Concept Maps: Theory, Methodology, Technology
Proc. of the First Int. Conference on Concept Mapping
Pamplona, Spain 2004

EXAMINING CONCEPT MAPS AS AN ASSESSMENT TOOL

Maria Araceli Ruiz-Primo
School of Education, Stanford University
Email: aruiz@stanford.edu

Abstract. In this paper, I describe concept maps as assessment tools for measuring one aspect of achievement, the organization of declarative knowledge in a domain. A concept map-based assessment consists of a task that elicits connected understanding, a response format, and a scoring system. Variation in tasks, response formats, and scoring systems produce different mapping techniques that may elicit different knowledge representations, posing construct-interpretation challenges. The paper provides a framework to guide the research on the technical quality of this tool. It will briefly describe some of the studies conducted and it will focus on some of the dimensions that can be used to define the assessment task.

1 Introduction

Cognitive psychologists posit that, "the essence of knowledge is structure" (Anderson, 1984, p. 5). Assuming that knowledge within a content domain is organized around central concepts, to be knowledgeable in the domain implies a highly integrated conceptual structure among those concepts. This means that as expertise in a domain grows, through learning, training, and/or experience, the elements of knowledge become increasingly interconnected (cf. Chi, Glaser, & Farr, 1988). Researchers have taken different representational approaches to capture this organizational property of knowledge (e.g., Goldsmith, Johnson, & Acton, 1991; Novak & Gowin, 1984; Novak, Gowin, Johansen, 1983; White & Gunstone, 1992). Among these approaches, concept maps have been proposed as a more direct approach (see Ruiz-Primo & Shavelson, 1996) for capturing the interrelatedness among concepts in a domain. It can be easily argued that the dimension of structure of knowledge yielded by concept maps is unique in comparison to traditional achievement tests.

The technology for developing, using, and evaluating concept maps as an assessment tool is currently being investigated. Over the past eight years, we have done research intended to inform a concept map-assessment knowledge base (Ruiz-Primo & Shavelson, 1996; Ruiz-Primo, Schultz, & Shavelson, 1996; Ruiz-Primo, Schultz, & Shavelson, 1997; Ruiz-Primo, Schultz, Li, & Shavelson, 2001; Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, in press). Our goals have been to provide not only evidence about reliability and validity of concept map assessments, but also to provide a framework that can guide further research in this area.

This paper focuses on issues related to the development and technical examination of concept maps as assessment tools. First, I present a framework that can be used to guide the design of concept map assessments. Then, I discuss the approaches we have used to examine the technical qualities of concept maps as assessment tools. In the third part of the paper, I discuss some issues related to the use of concept maps as an assessment tool. Finally, I drew some general conclusions about what we have learned about concept map assessments and what still needs to be done.

2 Conceptualizing Concept Maps as Assessment Tools

A concept map as an assessment can be thought of as a set of procedures used to measure important aspects of the structure/organization of a student's declarative knowledge. The term "assessment" reflects the belief that reaching a judgment about an individual's achievement in a domain requires an integration of several pieces of information. Concept maps based assessment is only one of those pieces (see Cronbach, 1990).

Formally, a concept map is a graph consisting of nodes and labeled lines. The nodes correspond to important terms (representing concepts) in a domain. The connecting lines denote a directional relationship between a pair of concepts (nodes). The label on the line (explanation) conveys how the two concepts are related. The combination of two nodes and a labeled line is called a proposition. A proposition is the basic unit of meaning in a concept map and the smallest unit used to judge the validity of the relationship drawn between two concepts (e.g., Dochy, 1996). Concept maps, then, purport to represent some important aspects of a student's declarative knowledge in a content domain (e.g., physics).

We (Ruiz-Primo & Shavelson, 1996) characterized concept maps assessments in terms of: (a) a task that invites a student to provide evidence bearing on his or her knowledge structure in a domain, (b) a format for the student's response, and (c) a scoring system by which the student's concept map can be accurately and consistently evaluated. Without these three components, a concept map cannot be considered as a measurement tool. This characterization made evident the variation in concept mapping techniques used in research and practice (Ruiz-Primo & Shavelson, 1996).

The importance of comparing different concept map assessment techniques might be better understood if we consider the relationship among the three assessment components (Figure 1). Threats to validity exist at almost every connection, because every connection is established based on assumptions and the model cannot work unless all the assumptions are valid. For example, assessment tasks not only elicit but also influence students' responses. That is, the characteristics of the assessment tasks may make students respond in ways that are not relevant (e.g., guessing) to the construct assessed, or may be too narrow that it fails to tap important aspects of the construct (e.g., the assessment task is too structured). The same can be said about the scoring system. A deficiency in the scoring system may preclude the ability to properly or sufficiently capture information about the quality of students' responses. Research on the variations in concept map assessments techniques is critical.

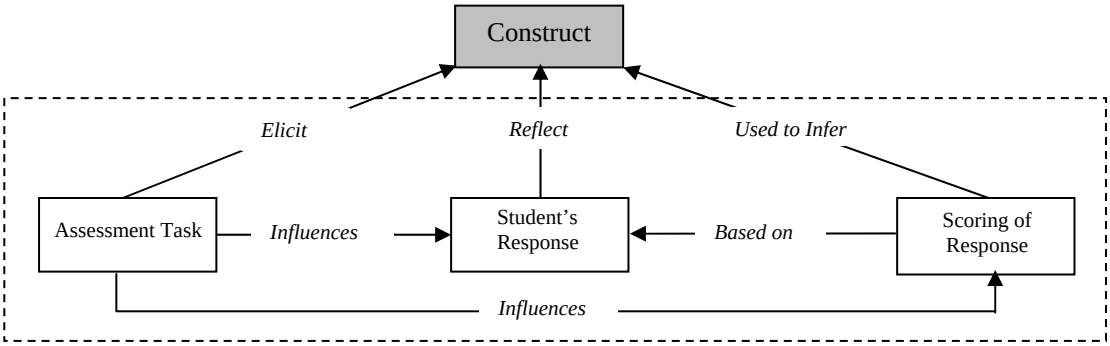

Figure 1. Interaction among assessment goals and assessment components (After Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, internal document).

2.1 The Nature of Concept Map Assessment Tasks, Response Formats, and Scoring Systems

In previous research, we identified the map components provided to the student as a dimension for determining the demands of the mapping techniques. A concept map task assessment could be characterized along a directedness continuum from high-directed to low-directed based on the information provided to the students. High-directed concept map tasks provide students with the concepts, connecting lines, linking phrases, and the map structure. In contrast, in a low-directed concept map task, students are free to decide which and how many concepts they include in their maps, which concepts are related, and which words to use to explain a relationship. The mapping techniques studied in our research have been selected at different points on a directedness continuum (Figure 2). Indeed, we have provided evidence that different mapping techniques do not provide the same information about the students' connected understanding (Ruiz-Primo, Schultz, Shavelson, 1996; Ruiz-Primo, Schultz et al., 2001; Ruiz-Primo, Shavelson et al., 2001; Yin et al., in press). Based on information collected across studies it is clear that directedness involves more aspects that need to be considered in deciding the characteristics of the mapping assessment tasks. In what follows, I explore other aspects that can be considered in deciding the demands of mapping assessment tasks.

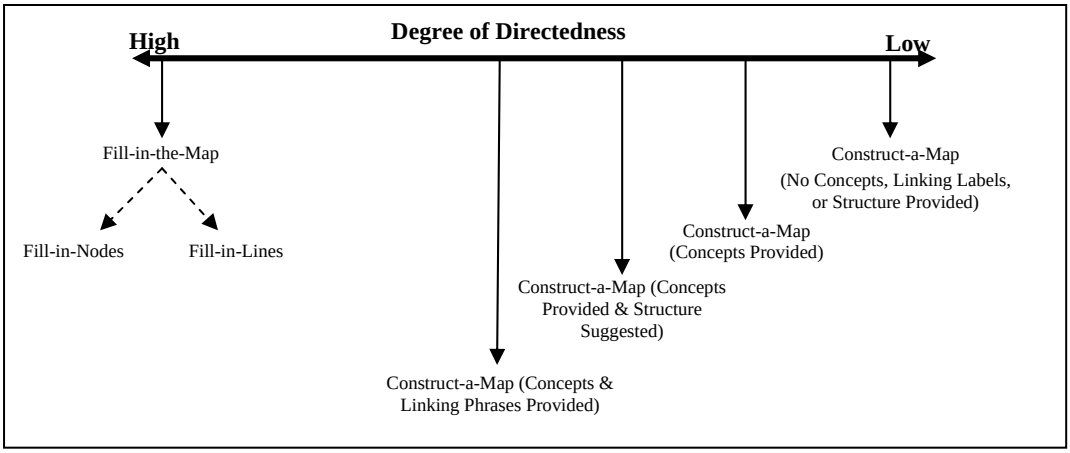

Figure 2. Concept map techniques studied according to directedness of the mapping tasks.

I believe that in order to better determine the cognitive demands evoked by different mapping techniques, it is important to consider not only what is provided, but also the extent/amount of what is provided, the significance of what is provided, and what is required from the examinees. In Figure 3, I propose a framework that takes into consideration different aspects of the nature of the tasks, the response format, and the scoring system. It provides a picture that considers some possibilities involved in deciding the characteristics of the assessment mapping techniques.

Admittedly, the interpretation of Figure 3 is complex. In the figure, the assessment components are shaded. The decisions made about the mapping task and the scoring system are directly related to the map components - what components are provided and with what characteristics, and what is scored. The response format is directly linked to the constraints imposed in the task. I provide an example that may help to understand the rationale of the figure. I focus only in one of the map components, *the terms* (concepts). In designing a concept map assessment, an assessor decides to provide the examinee with the terms to be used in the map, but she has to make another decision about whether to provide only a *sample* of terms or *all* of the terms to be used. Furthermore, if only a sample of terms is provided, the assessor needs to decide which terms should be provided among the ones considered relevant within a particular domain and topic. Terms provided by the assessor in the sample might be *key* (most relevant) terms or only *related* but not key terms. With only a sample of terms provided, the assessor may require from the examinee to provide only a few terms (say, ten) or all the terms the examinee considers as appropriate within a topic. In addition, the assessor should decide whether the terms the examinee provides should be key (most relevant) terms within the topic at hand or only related terms. All these decisions will guide some of the characteristics of the scoring system. If the examinee is to provide the terms, the assessor may decide to score them as correct or incorrect without considering the relevance of the terms. If amount of terms was not posed as a constraint, the assessor may score the quantity of terms provided (extent of the concepts the examinee connects/relates within a topic). If a criterion map is used for scoring, not only other dimensions of the examinee's response (e.g., similarity) should be considered; also, the correctness, relevance and quantity of the map will be determined, somehow, by the characteristic of the criterion map at hand. All these decisions may lead to different score interpretations. Rice, Ryan, and Samson (1998) have suggested that different methods of scoring maps within the same technique may be measuring different constructs of aspects of the domain.

An assessor, of course, can decide to provide all the components, fully or only partially (i.e., a sample of concepts, connections, explanations, and a partial structure of the map). The decisions made within each of the components will be also multiple within each component. Consider, for example, the scoring implications of providing linking phrases from which students will select the ones to be used to connect a pair of terms. If linking phrases are not selected carefully, students are very likely to obtain bipolar scores, which do not reflect the quality of their knowledge.

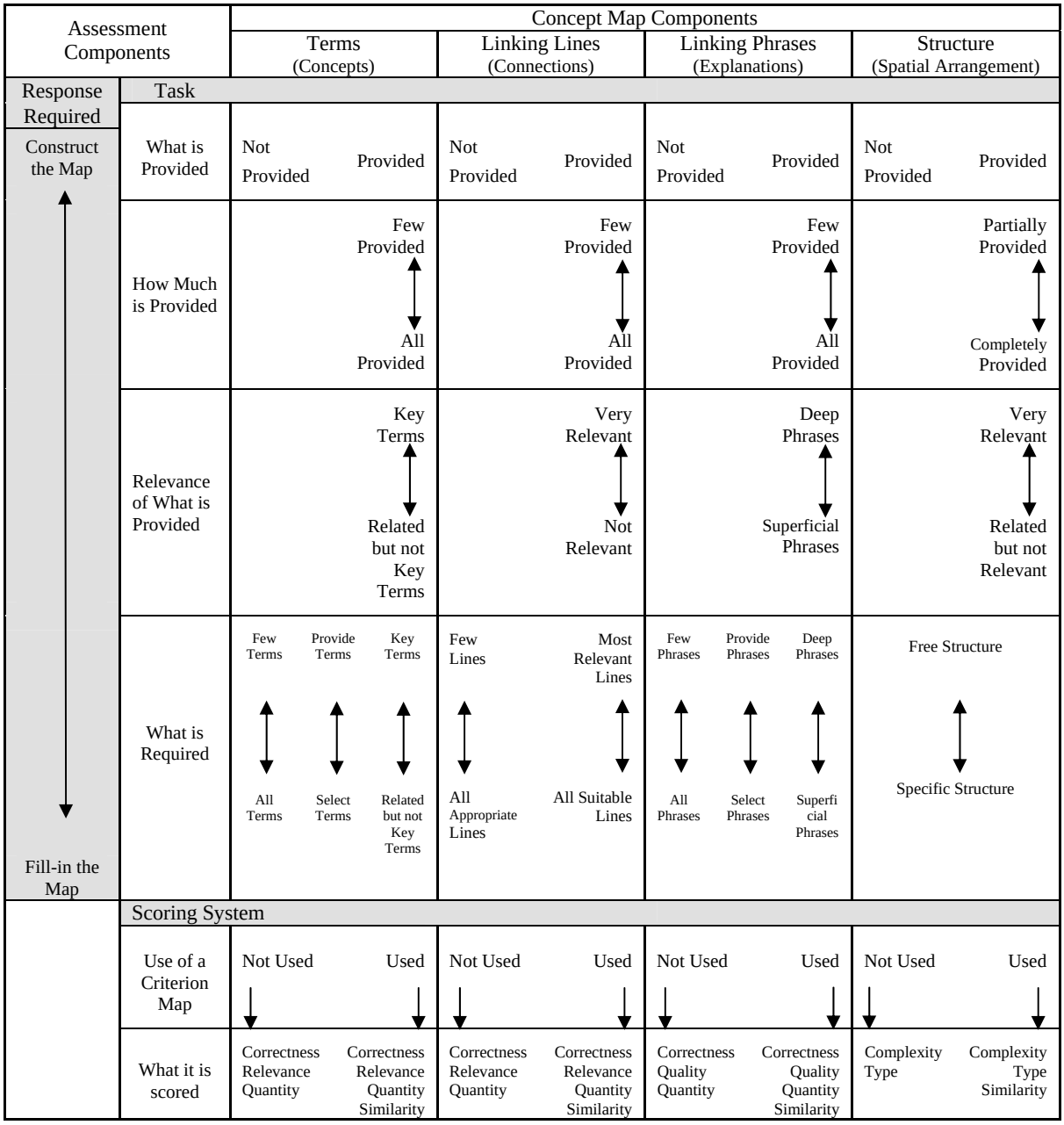

Figure 3. Framework considering some aspects about the nature of the mapping assessment tasks, response formats, and scoring system.

Making obvious the aspects that can be considered in designing a mapping assessment task makes two issues evident: (1) The *end product* of a concept map assessment – the map – can be conceived of as a continuum that goes from a map fully constructed by the students to a map only filled-in by the student,[1] and (2) within each type of end-product, the possibilities that can be considered in the design of a map assessment are multiple. A map completely constructed by the student may vary on what map components are provided, how much it is provided within each component, the characteristics of what is provided, and what is required from the student. Still, notice that although the end-product may be identical - a map completely constructed by the student - the cognitive demands evoked by diverse constraints are different.

In addition to the complexity of these decisions, there is the issue of response mode, what we have called the "method of assessment" in performance assessments. Are paper-and-pencil and computer-based concept

[1] Think, for example, of a mapping technique in which the examinee is provided with an "unfinished" map that s/he has to complete. Another example in the continuum closer to the fill-in-the-map can be the case in which the examinee fills-in the partially structure of the map provided but s/he will also finish the construction of the map.

map assessments exchangeable? Are interpretations across methods the same when using the same mapping technique? This is a research topic that also deserves a long research agenda in its own right.

3 Evaluating the Technical Quality of Concept Map Assessments

Intuitively, the use of concept maps to evaluate student declarative knowledge structure is appealing. A student's map directly reflects, to some degree, a student's understanding in a domain. Consensus needs to be reached not only about to what concept map assessments are but also about whether they provide reliable and valid measures of students' knowledge structure (Ruiz-Primo & Shavelson, 1996). That is, interpretation of assessment scores needs to be evaluated conceptually and empirically. In our work, these claims focus on both observed performance and cognitive aspects. We have conducted a series of studies to evaluate the information provided by concept maps. As mentioned, we have focused on the equivalence of different mapping techniques.

Most of the studies we have conducted involve repeated measures. Students are assessed across different mapping techniques and/or across the same mapping technique but with different samples of concepts. The former approach focuses on evaluating whether different mapping techniques provide a "similar" picture of students' connected understanding (a validity issue). The latter, examines concept-sampling variability (a reliability issue). We found that little attention has been paid to this latter issue (e.g., Ruiz-Primo & Shavelson, 1996). Hence, in our studies we have randomly sampled concepts whenever possible. We have conducted two types of empirical analyses, those based on quantitative analyses, and those based on cognitive analyses.

3.1 Empirical Quantitative Analyses: The Use of Generalizability Theory

Empirical quantitative analyses of concept map scores have been conducted within the context of Generalizability (G) theory. G theory recognizes that multiple sources of error contribute to the unreliability of a measure and hence to the estimate of student performance (e.g., Shavelson & Webb, 1991). The sampling framework we have used in our research defines and integrates the following facets: mapping techniques, raters, terms (i.e., concepts), and propositions. G theory has been used to evaluate the generalizability of students' average score map scores over mapping techniques, raters, concepts, and propositions. Different facets have been included in different studies; however, we have acknowledged that other facets, which we have not studied yet, can be included in the framework (e.g., occasions, method of response mode).

Results across all the studies using the construct-a-map technique suggest the following good news about concept map scores: (1) Students can be trained to construct concept maps in a short period of time with limited practice. (2) Raters do not introduce error variability into the scores. Concept maps can be reliably scored even when complex judgments such as quality of proposition are required (the interrater reliability on convergence score averaged across studies is .96). (3) Sampling variability from one random sample of concepts to another provides equivalent map scores when the concept domain is carefully specified. It is possible that the procedure we have followed in selecting the concept domain helped to create a list of cohesive concepts, therefore, any combination of concepts could provide critical information about student's knowledge about a topic. (4) The high magnitude of relative (.91) and absolute (.91) coefficients, averaged across types of scores and studies, suggest that concept maps scores can consistently rank students relative to one another and provide a good estimate of a student's level of performance, independently of how well their classmates performed. (5) The proportion of valid propositions in the students' map out of the possible propositions in a criterion map seems to better reflect systematic differences in students' connected understanding than other scores and it is the most effort and time efficient indicator. Other procedures have been carried out for supporting score interpretations (e.g., comparison between experts and novices scores).

3.2 Empirical Qualitative Analyses: Cognitive Validity

Cognitive validity studies help to link the intended assessment task demands, the cognitive activities evoked, and the student's performance level (Ruiz-Primo, Shavelson, Li, & Schultz, 2001). We have sought to bring empirical evidence to bear on the: (1) cognitive activities evoked by different mapping techniques, (2) relationship between the cognitive activities and performance scores, (3) impact of variation in assessment task on cognitive activities, and (4) correlations between assessment measuring similar and different constructs. The strategy is represented in Figure 4.

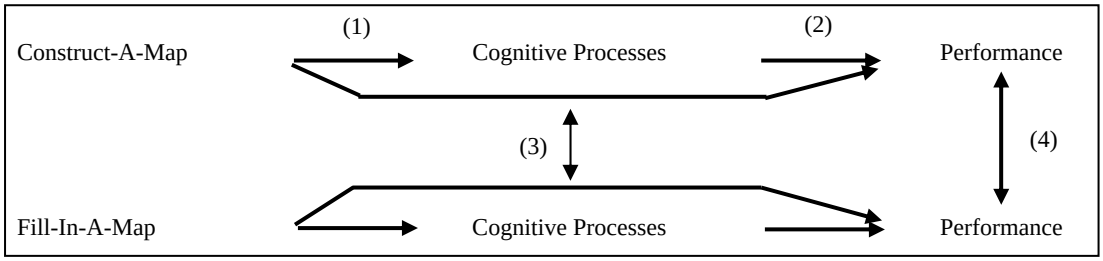

Figure 4. An approach to empirically evaluating cognitive interpretations of concept maps (Adapted from Ruiz-Primo, Shavelson et al., 2001).

We (Ruiz-Primo, Shavelson et al., 2001; Yin, et al., in press) have used talk-aloud protocols to capture experts' and novices' cognitive activities while concept mapping. We have briefly trained experts and novices to talk aloud while they are performing an unrelated task. Once they are comfortable talking aloud during the irrelevant task, we ask them to talk aloud as they are engaged in the concept-map assessment tasks while their verbalizations are recorded. If the student forgets to verbalize her or his thoughts, the researcher reminds her or him to talk aloud. In a recent study (Yin, et al., in press) we also video-taped participants' actions. The rationale behind this decision is that audio recoding may lead to ambiguous information. We have analyzed concept map talk aloud protocols using two approaches: In Approach 1, verbal protocols are transcribed, segmented, and coded. Codes follow from the construct definition and the logical analysis of the assessment demands. In Approach 2, we have examined the overall pattern of proposition generation.

In Approach 1, we (Ruiz-Primo, Shavelson et al., 2001) have distinguished between micro- and macro-levels of the protocol analysis. At the microlevel, we coded *explanations* (e.g., "N_2O_2 is a molecular compounds because they are both nonmetals," p. 115), *monitoring* ("I can't remember exactly what this is"; p. 115), *conceptual errors* reflecting misconceptions ("molecules are atoms"), and *inapplicable events* (e.g., reading instructions). At the macrolevel, we coded *planning* (verbalizations at the start of each protocol; e.g., "I will read the concepts and select the most important concept to put … in the center"; p. 116) and *strategy* (from entire protocol) for working through the mapping exercise. At the macrolevel, we used the entire verbal protocol evoked by a particular mapping technique. However, for the microlevel analysis, we segmented verbal protocols into fine-grained response entries (as small as phrases). On average, inter-coder reliability for the microlevel analysis was high across the different mapping techniques (construct-a-map, fill-in-the-nodes, and fill-in-the-lines). Both percent agreement and agreement adjusted for chance agreement (Kappa) were reported at the macrolevel with a range of 86 to 100 percent agreement and 71 to 100 percent adjusted.

In Approach 2, we have focused on two aspects of proposition generation: rate and procedures. *Proposition generation rate* refers to the speed with which students constructed propositions. *Proposition generation procedures* refer to the steps students take in proposition construction. We have attempted to infer how the nature of the assessment task influences students' cognitive processes by comparing their proposition generation rates across mapping techniques.

We (Ruiz-Primo, Shavelson et al., 2001) conducted an study to evaluate the interpretation of scores from three mapping techniques: (1) construct a map from scratch using concept provided, (2) fill-in-the-nodes in which students filled in a 12-blank-node skeleton map with concept provided, and (3) fill-in-the-lines, in which students filled a 12-blank-line skeleton map with linking phrases provided. To infer cognitive activities of three types of examinees (teachers, high proficient students, and low proficient students) we examined concurrent and retrospective verbalizations using Approach 1. We concluded that the three mapping techniques provided different pictures of students' connected understanding, and inferred cognitive activities across mapping techniques differed in relation to the directness of the task and the expertise of the examinees.

In another study, we (Yin at al., in press) examined the equivalence of another the two mapping techniques: Construct-a-Map (CMC) with own created linking phrases and Construct-a-Map with selected linking phrases (CMS). In this study, we used Approach 2 to infer cognitive activities. We used the students' think-aloud protocols to illustrate the cognitive procedures leading to the proposition generation rate differences. A comparison of the cognitive processes revealed that students in CMS spent extra effort and time matching what linking phrases they *wanted* to use with what they *could* use to construct their maps. Consequently, students in the CMS condition constructed their maps more slowly than students in the C condition. After all the evidence

was analyzed, we concluded that the CMC and CMS mapping techniques were not equivalent for the total accuracy score. The two concept-map task techniques elicited different student responses and representation of the students' declarative knowledge structure.

4 Conclusions

There is potential in using concept maps as assessment instruments, at least from the technical quality perspective. Nevertheless, there are still some issues that need to be resolved before we can conclude that they can reliably and validly evaluate students' connected understanding, especially if concept maps are to be used in high stake accountability contexts (cf. Lomask, Baron, Greig, & Harrison, 1992).

It is clear that we need to invest time and resources in finding out more about what aspects of students' knowledge are tapped by different forms of concept map assessments. Which technique(s) should be considered the most appropriate for large-scale assessment? Practical issues, though, cannot be the only criterion for selection. We have shown that the constraints and affordances imposed by different forms of assessments affect the student's performance. This means that different mapping techniques may lead to different conclusions about students' knowledge.

Many questions remain to be studied. For example, how large a sample of concepts is needed to measure a student's knowledge structure? How stable are concept maps scores? How exchangeable are concept mapping techniques that use different response modes (e.g., computer simulations versus paper-and-pencil). This research agenda is long, yet necessary, if we want to test in full the potential of concept maps as instruments for measuring different aspects of achievement in a particular domain.

5 References

Anderson, R. C. (1984). Some reflections on the acquisition of knowledge. *Educational Researcher, 13*(10), 5-10.

Chi, M.T.H., Glaser, R., & Farr, M.J. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlsbaum Associates, Inc.

Cronbach, L. J. (1990). *Essentials of psychological testing* (Fifth ed.). New York: Harper & Row Publishers.

Dochy, F. J. R. C. (1996). Assessment of domain-specific and domain-transcending prior knowledge: Entry assessment and the use of profile analysis. In M. Birenbaum & F. J. R. C. Dochy (Eds.) *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 93-129). Boston, MA: Kluwer Academic Publishers.

Duschl, R. A. (2003). Assessment of inquiry. In J. M. Atkin & J. E. Coffey (Eds.) *Everyday assessment in the science classroom* (pp.41-59). Washington DC. National Science Teachers Association Press.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*(1), 88-96.

Lomask, M., Baron, J. B., Greig, J. & Harrison, C. (1992, March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science.* Paper presented at the annual meeting of the National Association of Research in Science Teaching. Cambridge, MA.

Novak, J. D., & Gowin, D. R. (1984). *Learning how to learn*. New York: Cambridge Press.

Novak, J. D., Gowin, D. B., & Johansen, G. T. (1983). The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education, 67*(5), 625-645.

Pearsall, N.R., Skipper, J.E.J., & Mintzes, J.J. (1997). Knowledge restructuring in the life sciences. A longitudinal study of conceptual change in biology. *Science Education, 81*(2), 193-215.

Phillips, D.C. (1987). *Philosophy, science, and social inquiry. Contemporary methodological controversies in social science and related applied fields of research*. Oxford: Pergamon Press.

Rice, D.C., Ryan, J.M. & Samson, S.M. (1998). Using concept maps to assess student learning in the science classroom: Must different method compete? *Journal of Research in Science Teaching, 35*(10), 503-534.

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569-600.

Ruiz-Primo, M.A., Schultz, E. S., & Shavelson, R.J. (1996, April). *Concept map-based assessments in science: An exploratory study.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Ruiz-Primo, M.A., Schultz, E. S., & Shavelson, R.J. (1997, March). *On the validity of concept map-based assessment interpretations: An experiment testing the assumption of hierarchical concept maps in science.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38*(2), 260-278.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99-141.

Shavelson, R.J., & Ruiz-Primo, M.A. (1999). Leistungsbewertung im naturwissenschaftlichen Unterricht (On the assessment of science achievement). *Unterrichtswissenschaft. Zeitschrift für Lernforschung, 27* (2), 102-127.

Shavelson, R.J., & Ruiz-Primo, M.A., (2000). On the psychometrics of assessing science understanding. In J. Mintzes, J. Wandersee, J. Novak (Eds). *Assessing science understanding* (pp. 303-341). San Diego: Academic Press

Shavelson, R. J. & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. (2003). Reconsidering large-scale assessment to heighten its relevance to learnig. In J. M. Atkin & J. E. Coffey (Eds.) *Everyday assessment in the science classroom* (pp. 121-146). Washington DC. National Science Teachers Association Press.

White, R. T, & Gunstone, R. (1992). *Probing understanding*. New York: Falmer Press.

Yin, Y, Vanides, J, Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J. (in press) A comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching.*