzie, A.J., and Johnson, R.B., 2006, The validity issue in mixed research: Research in the Schools, v. 13, no. 1, p. 48–63.

Patton, M.Q., 2002, Qualitative Evaluation and Research Methods (3rd ed.): Thousand Oaks, California, Sage Publications, 598 p.

Phillips, D.C., and Burbules, N.C., 2000, Postpositivism and Educational Research: Lanham, Maryland, Rowman & Littlefield Publishers, 101 p.

Posner, G.J., Strike, K.A., Hweson, P.W., and Gertzog, W.A., 1982, Accommodation of a scientific conception: Toward a theory of conceptual change: Science Education, v. 66, no. 2, p. 211–227, doi:10.1002/sce.3730660207.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J.M., and Singer, E., 2004, Methods for testing and evaluating survey questions:

Public Opinion Quarterly, v. 68, no. 1, p. 109–130, doi:10.1093/poq/nfh008.

Rossman, G.B., and Wilson, B.L., 1985, Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study: Evaluation Review, v. 9, no. 5, p. 627–643, doi:10.1177/0193841X8500900505.

Sanchez, M.E., 1992, Effects of questionnaire design on the quality of survey data: Public Opinion Quarterly, v. 56, no. 2, p. 206–217, doi:10.1086/269311.

Simkin, T., Tilling, R.I., Taggart, J.N., Jones, W.J., and Spall, H., 1989, This Dynamic Planet: World Map of Volcanoes, Earthquakes, and Tectonic Plates: U.S. Geological Survey (in collaboration with the Smithsonian Institution) Map I-2800, scale 1:30,000,000.

Simkin, T., Ungar, J.D., Tilling, R.I., Vogt, P.R., and Spall, H., 1994, This Dynamic Planet: World Map of Volcanoes, Earthquakes, and Tectonic Plates (2nd ed.): U.S. Geological Survey (in collaboration with the Smithsonian Institution and U.S. Naval Research Laboratory) Map I-2800, scale 1:30,000,000.

Simkin, T., Tilling, R.I., Vogt, P.R., Kirby, S.H., Kimberly, P., and Stewart, D.B., 2006, This Dynamic Planet: World Map of Volcanoes, Earthquakes, Impact Craters, and Plate Tectonics: U.S. Geological Survey (in collaboration with the Smithsonian Institution and U.S. Naval Research Laboratory) Geologic Investigations Series Map I-2800, scale 1:30,000,000 (and companion website http://www.minerals.si.edu/tdpmap).

Siragusa, L., and Dixon, K.C., 2006, A research methodology: The development of survey instruments for research into online learning in higher education: Issues in Educational Research, v. 16, p. 206–225.

Solano-Flores, G., and Nelson-Barber, S., 2001, On the cultural validity of science assessments: Journal of Research in Science Teaching, v. 38, no. 5, p. 553–573, doi:10.1002/tea.1018.

Strahler, A.N., 1998, Plate Tectonics: Cambridge, Massachusetts, GeoBooks Publishing, 554 p.

Tashakkori, A., and Teddlie, C., 1998, Mixed Methodology: Combining Qualitative and Quantitative Approaches: Thousand Oaks, California, Sage Publications, 185 p.

Tashakkori, A., and Teddlie, C., 2003, Handbook of Mixed Methods in Social & Behavioral Research: Thousand Oaks, California, Sage Publications, 768 p.

Thomas, S.J., 2004, Using Web and Paper Questionnaires for Data-Based Decision Making: From Design to Interpretation of the Results: Thousand Oaks, California, Corwin Press, 194 p.

Trochim, W.M.K., and Donnelly, J.P., 2007, The Research Methods Knowledge Base (3rd ed.): Cincinnati, Ohio, Atomic Dog Publishing, 361 p. (companion website: http://www.socialresearchmethods.net/kb/).

U.S. Geological Survey Education webpage, 2010, USGS Map Databases: http://education.usgs.gov/common/map_databases.htm (accessed 12 March 2010).

Manuscript Accepted by the Society 23 June 2010