Historical Thinking: In Search of Conceptual and Practical Guidance for the Design and Use of Assessments of Student Competence

Josh Radinsky[1], Susan R. Goldman[1,2], James W. Pellegrino[1,2]

University of Illinois, Chicago

Learning Sciences Research Institute

Department of Curriculum & Instruction[1] and Department of Psychology[2]

To appear in K. Ercikan & P. Seixas (Eds.), *Assessment of historical thinking.* New York, NY: Routledge

Acknowledgments

The preparation of this commentary was supported, in part, by Project READI, a multidisciplinary, multi-institution collaboration aimed at research and development to improve complex comprehension of multiple forms of text in literature, history and science. The authors thinking on matters of assessment of historical thinking have benefitted from discussions with their READI colleagues. Project READI is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100007 to University of Illinois at Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

50 Word Bios

Josh Radinsky

Josh Radinsky, PhD, is Associate Professor of Curriculum Studies and the Learning Sciences at the University of Illinois at Chicago. His research focuses on teaching and learning with graphic and geographic data, in history and the social sciences, and the construction of roles and narratives in classroom discourse.

Susan R. Goldman

Susan R. Goldman, PhD, is Distinguished Professor of Liberal Arts and Sciences, Psychology, and Education at the University of Illinois at Chicago and Co-director of UIC's Learning Sciences Research Institute. Her research and scholarly interests focus on learning and assessment in subject matter domains such as literacy, mathematics, history, and science.

James W. Pellegrino – James W. Pellegrino, PhD, is Distinguished Professor of Liberal Arts and Sciences, Psychology, and Education at the University of Illinois at Chicago and Co-director of UIC's *Learning Sciences Research Institute*. His research and development interests focus on children's and adult's thinking and learning and the implications for assessment and instructional practice

Historical Thinking: In Search of Conceptual and Practical Guidance for the Design and Use of Assessments of Student Competence

The four papers in this section all attempt to clarify the nature of historical thinking for purposes of assessing students' disciplinary competence. Each details the design of assessments and their interpretive use, spanning educational contexts from classroom practices to international assessments. These papers make valuable contributions to the field, highlighting many conceptual and practical considerations needed to advance the assessment of historical thinking.

Our discussion is divided into three parts. Part 1 presents three conceptual frames regarding the nature of assessment and assessment design, providing an interpretive language for discussing the four chapters. Part 2 applies these frames to the chapters as a way to interpret the specifics of each case. Part 3 highlights challenges that remain in conceptualizing and operationalizing the assessment of historical reasoning.

Part 1: Three Conceptual Frames

We suggest three conceptual frames for analyzing the contributions of the chapters in this section: (1) the "C-I-A"; (2) assessment as evidentiary reasoning; and (3) evidence-centered design.

The "C-I-A": Curriculum, Instruction and Assessment. Assessment does not and should not stand alone in the educational system. Rather, it is one of three coordinated components – curriculum, instruction, and assessment. *Curriculum* refers to knowledge and skills in subject matter areas that teachers teach and students are supposed to learn. It generally consists of a scope of content in a given subject area and a sequence for learning. *Instruction* refers to methods of teaching and the learning activities used to help students master the content

and objectives specified by a curriculum. *Assessment* is the means used to measure the outcomes of education and students' achievements with regard to important competencies. Assessment may include large-scale formal methods (e.g., state or national assessments), or less formal classroom-based procedures (e.g., quizzes, class projects, and teacher questioning). Ideally, an assessment should measure what students are actually being taught, and what is taught should parallel the curriculum one wants students to master. Aligning the three components is often a challenge; each chapter in this section addresses this challenge in a different way.

Assessment as Evidentiary Reasoning. Assessment enables educators to learn about what students know and can do, but cannot offer a direct window into a student's mind. An assessment is a tool designed to observe students' behavior, in order to produce data that can be used to draw reasonable inferences about what students know. In the process of generating and interpreting evidence to support inferences about what students know, all assessment procedures operate from a chain of reasoning about learning. This is true for classroom quizzes, standardized achievement tests, computerized tutoring programs, and even the conversation between student and teacher as they work through a problem together, or discuss the meaning of a historical text.

This process of reasoning from evidence has been portrayed as a triad of three interconnected elements: the *assessment triangle* (Pellegrino, Chudowsky, & Glaser, 2001). The vertices represent three key elements underlying any assessment: a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence. For effective and valid assessment, the three elements must be in synchrony.

The assessment triangle provides a useful framework for analyzing the underpinnings of assessments to determine how well they accomplish intended goals, for designing assessments, and for establishing their validity (e.g., Marion & Pellegrino, 2006). Each of the elements of the triangle must make sense on its own, and must connect meaningfully to each of the other two, to lead to an effective assessment and sound inferences. Central to this process are theoretically-grounded and empirically-supported understandings of how students learn, what students know as they develop competence, and how students' performances reflect these competencies. Such considerations are reflected differently in each chapter in this section.

Evidence-Centered Design. The design of an actual assessment is a challenging endeavor that must be guided by theory and research about cognition in context, as well as practical prescriptions regarding the processes that lead to productive and potentially valid assessments for particular contexts of use. Design is always a complex process that applies theory and research to achieve near-optimal solutions under multiple constraints, some of which are outside the realm of science. Assessment design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a program); the context in which it will be used (e.g., classroom, district or international-comparative); and practical constraints (e.g., resources and time).

Recognizing that assessment is an evidentiary reasoning process, it is useful to consider this process of creating assessments as *evidence-centered design* (Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006). The process starts by defining the claims that one wants to be able to make about student knowledge in a disciplinary domain, such as particular historical thinking abilities, epistemological stances, or recall of historical facts. It is critical that these claims about disciplinary-learning targets be specified as precisely as possible, using verbs that

afford assessment, rather than vague, high-level cognitive superordinate verbs such as "know" and "understand." Example verbs might include *compare, describe, analyze, elaborate, explain, predict,* or *justify*. Guiding this process of specifying claims is a body of theory and research on domain-specific knowing and learning.

Each claim about a student's mastery of some aspect of disciplinary thinking must also be linked (by warrants, or rules of interpretation) to the forms of evidence that would provide support for such a claim. The *evidence statements* associated with given claims capture the features of students' performances that would substantiate the claims. This includes which features need to be present, and how they are weighted – what matters most, and what matters least or not at all. The tasks need to allow students to "show what they know" in a way that is as unambiguous as possible. The precision that comes from elaborating these claim-evidence statements pays off when designing assessments, because it is clear what forms of evidence the task design needs to produce if it is to support the intended range of claims. These criteria determine the inferences about student cognition that are permissible and sustainable from a given set of assessment tasks or items. The chapters in this section differ in the ways they balance these criteria, and in their approaches to incorporating evidence into assessment design.

Part 2. Consideration of the Four Chapters

Our comments on the chapters reflect our analysis from the perspective of each of these three conceptual frames: C-I-A, Evidentiary Reasoning, and Evidence-Centered Design. Key questions are what we want students to know and be able to do as learners of history, and how the knowledge and skills develop over time with appropriate curriculum and instruction. Assessment serves the function of making explicit – to students and their teachers, among others

– the nature of what is expected of students, and the types of performances that are associated with the development of competence and expertise.

1. The C-I-A frame: Addressing the alignment of curriculum, instruction, and assessment

All four papers mention curricular concerns that should inform assessment in some way, and this in itself is an important contribution. Körber & Hamme point out the assessment challenges created by the lack of curricular alignment across districts and countries, leading to the need for assessments that are topically "self-contained." Similarly, Waldis et al. connect curricular concerns with assessment concerns in discussing possible implications of students' differential familiarity with specific historical topics for the assessment of particular competencies. They found that students tended to jump to the evaluative and short-circuit the evidentiary aspects of historical thinking for topics that were highly familiar to them (or "over-determined").

VanSledright more specifically frames such curricular concerns as issues of the ecological validity of assessments, and models an approach for developing ecologically-valid assessment items for a particular historical topic. VanSledright also links assessment to instruction, arguing for the importance of formative assessments that "can be used by both [students and teachers] to make more robust, daily decisions that enhance historical thinking and understanding" (p. 19). He highlights the importance of *opportunities to learn* the assessed constructs, a key issue in the C-I-A relationship (Gresalfi, 2009; Moss, Pullin, Gee, Haertel & Young, 2008).

Going further, Seixas et al. point to "the problems of any assessment research that is not tied to instructional practice" (p. 18). The findings of their study are used to suggest instructional

strategies that might help develop particular historical thinking practices (e.g., p. 17: "Systematic teaching of such documents as opportunities to enter into a foreign environment could set marginal students up for a higher level of performance"). Indeed, their program of work is couched in the assumption that assessments should inform instruction.

Taken together, these chapters suggest a continuum: from awareness of ways curriculum and instruction might impact the design of assessments, to more integrated considerations of *how* curriculum, instruction, and assessment might more productively inform one another in educational design and practice.

2. The evidentiary reasoning frame: Addressing the three components of the "assessment triangle."

The "assessment triangle" hinges on an explicit model of domain cognition, i.e., historical thinking. Each paper articulates such a domain model, laying out the forms of knowledge and the reasoning practices that define the domain of history for the proposed assessments. These models are used in each paper to specify the scope of what is assessed, including what is *not* in the current scope of the investigation.

Across the four papers, there are clear overlaps among the domain models; many of the constructs can be easily articulated to one another. At the same time, there appears to be a continental difference: the two chapters by North American researchers (VanSledright, and Seixas et al.) are grounded in models of historical thinking that are informed primarily by the work of North American researchers in a cognitive research tradition following Sam Wineburg, and by the North American academic-standards movement. In contrast, the two chapters by European researchers (Körber & Hamme, and Waldis et al.) are grounded more squarely in a

model of historical thinking that is based on Jörn Rüsen's foundational work on narrative competencies.

These traditions are certainly not incompatible, and Seixas' (2004) edited volume, *Theorizing Historical Consciousness* (with a chapter by Rüsen), provides a useful reference for articulating these domain models to one another. Still, it is notable that the chapters do not more thoroughly cite a common base of literature. A diversity of frameworks is certainly a desirable thing; however, careful articulation of these frameworks to one another is an essential step in moving the field forward. The international research and education communities interested in the assessment of historical thinking must draw on a coherent and consistent common base of theoretical and empirical work. Other parts of the world not represented in these four papers need to be integrated into this conversation as well, if the kind of international-comparative assessment described by Körber & Hamme is indeed a goal.

There are also differences across chapters as to what constitutes the domain in terms of epistemological knowledge, content knowledge, and the practices of historical analysis and reasoning. Such differences can have substantial consequences with respect to how the assessment development process unfolds. The more clarity there is about the elements of the domain analysis, the easier it is to specify the claims one wishes to make about what students are supposed to be able to do and the scope of those performances – the types of materials, the types of activity, the specific cognitive processes and socio-cultural practices – as well as the types of evidence that would support claims that students have the desired competencies under consideration. Again the diversity of models can be productive, but where there are overlaps, clarity will support better assessment design.

Each paper's domain model is mapped to student performances that could serve as evidence of key competencies, though this is done at very different levels of detail. VanSledright's model names a set of "procedural organizing concepts" (evidence, accounts, significance, context, and causation) and "strategic thinking capabilities" (careful reading, identifying and attributing accounts, perspective assessing, reliability judging, corroborating evidence) that work together to construct historical accounts. Körber & Hamme's "FUER" model consists of four competencies: devising historical questions; synthesis and analysis of historical statements ("re-constructive competence"); perspective-taking ("orientation competence"); and "Sachkompetenz," a broad constellation of concepts and categories that make up the domain of historical cognition, procedures, and epistemology. The model incorporates three levels of mastery for each of these four dimensions (a-conventional, conventional, and trans-conventional), constituting an ambitious framework to inform assessment design.

A key challenge for such global models is the alignment of each competency with the kinds of observations that can be used to assess them. Körber & Hamme focus on the assessment of "re-constructive competence," proposing three types of tasks for eliciting and evaluating different aspects. While the authors point to the need for an empirical base for validating such models and items, the process of interpreting actual student performances (the third leg of the triangle) remains unclear. The key question for an assessment item like their proposed story-construction task (p. 13) is how students' stories would be scored as evidence of particular constructs within the FUER model.

Waldis et al. map aspects of students' narrative performances to aspects of their domain model, which incorporates four "sub-operations." In the study, they correlate evidence of these sub-operations with a separate evaluation of quality of the students' narratives, broken down into

specific "qualities" and "quality features." While interesting and relevant, these correlations between quality features and sub-operations do not explicate the evidentiary reasoning for assessment. Interpretation of these findings needs a clearer articulation of the claims to be made based on students' narratives, including how each scoring process yields evidence in support of particular claims.

Seixas et al. use the "assessment triangle" explicitly, to attempt to specify this process of aligning student performances with aspects of their model. As this chapter shows, this specification is important for more than validation of assessments: it also serves to clarify and nuance the domain model itself. As noted on p. 17, "[i]n the analysis of students' responses to the test items, we gained new insights about confounding factors in the cognition model and, more so, in the task items themselves." Like VanSledright, examples of possible answers to each item are articulated to those aspects of the domain model that would (and would not) be made visible.

3. The evidence-centered design (ECD) frame: Addressing the purposes, contexts of use, and practical constraints shaping assessment design.

A unique and valuable contribution of these four chapters is the opportunity to gain insight into the process of designing historical-thinking assessments. The ECD framework highlights design decisions as valuable opportunities to make the evidentiary logic of an assessment clearly visible.

Several important design challenges are described in these chapters, and the authors leverage them differently to highlight key assessment issues. All four describe aspects of the envisioned contexts of use for each assessment that constrain and shape the design: the amount of time and effort required of teachers and students (Seixas et al., VanSledright, Waldis et al.);

considerations of historical topics likely to be covered in the curriculum (Körber & Hamme, VanSledright, Waldis et al.); and considerations of text difficulty in the selection of sources (Seixas et al., Waldis et al.).

At a finer level of detail, specifying the design details of the assessment tasks and items has particular value. Körber & Hamme's proposed structure for a narrative-construction task (their Figure 1) suggests a large number of design decisions that would need to be made in applying this template to the construction of rubrics for scoring students' work. Waldis et al. explore assessment possibilities of multiple genres of writing tasks (panel discussion, article and blog), and map examples of writing to competencies, but do not design a scoring scheme for making assessment decisions. VanSledright focuses on the design of *weighted multiple-choice items* (WMCs), specifying the rationales for particular types of answers and distractors, and the logic of the weighting of the answers in several examples. At this level of detail the deeper challenges of this work become clear: for example, the logic of VanSledright's specific weighting decisions invite debate and challenge.

Seixas et al. most closely approximate an ECD analysis, in that the focus of the design discussion is specifically on the value of the evidence that each item would produce. They point to important issues related to the interactivity of different test items with each other, the ways a test can build toward more complex evidence of mastery, and limitations of their approach. Still, the "architecture" of the 1-hour test described on p. 19 remains a fairly general design guide, and key issues for an ECD analysis remain to be articulated, including (most importantly) the design of detailed guides for the interpretation of students' work in light of the model.

Part 3. Challenges for Future Research and Development

Several tensions emerge in these chapters, as well as challenges for the field as the development of assessments of historical thinking continues to evolve.

Content Versus Historical Thinking Practices. The four chapters agree on the need to move away from conceptions of history as static bodies of dated events, and on the importance of a citizenry versed in historical thinking practices. Yet, all four recognize that engaging in (and assessing) historical thinking necessarily involves some historical content. The Waldis et al. findings suggest that familiarity with a historical topic may change the competencies that students make evident in an assessment. The resolution of exactly how to incorporate historical content into assessments depends, in part, on the purposes of the assessment and the claims for which evidence is sought.

At one end of the spectrum of purposes is VanSledright, who is explicit about the formative purpose of his assessment development work. These assessments target only content and practices that students have had opportunities to learn. At the other end of the spectrum, Körber & Hamm discuss the design of large scale assessments in which the fairness of the assessment across a wide range of students and curricula is of paramount importance. They advocate assessment of historical thinking using facts that are unfamiliar, in assessments that are "self-contained," with all the facts students need to know embedded in the assessment itself.

Somewhere between the two is classroom-based summative assessment, as reflected in the Seixas et al. chapter. Their content is drawn from historical periods that students had studied, but the assessment dives deeply into particular incidents. The issue of studied-versus-unknown content in assessments of historical thinking needs to be explicitly considered in interpreting the observed performances, to support claims about what students know and can do. These chapters offer useful models of clearly articulating different logics for this consideration.

Literacy Demands of the Assessments. The literacy demands of an assessment are a necessary set of considerations in the design of assessment tasks. This applies to production as well as comprehension. Have students been provided with opportunities to learn to write like historians, as well as to read like historians (cf. de La Paz, 2005; MonteSano, 2011; Reisman, 2012)? Using an ECD process, the literacy demands would be addressed in the design of the assessment task models (Mislevy & Riconscente, 2006).

Comprehension: Authentic historical documents versus adapted materials. Not unsurprisingly, adapting source texts has been a serious bone of contention among historians, history teachers, and assessment developers. On the one hand, students need to be able to access the content in order to reason with it. If documents are too complex or contain high proportions of vocabulary or syntax from bygone ages, students will simply not read the material and it will not be possible to obtain observations of their historical thinking. Accordingly, designers may justify the use of extracts or adapted materials that do the "translations" for the students. On the other hand, if students are never confronted with historical documents that they have to struggle to make sense of, they will develop neither strategies (cognitive and interpersonal) for dealing with complex and challenging traces of the past, nor the confidence to tackle them. Students' development as historical thinkers will be dependent on the presence of document translators. This issue is explicitly addressed only by Seixas et al. It is an area that needs a great deal more attention, especially regarding ways in which instructional supports can assist students in tackling challenging texts when reading like a historian (cf. Goldman, 2012; Goldman & Snow, in press; Reisman, 2012; Schoenbach, Greenleaf, & Murphy, 2012).

Production: Constructed responses versus Multiple Choice. The assessment tasks that were featured in the chapters varied from those with high production demands (e.g., essay

writing) to those with lower production demands (e.g., multiple choice items). Several issues rest on these choices. First, essays are difficult to score, especially if the assessment developers have no exemplar responses. Rubric development is time consuming, and reliability across scorers is often difficult to obtain without clear criteria. The Waldis et al. chapter presented the tip of the iceberg on this process, and their experience is typical (cf. De La Paz, 2005; Monte-Sano & De La Paz, 2012). The scoring issue is less complicated for short-answer constructed responses, but unless the criteria for reasoning are clear, reliability issues are just as problematic. Also, short answer responses may not afford students rich opportunities to demonstrate historical reasoning. Students are also often able to demonstrate more sophisticated reasoning orally than in written form. Thus, the literacy demands of written production may mask the historical thinking and reasoning that students can do.

In contrast to the high-output demands of constructed responses, multiple-choice items do not require students to produce language. They do, however, require comprehension – at least for the more well-constructed multiple-choice items like those of Seixas et al. and VanSledright. Even here, the comprehension demands of items with lengthy alternatives can sometimes obfuscate the historical thinking the item was intended to assess. We find VanSledright's approach to weighted alternatives a provocative one. However, we suspect that many readers, ourselves included, could argue with some of the weighting decisions.

What seems called for in efforts to develop multiple-choice alternatives are the rigorous kinds of evaluation techniques that VanSledright, Körber & Hamme, and Seixas et al. discuss. These need to be conducted with individuals who reflect different levels in the development of expertise with respect to historical thinking. We speculate that what more knowledgeable students of history see as clear distinctions among various alternatives are much more difficult to

see for those less knowledgeable. Thus, there is a strong need to validate assessments on the populations for whom they are intended.

Developmental Issues and Learning Progressions in Historical Thinking. The question of whether alternative tasks or items tap the "same" historical thinking competencies across a wide range of individuals raises a larger set of issues regarding the development of historical thinking practices, and how they are introduced and then deepened through successive learning experiences. We know from the literature on expertise (e.g., Ericsson, 2006) that hundreds to thousands of hours of practice are needed to achieve expert status. This implies the need to coordinate the teaching of historical thinking within and across grades, with an eye toward making visible the processes of interpretation of the historical record and construction of historical narratives. To date, we know of little work that has attempted to examine this issue or trace out such progressions (but see Goldman, et al., 2009*)*.

In the context of the assessment triangle and the ECD process, this would mean unpacking the domain model into student models that are developmentally appropriate for a given age range of students. For example, while we might want 7 year olds to be aware that different people have different perspectives on events, the cognitive model would differ for the performance we expect from a 17 year old. The four chapters in this section do not address the issue of such progressions, nor the developmental level of the participants in their studies. This is an important direction for this work going forward.

Assessment development in any disciplinary domain is a challenging endeavor. We applaud the efforts of these authors to specify the domain of historical thinking for purposes of assessment, and their articulation of design models and cases. While they share a common goal, it is clear that there is far from a consensus view of the domain and how to assess student

competence. These cases are instructive, and offer the opportunity for further dialogue about how to meet the conceptual and practical challenges in the assessment of historical thinking.

References

De La Paz, S. (2005). Effects of Historical Reasoning Instruction and Writing Strategy Mastery in Culturally and Academically Diverse Middle School Classrooms. *Journal of Educational Psychology, 97*, 139.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, and R. R. Hoffman, R. R. (Eds.). *Cambridge handbook of expertise and expert performance* (pp. 685-706). Cambridge, UK: Cambridge University Press.

Goldman, S. R., et al. (2009). *Reading for understanding across grades 6-12: Evidence based argumentation for disciplinary learning.* Funded July, 2010 by Institute for Educational Sciences, U.S. Department of Education, Grant # R305F100007.

Goldman, S. R. (2012). Adolescent Literacy: Learning and Understanding Content. *Future of Children, 22*, no. 2, 89 – 116.

Goldman, S. R., & Snow, C. (in press). Adolescent literacy: Development and instruction. To appear in A. Pollatsek & R. Treiman (Eds.), *The Oxford Handbook of Reading*. New York, NY: Oxford University Press.

Gresalfi, M. S. (2009). Taking up opportunities to learn: Constructing dispositions in mathematics classrooms. *The Journal of the Learning Sciences*, *18*(3), 327-369.

Körber, A., & Hamme, J. M. "Historical Thinking, Competencies and their Measurement: Challenges and Approaches," this volume

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, *25*(4), 47-57.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Monte-Sano, C., & De La Paz, S. (2012). Using Writing Tasks to Elicit Adolescents' Historical Reasoning. *Journal of Literacy Research, 44*, 273-299.

Moss, P., Pullin, D., Gee, J., Haertel, E., & Young, L. (Eds.). (2008). *Assessment, equity, and opportunity to learn.* Cambridge, England: Cambridge University Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction, 30*, 86-112.

Rüsen, J. (2004). Historical consciousness: narrative structure, moral function, and ontogenetic development. In P. Seixas (Ed.), *Theorizing historical consciousness* (pp. 63-85). Toronto, Canada: University of Toronto Press.

Schoenbach, R., Greenleaf, C., & Murphy, L. (2012). *Reading for understanding: How Reading Apprenticeship improves disciplinary learning in secondary and college classrooms*. John Wiley & Sons.

Seixas, P., Gibson, L., & Ercikan, K. "A Design Process for Assessing Historical Thinking: The case of a one-hour test," this volume.

VanSledright, B. "Assessing for Learning in the History Classroom," this volume.

Waldis, M., Hodel, J., Thünemann,, H., Zülsdorf-Kersting, M., & Ziegler, B. “Material-based and Open-ended Writing Tasks for Assessing Narrative Competence Among Students,” this volume.