Development and validation of a practical classroom assessment of students' beliefs about differing historical accounts

D. Kevin O'Neill
Sheryl Guloy
Fiona M. MacKellar
Dale R. Martelli
Simon Fraser University, Canada

ABSTRACT

History teachers in multicultural societies are increasingly responsible for facilitating students' awareness of and understanding of multiple accounts of the same, or related past events. The primary goal of the Historical Account Differences questionnaire is to help history teachers assess their own students' beliefs about why accounts can differ, and the effectiveness of lessons and units aimed at developing students' epistemological conceptions about such accounts. The theoretical underpinnings, design and validation of the questionnaire are discussed. As part of the validation, responses were provided by 899 Canadian students from 8th grade through postsecondary studies. Findings failed to support the hypothesis of strict stage-like progression in students' conceptions claimed by the developmental theory on which the instrument was based. However, other claims implicit in the theory were supported. Theoretical and practical implications are discussed.

KEYWORDS

Classroom assessment, epistemological conceptions, historical accounts, history teaching, developmental theories, dimensional theories

CITATION

COPYRIGHT

Introduction

> The past makes itself accessible to us in the present via residue in the form of accounts. (VanSledright 2015, p. 81)

Historical accounts often have a central place in history teachers' practice. Narrative is a central way of understanding human experience in time (Cronon, 1992); but helping students to understand the nature of historical accounts has always been challenging work. This task is particularly challenging in multicultural societies today, which arguably have multiple pasts (Conrad, Ercikan, Friesen, Létourneau, Muise, Northrup, & Seixas, 2013). As societies have better recognized their diversities, educators have faced growing demands to facilitate their students' awareness and understanding of multiple accounts of the same, or related past events (Banks, 2008; Seixas, 2004; Takaki, 1993).

Textbooks have traditionally offered a single, univocal narrative in the impersonal "voice of history" (Wineburg, 1991); but even there, sidebars today invite students to consider different, related stories – to say nothing of accounts from popular media and other sources. For example in Canada, high school students read textbooks that tell the story of their history as beginning when Europeans boldly set forth to tame a "new" land, heroically surmounting many hardships, including resistance from Indigenous peoples. At the same time, there is growing recognition (including in official federal government reports) that these "heroes" were systematically eradicating culturally advanced Indigenous nations that pre-dated European colonization by thousands of years, and repeatedly violated agreements they made with them. Such differing accounts bring urgency to the question: How well are students equipped to appreciate differing historical accounts, and their increasing prevalence in history teaching and society at large?

In practice, openness to the multiplicity of historical accounts (or what Chapman (2017) alternatively terms "interpretations") carries risks which educators must manage in an informed manner. Without a mature understanding of why various carefully-researched historical accounts might differ, there is a chance that students will respond to them by hardening to either a cynical view of the accounts they are presented with (e.g. "history is written by the victors") or an "anything goes" conception according to which all accounts are equally valid in principle, and everything is a matter of perspective (Lee & Shemilt, 2004; Maggioni, VanSledright, & Alexander, 2009; Chapman, 2017). Such conceptions not only render the discipline of history meaningless for students, but are inconsistent with democratic participation in a pluralistic society, which many educators make it their aim to enable (Castro, Field, Bauml, & Morowski, 2012; Miller & Toth, 2012).

How can teachers refine curricula and teaching strategies that will lead students toward mature ideas about why and how historians have produced differing accounts of the past? We believe that any complete response to this question must include tools of assessment that are practical for teachers to use, without the support of researchers. Without knowing what beliefs a specific classroom full of students already holds about differing historical accounts, teachers are essentially left guessing what lessons and activities might be effective for them. In the work described below, we pursued the design of an assessment that would, to borrow VanSledright's words, "assist teachers in making the pedagogical adjustments necessary to grow possibilities for understanding" of history (VanSledright 2015, p. 80).

The Historical Account Differences (HAD) questionnaire is designed for use with students in both secondary and postsecondary classrooms. Here we discuss the theoretical underpinnings of the questionnaire, explain how it was developed, and present analyses of 899 responses provided by students spanning Grade 8 through postsecondary studies, which were gathered as part of our validation effort.

Epistemological Conceptions in History

In the North American context, the scholarship of history learning has focused for decades on the development of historical thinking skills (Lightning, 2021). So, to situate this work it is important to underscore the distinction between *historical thinking* and *domain-specific epistemological conceptions in history*. Historical thinking refers to a complex and theoretically-contested set of cognitive processes. One example is *sourcing* – the process of considering the origins of a source when interpreting it (Britt, Perfetti, Van Dyke & Gabrys, 2000), such as who wrote it and for what purpose. Another important historical thinking skill is *contextualization,* which involves considering the time period and circumstances surrounding an event when interpreting evidence about it (Mitrovic, 2015).

While undoubtedly important, historical thinking skills have proven challenging to capture in assessments, for a variety of reasons (Ercikan & Seixas, 2015). For example, since authentic historical thinking tasks often involve students reading and interpreting difficult texts, historical thinking assessments may be unable to avoid conflating students' facility in historical thinking with their general reading skill (Reisman, 2015). Further, such assessments can become so time-intensive to administer and score that it becomes impractical to use them very often in the classroom context.

History-specific epistemological conceptions, on the other hand, include an individual or group's beliefs about the discipline of history and the nature of historical knowledge at a given moment in time. They are distinct from historical thinking skills, but are equally important because they shape how those skills are used (Chapman, 2017; Wiley, Griffin, Steffens, & Britt, 2020). For instance, in order for a student to use a set of varied historical accounts to make a judgement about the significance of a particular battle in World War I, she needs *both* appropriate skills in contextualization and sourcing, *and* beliefs about why accounts can differ that will promote the relevant application of those skills. There is good evidence that students' epistemological conceptions shape their performance on historical reasoning tasks, and that some conceptions are more "activity appropriate" than others when engaging with multiple texts (Wiley et al., 2020, p. 14).

Thus, history students' domain-specific epistemological conceptions are of practical concern for teachers of history. In many instances, attempts to involve students in historical thinking may not be fruitful without first addressing their epistemological conceptions. For example, document-based questions (DBQs) (Rothschild, 2000) can be fruitful skill-building opportunities for students who already hold sufficiently mature conceptions about the nature of historical evidence; but if a student approaches a DBQ with the naïve belief that every account must be either unbiased and true, or biased and untrue, he is not likely to have a fruitful encounter (Chapman, 2017). Indeed, he may simply be confused as to why his teacher has chosen to mislead him with competing accounts, some of which must be untrue by definition. Such responses from students have been documented (Lee, 2004, 2005; Lee & Shemilt, 2004; Seixas, 2006; Wineburg, 2001).

In recent work, Wiley et al. (2020) developed an instrument aimed specifically at gauging students' ideas regarding the value of integrating evidence across multiple documents in history, which proved predictive of student performance on document-based questions. The agenda we pursue here can be seen as related. We chose to focus our development effort specifically on students' beliefs about why historical accounts may differ. We chose narrative accounts as the focus for our work because they are focal to the potentially urgent issue in many instructors' practice alluded to earlier – the differing accounts that can be at issue in modern, multicultural classrooms. Further, as we have suggested, students' beliefs about why accounts differ may strongly influence the effectiveness of teachers' attempts to teach historical thinking. As Lee & Shemilt (2004) wrote,

> Tackling historians' reasons for writing one story rather than another before students grasp that accounts cannot be copies of the past is a recipe for leaving them helplessly shrugging their shoulders in the face of competing stories. (p. 31)

The central goal of the work described below has been to provide a time-efficient, logistically simple and pedagogically useful way for history instructors to assess the beliefs that their *own* students hold at a *particular* time about why historical accounts may differ. As Smith and Breakstone (2015) argued,

> [L]engthy essays that students compose…make it difficult for teachers to quickly identify particular skills or concepts that need further attention. Teachers with classes of more than 30 students have to wade through hundreds of pages of student writing to determine next steps for teaching. It is unrealistic to expect teachers to frequently assign DBQs to all of their students or to use them to make quick adjustments to classroom instruction (Smith and Breakstone 2015, p. 234).

Given these practicalities, our goal was to develop a zero-stakes assessment tool that instructors would be able to use as often as they needed to, without researchers' help. Ideally the assessment would be understandable to students as young as 13, would be quickly completed, and easy to score.

The challenge of this agenda should not be underestimated, because as Maggioni et al. (2009) wrote, "in trying to model individuals' epistemic beliefs, we are attempting to categorize commitments that people, for the most part, may perceive as fuzzy and somewhat vague" (p. 189). Understandably, there is a diversity of theoretical approaches to the problem of assessing epistemological conceptions. Theories of personal epistemology presented in the literature have differed in two broad ways: they have been either domain-general or domain-specific, and they have been either developmental or dimensional in nature. Theories are described as domain-general if they attempt to make claims that hold true across disciplines, rather than confining themselves to a single one. Theories are developmental in nature if they predict that epistemological conceptions will change predictably in a series of discrete, ordered stages along a fixed trajectory, while dimensional theories model epistemological conceptions using multiple dimensions that can vary independently.

In early work, Perry (1970) presented a domain-general developmental theory of personal epistemology that described university undergraduates as progressing in a series of nine ordered stages from what he termed "dualist thinking" to "commitment." The theory is domain-general in the sense that it does not pertain to students in any one discipline, and developmental in that it predicts advancement along a fixed trajectory of ordered stages. Shemilt (1987) also presented a developmental theory, though in contrast his was domain-specific and addressed students at a different level of schooling. This theory described adolescent history students progressing through a series of four ordered stages relating to their ideas about evidence and methodology in history. More will be said of this theory later.

Over time, skepticism has grown regarding the ability of stage theories to adequately represent the complexity of students' epistemological conceptions and their transformation over time (Chandler, Hallett & Sokol, 2002). Empirical findings have problematized the assumption of strict stage-like progression, demonstrating inconsistency in learners' conceptions. This is sometimes described as "epistemic wobbling" (VanSledright and Reddy, 2014, p. 43), and may in part be an artifact of the instruments that researchers have developed and used.

Partly in reaction to such findings, models of epistemic beliefs have been developed that represent them not on a single scale, but using a number of independent dimensions. For example, Schommer-Aikins' (2004) Epistemological Belief System model comprises five dimensions relating to people's beliefs about the certainty of knowledge, the simplicity of knowledge, omniscient authority, the extent to which learning should be quick, and whether intellectual

ability is innate. Such dimensional models do not assume the existence of an apex of development or a series of ordered stages; thus they are less subject to the challenge of "wobbling."

The issue of domain generality or specificity in epistemological beliefs was explored in depth by Muis, Bendixen and Harle (2006). They noted that, "research on problem solving and critical thinking supports the notion that expert knowledge is predominantly domain specific" (p. 4). Thus, it seems reasonable to posit that epistemological beliefs are also domain specific in important ways. In their extensive review of literature, Muis et al. (2006) examined 19 empirical studies designed to examine the domain specificity or generality of epistemological beliefs, noting that most "found predominant support for domain-specific beliefs across varying levels of education." (p. 24) For the purposes of analysis, the researchers involved in these studies assessed epistemological beliefs using instruments based on a consistent set of dimensions; but outcomes often showed that the domain asked about (e.g. science versus history) significantly altered the beliefs that students reported.

To conclude this section we would like to note that much of the recent research on epistemological conceptions in history has taken the approach of adopting a domain-general theoretical model of epistemological beliefs, and then localizing it to the domain of history, so to speak, when designing instruments. We elaborate on this observation in the following section.

Quantitative Measures of Epistemological Conceptions in History

While there is a longstanding body of work on quantitative measures of epistemological beliefs (e.g. Schommer-Aikins, 2004), studies of epistemological beliefs in history have been conducted using purely qualitative methods until relatively recently. The first scholars to publish a quantitative instrument to gauge epistemological beliefs in history appear to have been Maggioni, VanSledright, & Alexander (2009). They developed a questionnaire called the Beliefs about Learning and Teaching of History Questionnaire (BLTHQ), using theoretical underpinnings that were domain-general and developmental in nature. This 21-item questionnaire was designed specifically for use with K-12 teachers, to evaluate the effectiveness of professional development in the domain of history. Using the questionnaire, Maggioni et al. (2009) were able to document changes in the epistemological beliefs of 72 fourth- and fifth-grade student teachers over the course of a summertime professional development opportunity.

Items from the BLTHQ were adapted by Nitsche and Waldis (2018) for their own instrument, the Epistemological Beliefs Questionnaire in History (EBHQ). Using this instrument they investigated the relationships among history teachers' epistemological beliefs in history, their pedagogical beliefs, and several other internal and external variables such as age, sex, level of education, socioeconomic status and teaching experience. Analyzing responses from 177 pre-service teachers in German-speaking Switzerland, they found primarily small effects. For example, it was found that the number of history courses a student teacher had completed had a small but signficant relationship to narrative constructionist epistemological beliefs in the domain of history. Similarly small effects were observed with respect to age, sex and socio-economic status (SES), with increasing age and higher SES predicting skeptical beliefs. Intriguingly, the researchers also found that a larger number of courses in history education taken by a pre-service teacher predicted transmissionist perspectives on teaching and learning, while a larger number of semesters of study overall predicted less transmissionist views (Nitsche & Waldis, 2018).

Stoel, Logtenberg, Wansink, Huijgen, van Boxtel, & van Drie (2017) extensively modified Maggioni's instrument in an effort to develop one that could be used with school-aged students. Their 26-item instrument was designed to distinguish between naive and nuanced views with regard to the nature of historical knowledge and historical knowing. Using responses from 922 students enrolled in one of the two highest educational tracks in the Netherlands, they were able to develop a 3-factor scoring scheme that allowed them to assess differences between school tracks, and relationships between epistemological beliefs, interest in history, and history grades. However, some of their findings were confusing considering the developmental theoretical basis

of their instrument. They reported being "puzzled...that students could simultaneously report a positive score for both the naïve and the nuanced nature-of-knowing items." (p. 131) This finding is reminiscent of the "wobbling" mentioned above.

Finally, Wiley et al. (2020) developed and validated a 12-item measure of epistemological beliefs that was specifically tailored to measure beliefs relevant to integrating knowledge across documents in the domain of history. This work was a departure from previous research on epistemic cognition in history, in that it did not adopt wholesale a theory of personal epistemology and then attempt to capture all of its facets in the research instrument. Instead, the researchers adapted items from previously-published research that gauged beliefs which prior research suggested were specifically relevant to students' work on document-based questions. Over a series of three experiments conducted with hundreds of university undergraduates, high school students and middle school students, Wiley et al. (2020) demonstrated that "activity appropriate" (p. 14) beliefs contributed unique value in predicting students' performance on document-based questions, over and above students' academic track (AP or non-AP) and level of schooling (high school or middle school). This task-specific approach to measuring epistemological beliefs bears similarity to what we have done in the work reported below.

Developing the Historical Account Differences Questionnaire

In this section we describe the theoretical basis for HAD and the methods we used in developing and validating the questionnaire.

Theoretical Underpinnings

Counter to the approach taken in much of the prior research, we opted for a domain-specific theoretical foundation for HAD. Due to the research on expertise referred to by Muis et al. (2006), we speculated that a domain-specific foundation was a potentially fruitful route to an instrument that could provide the most pedagogically-useful information to instructors with the minimum number of items and demand on instructional time. The theoretical basis of our instrument came from Shemilt (1987).

As part of the Schools Councils History project in the United Kingdom, Shemilt conducted many hours of interviews with British adolescents in order to understand the development of their ideas about the discipline of history (Shemilt, 1987, 2000). Based on these cross-sectional interviews, Shemilt then formulated a 4-stage theory that speculatively described the development of students' understanding of historical evidence, methodology and accounts in a series of four ordered stages.

At Stage 1 of Shemilt's model, the lowest level of understanding, students are described as taking knowledge of the past as given. The only difficulty they associate with history is the difficulty of reading or listening to stories and remembering them. By extension, they think of historians as no more than good memorizers of stories about the past.

Students attain a Stage 2 understanding when they realize that the past does not speak with a single voice: knowledgeable people do disagree in their accounts of the past. However, the only explanations a Stage 2 thinker can muster for this disagreement are that the evidence may not be reliable (e.g., some reporters witnessed the events in question while others did not) or that some reporters may be biased. At this stage, students think of historians as people who are somehow able to sniff out false or biased stories about the past, "read off" the truth from sources, and piece together the one true account.

Students at Stage 3 are distinguished by the understanding that historical knowledge can never be absolutely certain. At best, we can use the evidence available about the past (some of which is other peoples' stories, but which also includes non-narrative relics such as tax records and maps) to reduce the uncertainty of our knowledge. In line with this view, Stage 3 thinkers understand historical scholarship to involve reasoning methodically with evidence to come up with an

account that represents the most likely and believable *reconstruction* of events. Stage 3 thinkers understand historians as people who know how to do this methodical work.

In the fourth and most advanced stage of Shemilt's model, students are described as viewing historical knowledge like a kaleidoscope. There is no single true account, or even necessarily a single most likely account. Rather, students recognize that it is possible to have several equally defensible accounts of the past – particularly if they have been constructed to address different questions. In this view, a historical account is understood to be a creature of its time, constructed to address a need. It is constrained by the available evidence, but is also shaped by the questions it seeks to answer. Like a kaleidoscope, history's "patterns are ordered and determinate, but do not yield a single stable picture" (Shemilt, 2000).

Shemilt's 1987 model was elaborated to a six-stage model in later publications (Lee & Shemilt 2004); yet we chose the 4-stage model for its parsimony, which is particularly important for a practical classroom instrument. An instrument based on a 6-stage model would clearly need to involve a larger number of items to demonstrate reliability, and would thus take longer to complete.

Note that while our instrument built on the ideas in Shemilt's 1987 developmental model, we did not assume that the epistemological conceptions of our secondary and postsecondary participants in British Columbia would progress in the particular way described by Shemilt for British adolescents in the 1970s. Our research was planned to provide opportunities to empirically test, on a substantial scale, whether students' conceptions progressed in the stagelike way predicted in Shemilt's theory. To our knowledge, such a test has never been attempted.

Question Structure

While we have published about previous versions of the Historical Account Differences questionnaire (O'Neill, Guloy and Sensoy, 2014), the version examined here (Appendix A) was more sophisticated than those used earlier. It is composed of six questions, each of which was inspired by Shemilt's original interviews, and designed to provide a context for examining students' ideas about differing accounts:

- What makes somebody a historian?
- How do historians develop their stories about events in the past?
- Why do historians write or tell stories about the past?
- Why do historians write new books about events that were already written about before?
- What makes a story about the past useful?
- If a historian is learning about the events of a period and finds two stories about them that disagree, what should she do?

Students respond to each question by rating a set of four belief statements on a 5-point scale from "strongly agree" to "strongly disagree." Each statement is written to reflect the beliefs that Shemilt (1987) hypothesized students to hold at each of the four stages. For example, the question "What makes somebody a historian?" asks students to rate the following four statements:

- They have studied and memorized what happened in the past *(Stage 1)*
- They can figure out which stories about the past are biased or untrue, and put together the one true story *(Stage 2)*
- They can use evidence to figure out what probably happened in the past *(Stage 3)*
- They re-consider past events, and think about what those events mean to us today *(Stage 4)*

Here, the four statements are presented in stage order for ease of reading. In the questionnaire presented to students, the four statements were presented in a different, random order for each question, without the stage numbers appended.

The approach of framing rating scales with questions is not common in the literature, though it bears some similarity to the scenario-based approach of Barzilai & Weinstock (2015) in their Epistemic Thinking Assessment instrument. In that work it was assumed that brief scenarios could enhance the meaning of abstract scales and make them easier for students to respond to. In HAD, the six questions help give meaning to otherwise abstract scales.

Respondents are free to rate each of the statements as they choose. Thus, in responding to a question a student could hypothetically choose to strongly disagree with all 4 statements, or strongly agree with all 4 statements – though in practice they rarely do.

Methods

Construct Validity

To ensure that each of the four statements specifically targeted one stage from Shemilt's theory, we recruited four history professors and one historical archaeologist to perform a sorting validation task. The five professors, who represented varying specialties including Canadian history, Middle East history, Latin American history and U.S. history, had not collaborated with the research team previously.

Two rounds of hour-long, one-on-one interviews were carried out with each professor between August 2012 and October 2012. During each round, a researcher explained that the purpose of the interview was to check the language of the questions and belief statements for their correspondence with Shemilt's theory. The professor was then asked to review Shemilt's 4-stage theory, using a one-page description very similar to the summary provided earlier in this paper. Interviewees had no difficulty understanding or distinguishing the four stages, but kept the one-page description for reference while they were presented with the draft statements for the questionnaire, one by one, in random order.

One at a time, items from the draft questionnaire were presented, with the instruction, "If a student strongly agreed with this statement, which stage would he or she be at?" The professor then sorted the item into one of four piles corresponding to Shemilt's hypothesized stages. After all the draft statements had been sorted, the interviewer recorded the professor's sorting of the cards. When the professor's sorting of a card was observed to disagree with the stage intended by the questionnaire authors, he or she was asked to recommend improvements to the formulation of the item.

Improvements suggested by the panel of professors in round 1 of the interviews were tested in round 2 using the identical procedure. Cohen's Kappa was then calculated for each pile of cards sorted by each participant, comparing the intended stages of the statements to the stages they were sorted into.

Known-groups Validity

The questionnaire was made available on a web site hosted at our home institution. In collaboration with teachers and professors of history across the greater Vancouver area of British Columbia, Canada, we solicited responses to the HAD questionnaire from a total of 899 students between September 2013 and May 2014: 566 secondary school students and 342 postsecondary students. The secondary school students spanned grades 8 to 12, with approximately 100 participants per grade (see Table 1).

TABLE 1: Makeup of the secondary students by grade level

Grade	8	9	10	11	12	Total
# Responses	138	129	88	85	118	558

Secondary school students were invited to complete the questionnaire during the school day in a computer lab. Completion of the questionnaire required a unique participant ID code that was provided to each student by a researcher only after parental consent for the research had been verified. One or two researchers were present during each administration of the questionnaire to verify parental consent, and provide any technical support that might be required. In the course of data collection, it was observed that nearly all students could complete the HAD questionnaire in 15 minutes or less.

At the postsecondary level, student participants were recruited in class (usually through a brief announcement in a lecture), but completed the questionnaire on their own time. Postsecondary students provided research permission electronically before proceeding to the questionnaire. In addition to the questions described earlier that formed the core of the HAD questionnaire, postsecondary students answered questions about their involvement with history as a discipline, e.g. as a minor or major.

The postsecondary participants had widely varying exposure to history as a discipline (see Table 2). While the majority of the postsecondary students were recruited from 100- and 200-level history courses at three different institutions, the sample also included 31 students enrolled in postsecondary engineering courses requiring no background in history beyond high school, and a handful of students pursuing masters and doctoral degrees in history.

TABLE 2: MAKEUP OF THE POSTSECONDARY STUDENTS BY INVOLVEMENT WITH HISTORY

Involvement in history as a discipline	Not taking history	History elective course	Completing history minor	Completing history major	Completing history masters	Completing history PhD	Total
# Responses	22	180	38	68	5	9	322

We conducted analyses to test four hypotheses based on Shemilt's theory:

- H1: Postsecondary students' scores will reflect more sophisticated beliefs than those of secondary students
- H2: Secondary students' scores will reflect more sophisticated beliefs as they progress through the grades
- H3: Among postsecondary students, those with greater involvement in history as a discipline will demonstrate more sophisticated beliefs
- H4: Students will not at the same time hold beliefs that Shemilt's theory describes as naïve and sophisticated

Data Analysis and Results

Construct Validity

At the end of round 2 of the sorting task, Cohen's Kappa was calculated for each professor's sorting of the questionnaire items. In this step, the history professors' categorization of the statements was compared to the stages the research team had intended these items to capture. Values of Kappa ranged from a high of .78 for one professor (83% agreement with intended stage) to a low of .61 (71% agreement with intended stage). The average value of Kappa across the five professors was .71 (78% agreement). Given that HAD is not intended as a high-stakes assessment, these levels of reliability were judged acceptable. Further small refinements suggested by professors in round 2 were made before the data collection for known-groups validity testing was carried out.

Known-Groups Validity

Students' responses were scored on two scales. The "Knowledge and Reliability" (K&R) scale reflects the expressed strength of a student's agreement that historical accounts only differ due to the the fallibility of some storytellers due to ignorance or bias. In Shemilt's theory, these are considered naïve ideas. The K&R score is calculated by summing each student's ratings of the items developed to address Stage 1 & 2 conceptions. The "Reconstruction and Purpose" (R&P) scale reflects the expressed strength of a student's agreement that historical accounts differ due to differences in the interpretation of limited evidence or differences in storytellers' purposes. These Stage 3 & 4 beliefs are considered sophisticated in Shemilt's developmental theory. The R&P scale is calculated by summing each student's ratings of the items developed to address Stage 3 & 4 conceptions. Chronbach's alpha was calculated at .760 for the K&R scale, and .669 for the R&P scale.

Note that since the instrument was intended for use by classroom teachers, we opted to compute scale scores as sums rather than means, to keep the scoring simple. Since each scale includes 12 of the 24 items, and students' responses to each statement take the form of a rating from 1 to 5 (1= strongly disagree to 5=strongly agree), the range of possible scores for each of the two scales lies between 12 and 60.

Using scores calculated for the two scales, analyses were carried out to test the four hypotheses stated in the previous section. Since the distribution of students' scores on both scales were found to be normal (see Figures 1 and 2), parametric tests were used.

Figure 1: COMPARISON BETWEEN SECONDARY AND POSTSECONDARY STUDENTS' SCORES ON THE K&R SCALE ITEMS
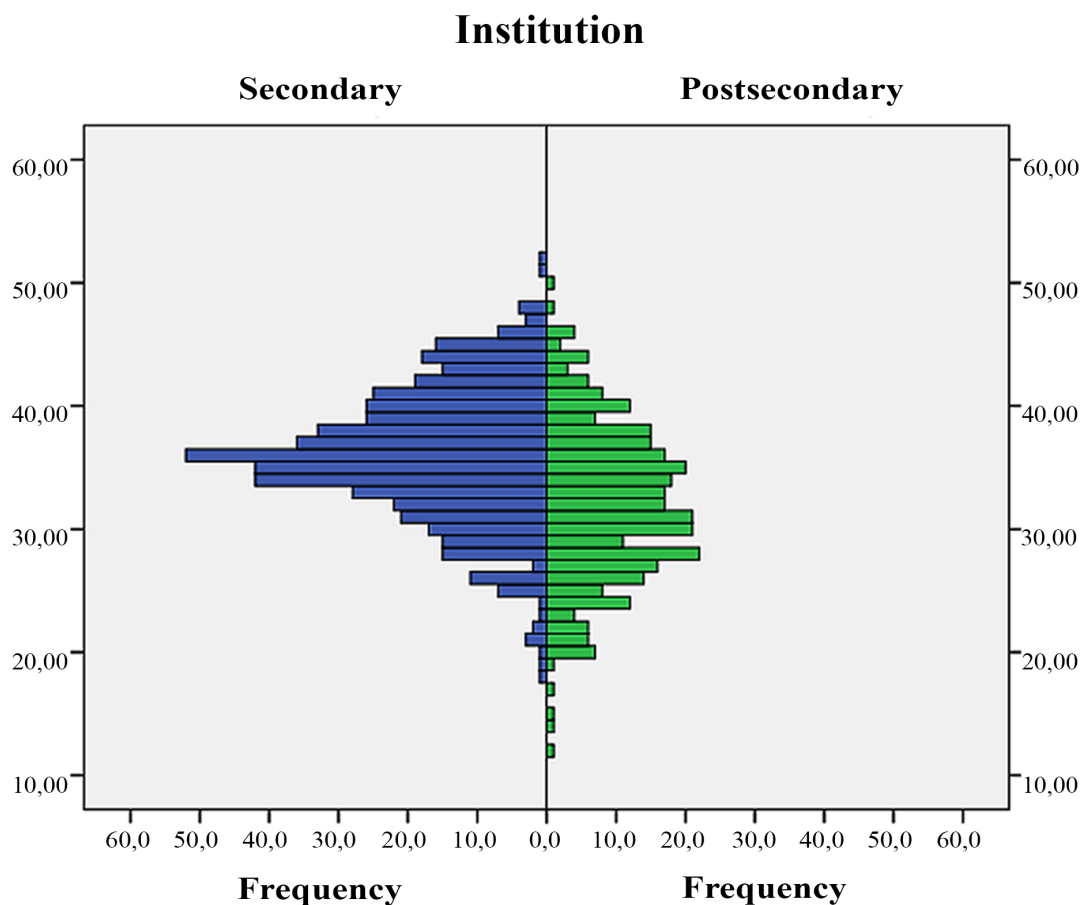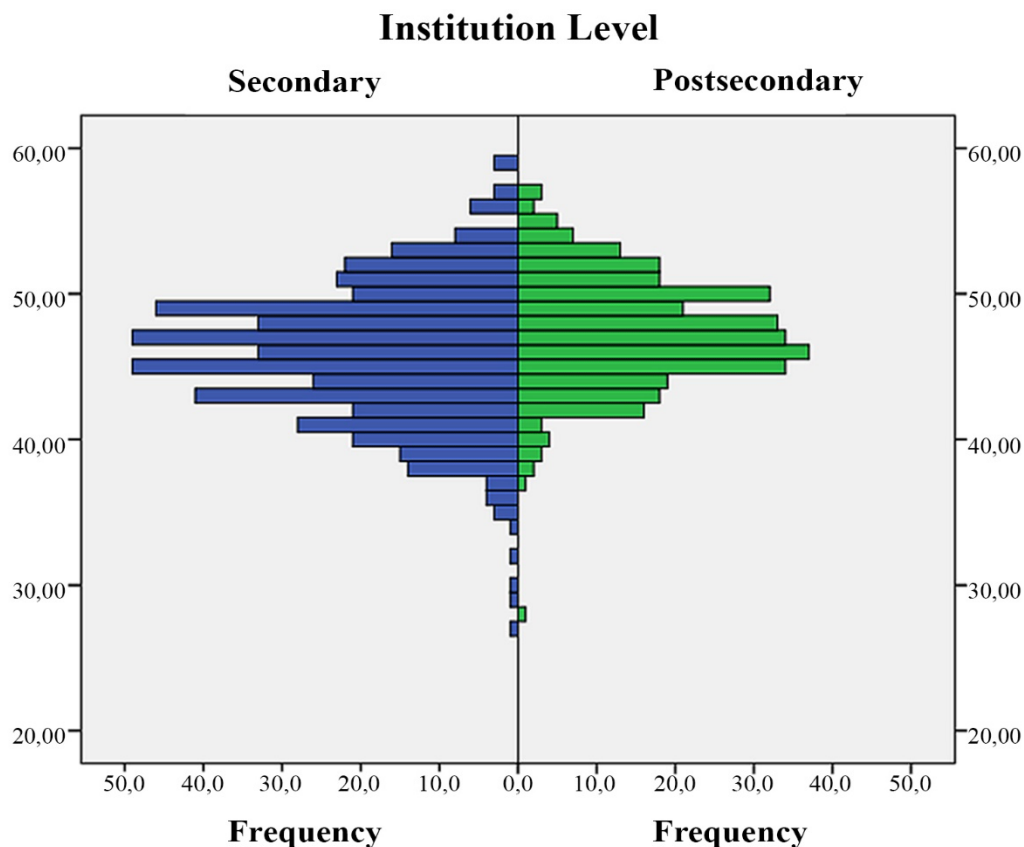
FIGURE 2: Comparison between secondary and postsecondary students scores on R&P scale items

With regard to H1, we found that as expected, the average scores of secondary and postsecondary students were significantly different on both the K&R and R&P scales. On the K&R scale, a statistically significant difference was found between secondary students (n=514; M= 36.03; SD=5.49) and postsecondary students (n=322; M=32.05; SD=6.47); t(599.28)=9.17; p<.001, 95% CI [3.13, 4.83], with secondary students' higher average scores reflecting more naïve beliefs. The effect size of level of schooling, d=.66, would traditionally be interpreted as moderate.

On the R&P scale, a statistically significant but smaller difference was found between secondary students (n=494; M=45.83; SD=4.79) and postsecondary students (n=324; M=47.36; SD=3.87); t(781.34)=-5.04; p<.001, 95% CI [-2.13, -.94]. The effect of level of schooling would traditionally be interpreted as small (d=.35), but the expected direction was in line with what Shemilt's theory would predict – ie. postsecondary students were assessed as expressing more sophisticated beliefs.

Analysis did not produce evidence to support H2. Despite gathering complete HAD responses from approximately 100 students at each grade level from 8 to 12, we found no significant differences between grade levels on either the K&R or R&P scales.

At the postsecondary level, the number of research participants we were able to recruit varied widely by involvement with the discipline of history. For this reason, before conducting analyses to test H3, Levene's test was used to ensure the homogeneity of variances on each scale among four groups represented in the data: those not taking history, those taking a history elective, history minors and history majors. (Given the small numbers of history Masters and Ph.D. students included in our sample it seemed likely that their responses would be idiosyncratic; therefore these two groups were excluded from this particular analysis.)

TABLE 3: DIFFERENCES IN K&R SCALE SCORES BY INVOLVEMENT WITH HISTORY

Involvement Comparison	Mean Difference (out of 60)	Significance	Cohen's *d*
Not taking history vs. history elective	1.18		
Not taking history vs. history minor	3.79		
Not taking history vs. history major	5.09	*p*<.01	0.88
History elective vs. history minor	2.61		
History elective vs. history major	3.92	*p*<.001	0.67
History minor vs. history major	1.31		

An omnibus ANOVA test was used to determine whether differences existed among the four groups. *F*-tests showed that significant differences ($p <. 05$) did exist between these groups for both the K&R and R&P scales; so we used the Tukey-Kramer test to examine the differences on each scale between those not studying history, the elective takers, history minors, and history majors.

As indicated in Tables 3 and 4, the differences found between the groups were generally non-significant; but the significant differences were in the directions that would be predicted by Shemilt's theory. On the K&R scale, students not taking history or taking a history elective had average scores that suggested greater naivete than those of students completing a history major. Using Cohen's *d*, the effect size of taking a history major versus not taking history was calculated at .88. This would traditionally be interpreted as a large effect. The effect of taking a history major versus a history elective was calculated at .67, which would traditionally be interpreted as a moderate effect. On the R&P scale, students taking a history elective course were found to have significantly lower scores on average than those working on a history minor. The effect of taking a history minor versus an elective course was calculated at .60, a moderate effect.

TABLE 4: Differences in R&P scale scores by involvement with history

Involvement Comparison	Mean Difference (out of 60)	Significance	Cohen's *d*
Not taking history vs. history elective	- 0.43		
Not taking history vs. history minor	- 2.51		
Not taking history vs. history major	- 1.30		
History elective vs. history minor	- 2.09	*p*<.05	0.60
History elective vs. history major	- 0.88		
History minor vs. history major	1.21		

Shemilt's theory described a stagewise progression in which naïve conceptions about differing historical accounts are abandoned as more sophisticated ones are taken up. If the theory were correct, students should not appear to hold naïve and sophisticated beliefs at the same time. Logically according to H4, students' scores on the K&R and R&P scales should be inversely correlated; but we did not find this to be the case. Pearson's *r* was calculated at .089 (*p*=0.13) between the scores for all participants on the K&R and R&P scales. Therefore, H4 was not supported.

Discussion

In order to refine their practices of inquiry-oriented history teaching, instructors would benefit from practical and time-efficient ways to assess their students' epistemological beliefs. HAD shows promise for addressing the specific need of assessing students' beliefs about why historical accounts may differ.

In important respects, the findings of our validation effort were consistent with the predictions of Shemilt's original developmental theory. The validity of scores derived from HAD gains credence from statistically significant differences detected between secondary and postsecondary students' scores on the questionnaire, and statistically significant differences detected between students with varying involvement in history as a discipline at the postsecondary level. We were surprised not to find significant differences in secondary students' scores between grades 8 and 12; but it is important to note that Shemilt did not claim that students' ideas would develop simply as a function of age, but rather as a function of experience and history teaching. When our data were collected, the social studies and history curriculum in British Columbia focused on content coverage, largely to the exclusion of historical research methods. Therefore, the educational treatment across these grades was not necessarily of a kind that would be expected to increase the sophistication of students' epistemological beliefs in history.

A key finding that ran counter to Shemilt's 4-stage theory was the lack of a significant *inverse* relationship between students' scores on the K&R and R&P scales. Students do not appear to abandon altogether the beliefs that Shemilt's theory describes as unsophisticated as they gain experience with history as a discipline. Rather than a rigid stagewise progression in students' conceptions regarding differing accounts, our data suggest instead that greater exposure to the discipline of history tends to reduce students' *exclusive* reliance on bias and storyteller ignorance as explanations for differing accounts – broadening the range of possible explanations they consider. This finding echoes those of other researchers, who have noted signs of "epistemological wobbling" in their own studies (VanSledright and Reddy, 2014; Stoel et al., 2017). In the context of the broader research on personal epistemology these findings might be considered unsurprising; but they beg consideration as to whether researchers focused on epistemic cognition in history should be as committed as they are to developmental theories rather than dimensional ones.

Future Directions

The main purpose of the HAD questionnaire is to serve as a tool in history teachers' curriculum planning and evaluation. As Chapman (2017) noted,

> There are impressive examples, in the pedagogic literature, of strategies that seek to engage students in the analysis of historical interpretations as 'deliberate' and 'crafted' artefacts…designed to achieve particular effects… (p. 101).

The important question for classroom teachers is whether and when to deploy the strategies referred to by Chapman. In the future, we plan to make the HAD questionnaire freely available on a web site so that instructors can have their own students complete it anonymously, and be immediately presented with automatically-scored results. Ideally, the web site would also run statistical tests automatically, so that for example, a teacher would be able to know whether her students' beliefs had, in aggregate, significantly shifted after her deliberate attempts to enrich their understanding of differing accounts using the kinds of activities that Chapman refers to.

Another potential use of HAD is to gauge students' readiness for activities that involve working with differing accounts. For this to be feasible, further research needs to be done to establish benchmarks for the instrument. At this point we have no empirical basis to assert what a particular set of scores on the K&R and R&P scales implies in terms of the kinds of tasks that students are likely to be capable of. However, the work of Wiley et al. (2020) is encouraging, for it

suggests that a targeted measure of epistemological beliefs like HAD may help predict when a particular group of students is in a position to benefit from particular classroom tasks.

Through carefully planned investigations, we hope to develop HAD into a useful tool for instructors as they experiment with novel curriculum and pedagogical approaches in modern, multicultural history classrooms.

Limitations

A number of limitations are worthy of mention. First, since our primary objective was to develop a practical classroom assessment that could be completed in approximately 15 minutes, our instrument did not include a large enough number of items to reliably assess students' commitment to beliefs associated with each individual stage of Shemilt's developmental theory. If the primary objective of the research had been to test the four-stage theory in detail, more items would have been required.

Second, this study did not involve gathering data on the specific content or methods of instruction that our questionnaire respondents had been exposed to in history classrooms. For this reason, we cannot make detailed claims about *why* students' HAD scores differed (or did not differ) according to the depth of their involvement in history as a discipline, or their exposure to more or less history instruction. The secondary school data represent a partial exception to this limitation, since we know that during the time of our data collection (and for many years prior) secondary school curriculum in the area where our data were collected did not mandate any study of historical research methods. Due to curriculum revisions this situation is now changing, so in years to come our data may provide a baseline against which to examine the efficacy of the curriculum changes.

Third, the differences we detected between secondary students' and postsecondary students' scores on HAD cannot be attributed to the influence of postsecondary instruction alone. While the postsecondary institutions involved in our study do not select students based on the sophistication of their epistemological beliefs, it may be the case that those who are admitted to and choose to pursue postsecondary education hold more sophisticated beliefs due to their socioeconomic status and other influences, rather than by virtue of postsecondary instruction alone. Further studies will be needed to disentangle students' epistemological beliefs from their socioeconomic status, institutions' choices to admit, students' long-term goals, and other variables.

Fourth, limitations on both our access to potential research participants and the time available to conduct our study resulted in a small number of responses for some groups of strong potential interest in a study of this kind – particularly Masters and Ph.D. students in history. In future work it will be important to better represent these groups.

Finally, as with Shemilt's original 1987 study, the work presented here involved using cross-sectional data to make inferences about a process of change in students' epistemological beliefs that unfolds over years. In the absence of instruction aimed at improving students' epistemological beliefs, change may never occur, or may take a long time; however our large-scale data were collected in just a few months. To directly examine change in students' epistemological beliefs, one would ideally follow the same cohort of students longitudinally over a period of years, gathering simultaneous data regarding the nature of the history instruction they received, extracurricular encounters with history (e.g. museums, popular media) and other possible influences.

Conclusion

To our knowledge, the research reported here represents the first attempt to design and validate a domain-specific quantitative measure of students' beliefs about differing historical accounts that would be practical for instructors to use independently in their own classrooms. Acknowledging their limitations, the data and analyses presented above suggest that the HAD questionnaire,

which is quick for students to complete and can be scored easily, holds promise. With further research we hope to demonstrate its utility for classroom teachers. With regard to theory, our failure to find evidence of stage-like developmental progression in our data resonates with the findings of other researchers, and highlights limitations in the ability of developmental theories to describe the changes that take place in students' beliefs about historical knowledge over the course of their education.

Acknowledgements

References

Banks, J. A. (2008). Diversity, group identity and citizenship education in a global age. *Educational Researcher, 37*(3), 129-139. http://dx.doi.org/10.3102/0013189X08317501

Barzilai, S., & Weinstock, W. (2015). Measuring epistemic thinking within and across topics: A scenario-based approach. *Contemporary Educational Psychology, 42,* 141–158. https://doi.org/10.1016/j.cedpsych.2015.06.006

Britt, M. A., Perfetti, C. A., Van Dyke, J. A., & Gabrys, G. (2000). The Sourcer's Apprentice: A tool for document-supported history instruction. In P. Stearns, P. Seixas, & S. Wineburg (Eds.), *Knowing, teaching and learning history: National and international perspectives* (pp. 437-470). New York: New York University Press.

Castro, A. J., Field, S. L., Bauml, M., & Morowski, D. (2012). "I want a multicultural classroom": Preparing social studies teachers for culturally diverse classrooms. *The Social Studies, 103*(3), 97-106. http://dx.doi.org/10.1080/00377996.2011.575419

Chandler, M. J., Hallett, D., & Sokol, B. (2002). Competing claims about competing knowledge claims. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing.* Mahwah, NJ: Erlbaum.

Chapman, A. (2017). Historical interpretations. In I. Davies (Ed.), *Debates in history teaching* (pp. 96-108). London: Taylor and Francis. http://dx.doi.org/10.4324/9780203831458

Conrad, M., Ercikan, K., Friesen, G., Létourneau, J., Muise, D., Northrup, D., & Seixas, P. (2013). *Canadians and their pasts.* Toronto: University of Toronto Press.

Cronon, W. (1992). A place for stories: Nature, history and narrative. *The Journal of American History, 78*(4), 1347-1376.

Ercikan, K., & Seixas, P. (Eds.). (2015). *New directions in assessing historical thinking*. New York: Routledge. http://dx.doi.org/10.4324/9781315779539

Lee, P. (2004). Understanding history. In P. Seixas (Ed.), *Theorizing historical consciousness* (pp. 129-164). Toronto: University of Toronto Press.

Lee, P. (2005). Putting principles into practice: Understanding history. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics and science in the classroom* (pp. 31-77). Washington, DC: The National Academies Press.

Lee, P., & Shemilt, D. (2004). "I just wish we could go back in the past and find out what really happened": Progression in understanding about historical accounts. *Teaching History, 117,* 25-31.

Lightning, J. (2021). Revisiting Lexington Green: Implications for teaching historical thinking. *Cognition & Instruction, 39*(3), 306-327. http://dx.doi.org/10.1080/07370008.2021.1880410

Maggioni, L., VanSledright, B., & Alexander, P. A. (2009). Walking on the borders: A measure of epistemic cognition in history. *The Journal of Experimental Education, 77*(3), 187-213. http://dx.doi.org/10.3200/JEXE.77.3.187-214

Miller, G. R., & Toth, S. L. (2012). To dismantle an idle past: Using historiography to construct a digital learning environment. *The Social Studies, 103*(2), 73-80. http://dx.doi.org/10.1080/00377996.2011.571303

Mitrovic, B. (2015). Historical understanding and historical interpretation as contextualization. *History and Theory, 54,* 311-332. http://dx.doi.org/10.1111/hith.10762

Muis, K., Bendixen, L. D., & Haerle, F. C. (2006). Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review, 18*(1), 3-54. http://dx.doi.org/10.1007/s10648-006-9003-6

Nitsche, M., & Waldis, M. (2018). *Examining History Teachers' Epistemological and Pedagogical Beliefs.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

O'Neill, D. K., Guloy, S., & Sensoy, Ö. (2014). Strengthening methods for assessing students' metahistorical conceptions: Initial development of the Historical Account Differences survey. *The Social Studies, 105*(1), 1-14. https://doi.org/10.1080/00377996.2013.763524

Perry, W. G. (1970). *Forms of intellectual and ethical development in the college years: A scheme.* New York: Holt, Rinehart and Winston.

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction, 30*(1), 86-112. http://dx.doi.org/10.1080/07370008.2011.634081

Reisman, A. (2015). The difficulty of assessing disciplinary historical reading. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 29-39). New York: Routledge. http://dx.doi.org/10.4324/9781315779539

Rothschild, E. (2000). The Impact of the Document-Based Question on the Teaching of United States History. *The History Teacher, 33*(4), 495-500. http://dx.doi.org/10.2307/494945

Schommer-Aikins, M. (2004). Explaining the Epistemological Belief System: Introducing the Embedded Systemic Model and Coordinated Research Approach. *Educational Psychologist, 39*(1), 19-29. http://dx.doi.org/10.1207/s15326985ep3901_3

Seixas, P. (2006). *Benchmarks of historical thinking: A framework for assessment in Canada.* Retrieved from Vancouver, BC: http://www.histori.ca/benchmarks/

Seixas, P. (Ed.) (2004). *Theorizing Historical Consciousness*. Toronto: University of Toronto Press.

Shemilt, D. (1987). Adolescent ideas about evidence and methodology in History. In C. Portal (Ed.), *The History curriculum for teachers* (pp. 39-61). London, England: Falmer.

Shemilt, D. (2000). The caliph's coin: The currency of narrative frameworks in History teaching. In P. N. Stearns, P. Seixas, & S. Wineburg (Eds.), *Knowing, teaching and learning History: National and international perspectives* (pp. 83-101). New York: New York University Press.

Smith, M., & Breakstone, J. (2015). History assessments of thinking: An investigation of cognitive validity. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 233-245). New York: Routledge. http://dx.doi.org/10.4324/9781315779539

Stoel, G., Logtenberg, A., Wansink, B., Huijgen, T., van Boxtel, C., & van Drie, J. (2017). Measuring epistemological beliefs in history education: An exploration of naïve and nuanced beliefs. *International Journal of Educational Research, 83,* 120-134. http://dx.doi.org/10.1016/j.ijer.2017.03.003

Takaki, R. T. (1993). *A different mirror: A history of multicultural America.* Boston: Little, Brown & Co.

VanSledright, B., & Reddy, K. (2014). Changing epistemic beliefs? An exploratory study of cognition among prospective history teachers. *Tempo e argumento, 6*(11), 28-68. http://dx.doi.org/10.5965/2175180306112014028

VanSledright, B. (2015). Assessing for learning in the history classroom. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 75-88). New York: Routledge. http://dx.doi.org/10.4324/9781315779539

Wiley, J., Griffin, T. D., Steffens, B., & Britt, M. A. (2020). Epistemic beliefs about the value of integrating information across multiple documents in history. *Learning and Instruction, 65,* 101266. http://dx.doi.org/10.1016/j.learninstruc.2019.101266

Wineburg, S. (1991). On the reading of historical texts: Notes on the breach between school and the academy. *American Educational Research Journal, 28*(3), 495-519. http://dx.doi.org/10.3102/00028312028003495

Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past*. Philadelphia: Temple University Press.

About the Authors

D. Kevin O'Neill, B.Sc., PhD is Associate Professor of Education and Technology at Simon Fraser University, where he helps to prepare future teachers, and teaches and mentors students in the Educational Technology and Learning Design graduate programs. He has had a longstanding interest in history teaching and design-based research.

Email: kevin_oneill@sfu.ca ORCiD: 0000-0001-6880-9146

Sheryl Guloy, PhD, provides educational and knowledge translation consulting services and leads a sleep coaching and professional education corporation in Montreal, Canada. She holds an MA and a PhD in Educational Technology as well as a BComm in International Business. Her research involves virtual mentoring environments, leadership development, higher education, program development, and mobile apps to aid sleep.

ORCiD: 0000-0003-4735-8344

Fiona M. MacKellar, MA is an instructional designer and PhD candidate who came to history education research sideways-fashion during work on her MA thesis. The focus of her current dissertation research is how secondary students come to grasp and to advance their understanding of historical evidence.

ORCiD: 0000-0002-0534-3892

Dale R. Martelli, MA is a PhD candidate in Educational Theory and Practice: Philosophy of Education at Simon Fraser University, and has been a practicing teacher for more than 30 years, in Canada, Nigeria and Israel. He currently teaches secondary history and philosophy in Vancouver, and is the sitting president of the British Columbia Social Studies Teachers' Association.

ORCiD: 0000-0002-5897-0128

Appendix A: The Historical Account Differences Questionnaire[1]

Instructions: Carefully read each question, and rate each of the four possible answers that follow it, according to how strongly you agree or disagree with it. This is not a test, and there are no right answers.

Q: What makes somebody a historian?

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
They re-consider past events, and think about what those events mean to us today. (4)					
They can figure out which stories about the past are biased or untrue, and put together the one true story. (2)					
They can use evidence to figure out what probably happened in the past. (3)					
They have studied and memorized what happened in the past (1)					

Q: How do historians develop their stories about events in the past?

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
They consider several possibilities and make an educated guess about what happened based on the evidence available. (3)					
They add new stories by re-examining evidence to answer questions that may not have been asked before. (4)					
They look at what several people have said about the events, and decide which one's story is the least biased. (2)					
They read only what has been written or recorded about the event by the witnesses who were there to see it. (1)					

---

[1] The Shemilt stage that each belief statement is intended to capture is marked in parentheses for the convenience of readers. These stages do not appear on versions used with students.

Q: Why do historians write or tell stories about the past?

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
Because questions come up today that can be answered by looking at the past (4)					
So that other people can know what happened in the past (1)					
To provide other people with a careful examination of the best evidence available today (3)					
Because they want to provide other people with stories that are true and unbiased (2)					

Q: Why do historians write new books about events that were already written about before?

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
Historians write new books because old historical evidence can be examined in new ways that give us a better understanding. (3)					
Historians should not write new books. Older books are more trustworthy because they were written closer in time to the events themselves. (2)					
Historians writing new books makes no sense because there should only need to be one book about each event. (1)					
Historians write new books because as time goes on, people have new questions that need to be answered. (4)					

Q: What makes a story about the past useful?

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
It comes from a witness who intends to be fully truthful and unbiased (2)					
It gives you a straightforward description of what happened in the past (1)					
It answers your questions about why things turned out the way they have (4)					
It is based on careful examination of a variety of evidence from the past (3)					

Q: If a historian is learning about the events of a period and finds stories that seem different, what should she do?

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
She should try to figure out which story writer is less biased or was closer to the events, and use only that one. (2)					
She should try to understand the questions that the people who wrote the stories were trying to answer. (4)					
She should make an educated guess about what most likely happened, based on other evidence. (3)					
She should disregard both stories because it does not make sense for there to be different stories. (1)					