13

Using Artificial Intelligence to Assess the Level of Cognitive Complexity Involved in Didactic Tasks. A case study on Historical Thinking

Mg. Connie Cofré-Morales
Universidad Viña del Mar, Chile
connie.cofre@docente.uvm.cl
https://orcid.org/0009-0000-1442-1218

Eduardo Puraivan
Universidad Viña del Mar, Chile
Universidad de Playa Ancha, Chile
epuraivan@uvm.cl
https://orcid.org/0000-0003-2134-8922

Ing. Karina Huencho-Iturra
Universidad Viña del Mar, Chile
karina.huencho.iturra@alumno.uvm.cl
https://orcid.org/0009-0009-7043-2929

Mg. Macarena Astudillo-Vásquez
Universidad Viña del Mar, Chile
mastudillo@uvm.cl
https://orcid.org/0000-0001-5840-0865

PhD Carlos Hervás-Gómez
Universidad de Sevilla, Spain
hervas@us.es
https://orcid.org/0000-0002-0904-9041

PhD M. Dolores Díaz-Noguera
Universidad de Sevilla, Spain
noguera@us.es
https://orcid.org/0000-0002-0624-4079

https://doi.org/10.36006/09643-1-13

Abstract
The design of didactic activities is aimed at developing learning objectives at different levels of cognitive complexity. In the field of History teaching, one of the purposes is to develop Historical Thinking, which requires specific tasks. However, several investigations report that many of the activities presented in school textbooks are limited to the use of the lowest cognitive levels, not achieving the intended Historical Thinking of students. Faced with this singularity, it is valid to ask how we can use some tools to validate that the activities fulfill the expectations of cognitive complexity. In this paper, we discuss the use of an AI to evidence whether it can help as a support tool for this task. The results show that the task of assigning the predominant cognitive level is challenging even for highly qualified experts, and that the AI results match at least with expert's assessment using some of the taxonomies considered, except for one activity. On the other hand, there is a high appreciation by experts of the potential of ChatGPT to both classify and argue its decisions, although there are also some risks to be considered.

Keywords: Artificial Intelligence, ChatGPT, cognitive complexity, historical thinking.

13.1. Introduction

In the subject of History, from the perspective of the development of skills and competencies, in recent decades, the focus has been placed on students being able to develop "historical thinking", which is defined as the deployment of a set of specific skills of the discipline to ensure that the historical facts of the subject of History are not memorized, but constructed by the students themselves through the use of historical evidence or sources to generate an interpretation of the facts, emulating the work of a historian (Wineburg, 2001; Kitson, et al., 2015; Sáez-Rosenkranz, 2017). In this sense, the teaching of historical thinking consists in initially posing a problem or research question that guides the inquiry that students will carry out in the historical sources, in order to solve the problem posed or provide an answer to the question (Seixas & Morton, 2013; Kitson et al., 2015; Henríquez et al., 2018).

The levels of cognitive complexity identifiable in didactic tasks that seek to develop specific objectives and skills, as is the case of historical thinking, can be identified through the use of

taxonomies that describe what the student does at the cognitive level and, at the same time, allow us to show whether these activities really promote the development of this specific skill of the subject. Bloom's Taxonomy (in the Anderson and Krathwohl's revision) has been used as a means in several studies (Sáiz, 2014; Gómez & Miralles, 2016; Martínez & Gómez, 2018) to identify the levels of complexity of textbook tasks. This taxonomy proposes 6 levels that are presented continuously and ascending in complexity, although they are not necessarily interpreted as linear, in the sense that a student, when working at one of the highest levels, has necessarily passed through all the previous levels (Förster & Rojas-Barahona, 2017). These levels are: "remember", "understand", "apply", "analyze", "evaluate" and "create". Regarding the levels of cognitive complexity but specifically focused on the subject of History and historical thinking, Sáiz (2013) has formulated 3 levels in which there is correspondence with the levels of Bloom's Taxonomy (Anderson and Krathwohl's revision): "low", where students locate and repeat information extracted from historical sources, corresponding to Bloom's "remember" level; "medium", where students understand information from other sources by paraphrasing, summarizing or schematizing it, corresponding to Bloom's "understand" level; and "high", where students analyze, apply and/or evaluate information from sources in order to construct new knowledge about the past. This last level corresponds to Bloom's "apply", "analyze", "evaluate" and "create" levels.

Several studies have investigated the levels of cognitive complexity that are most frequently presented in the didactic tasks of textbooks for the subject of History, in order to verify whether it is possible to get students to develop historical thinking with the support of these activities. In these studies, it is observed as a tendency that the levels of complexity that present a greater presence in the textbooks are those that are lower, and it is highlighted that the higher levels present a very scarce presence (Sáiz, 2014; Gómez, 2014; Gómez & Miralles, 2016; Martínez & Gómez, 2018; Palacios, 2019; Bramann, 2021).

Given the advancement of Artificial Intelligence (AI) in education, recent studies (Bolaño-García & Duarte-Acosta, 2024; Jiménez et al., 2018; Kasneci et al., 2023; Küchemann et al., 2023) have explored its usefulness and risks. The capacity of AI

as a teaching support tool is recognized, motivating investigations into its potential applications. In our case, it is crucial to evaluate whether AI can effectively classify the levels of complexity of didactic tasks in History, and how reliable its work is according to experts in the field.

13.2. Background

The possible applications of AI in education have been actively explored in recent times due to its rapid advancement and progress, however, regarding the opportunities and challenges offered in this area by the use of chatbots such as ChatGPT, authors such as Kasneci et al. (2023) state that the studies conducted to experiment in this area are at an early stage, with few empirical studies being found in the literature. In the field of teaching, it is possible to find various uses of AI where it acts as a support tool for teaching. For example, according to Bolaño-García & Duarte-Acosta (2024) from a systematic review done to investigate the use of AI in education, it has been detected that AI is a tool that can be used for the personalization of learning in the sense that teachers can adapt teaching materials to the individual needs of their students immediately; to provide students with instant and personalized feedback on their learning tasks and activities; and to automatize administrative and evaluative tasks in which teachers spend a great amount of time, among other notable uses. Moreover, Kasneci et al. (2023) report that AI offers opportunities for teaching where it can be used for lesson planning or inclusive classes, as well as to generate questions or activities that promote the participation of people with different abilities and/or levels of knowledge; likewise, the mentioned authors point out that AI can semi-automatize the grading of students, highlighting the strengths and weaknesses of their work.

Regarding didactic activities for student learning, some studies have experimented with the use of AI to generate such activities and evaluate their quality. The study by Küchemann et al. (2023) worked with an intervention group that had to create activities for the subject of Physics using ChatGPT, which were subsequently compared with those created by a textbook-sup-

ported control group on the same content. In the relevant findings of the aforementioned work, it is highlighted that no significant differences were detected in the quality of the activities generated by both groups. It is mentioned, however, that one of the activities created by the group that used ChatGPT presented an important omission that did not allow determining the response to the activity. Regarding the taxonomic levels of the activities (based on Bloom's Taxonomy), it was detected that the activities generated by both groups were concentrated in the levels of "remember", "understand" and "apply", with very few activities at the level of "analyze" and "evaluate", and none belonged to the level of "create" in the activities of both groups.

On the other hand, Kwan (2024), using ChatGPT, generated an assessment script that included a test, scoring guideline, suggested solutions to the test, and classification of the difficulty levels of the test questions based on Bloom's Taxonomy levels. Among the main findings, it was detected that the assessment generated by ChatGPT was quite structured. However, it was observed that one question was incorrect; that there were inconsistencies in the scoring guideline related to the score of some questions and the total score given by the chat itself; and that, when requesting the solution to the test, the chat only generated the solution to one of the test sections (of the two existing in total), failing to suggest a solution for the other section. With respect to the taxonomic level classification (performed with ChatGPT) of the generated assessment questions, it is argued that most of them are at the "remember" and "apply" levels, and that two of them are at the "analyze" level. Therefore, of the questions generated by ChatGPT, no questions for the higher levels of "evaluate" and "create" are detected.

From the above mentioned, two aspects stand out regarding the generation of tasks using ChatGPT and its subsequent classification of the levels of cognitive complexity based on Bloom's Taxonomy. Firstly, it is striking that the activities generated using the chat possess the same frequency trends of taxonomic levels as the textbook activities that other studies have classified (Sáiz, 2014; Gómez, 2014; Gómez & Miralles, 2016; Martínez & Gómez, 2018), namely, activities concentrated in the first 4 levels of the Taxonomy (those of the lowest cognitive level), and very few of the "evaluate" and "create" levels (the highest

cognitive levels). Secondly, it is observed that ChatGPT has the ability to classify the levels of cognitive complexity of didactic activities, taking Bloom's Taxonomy as a reference; however, the referenced studies have not verified whether the classification made by the chat is correct, thus it is not possible to affirm that this ability, at present, is accurate and effective. This last point is of particular relevance, since, as has been suggested by some authors (Hashem et al., 2024; Kwan, 2024; Gill et al., 2024), in some cases ChatGPT can generate inaccurate or incorrect data.

Therefore, in this study, we proposed to use ChatGPT to perform a classification of the taxonomic levels of a set of didactic activities extracted from two textbooks circulating in Chilean schools in 2022, specifically from the subject of History, Geography and Social Sciences. We also performed a validation, under the criteria of experts in the discipline, of the classification and arguments provided by the chat to verify its accuracy and reliability

13.3. Methodology

We asked ChatGPT 4 (hereinafter ChatGPT) to analyze didactic activities and determine the predominant cognitive level, according to the Taxonomies of Bloom (Anderson and Krathwohl's revision) and Sáiz (2013), hereinafter Bloom's and Sáiz's Taxonomy, respectively. As a case study, we used didactic tasks in the field of Historical Thinking. The methodology used in this work considers the following stages.

- **Selection and Preparation of Didactic Activities.** For our study, we chose and prepared six didactic activities, all extracted from two textbooks used in Chilean schools during the year 2022. These activities belong specifically to the subject of History, Geography and Social Sciences. Each of them was stored in separate files to facilitate their analysis and management.
- **Elaboration of Prompts for ChatGPT.** We designed specific prompts to guide ChatGPT in the analysis of didactic activities. These prompts are oriented to determine the cognitive

level associated with each activity, according to Bloom's and Sáiz's taxonomies. In addition, ChatGPT was asked to provide arguments to justify each of its decisions in this classification process.

- **Obtaining Results.** In this phase, we compiled the classifications made by ChatGPT together with their corresponding justifications. These results are based on Bloom's and Sáiz's taxonomies used in the analysis. This process allowed us to evaluate how the chat determines the cognitive levels of the didactic activities.
- **Expert Critical Assessment.** For a detailed and critical review of the classifications made by ChatGPT, we convened six highly qualified experts, all with doctoral degrees. Initially, each expert was asked to determine the predominant cognitive level in the didactic activities, using Bloom's and Sáiz's taxonomies as a reference. Subsequently, we presented them the ChatGPT results and asked them to critically evaluate both the classifications and the arguments offered by the chat. This analysis focused on identifying the potential risks and benefits of using ChatGPT.

13.4. Results

Below we present the findings of our experiment with six didactic activities extracted from two textbooks used in Chilean schools during 2022, focusing on the area of History, Geography and Social Sciences. These materials are representative of those used in the first two years of secondary education (students aged 14-15 years).

Classification of activities

We prepared and presented the didactic activities extracted from the textbooks of the subject of History, Geography and Social Sciences to each of the experts. Figure 13.1 is an example of an activity presented (the question asks "What does the act of source A symbolize? Why was this date chosen?").

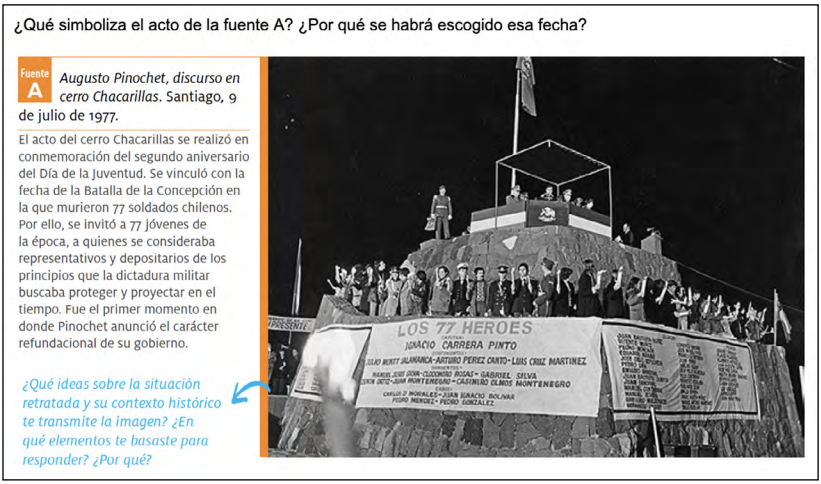

Figure 13.1. Example of an activity extracted from a textbook.

Each expert identified the predominant cognitive level of the didactic tasks, based on Bloom's and Sáiz's taxonomies. The assessment was carried out individually by two experts for each activity. After completing their classification, they were shown the results obtained with ChatGPT for direct comparison. The details of these evaluations and comparisons are summarized in Table 13.1. In the ChatGPT's column we present the classification of both taxonomies and in parentheses the percentage of agreement with the experts considering each taxonomy. Interesting situations emanate from these results. Firstly, note that there is no agreement among experts in determining the predominant cognitive level of each activity, except when using Sáiz's taxonomy in Activity 5 and when using Bloom's taxonomy in Activity 6. This empirical finding confirms the complexity of this task. On the other hand, we see that the classification performed by ChatGPT is in line with that of at least one expert (examining the taxonomies individually), except in activity 5. We have highlighted (gray background) the coincidences between ChatGPT and the experts.

Table 13.1. Experts and ChatGPT results (Ai: Activity "i", T: Taxonomy, *Ej: Expert "j"*.

A	T	ChatGPT	E_1	E_2	E_3	E_4	E_5	E_6
A_1	Bloom	Analyze (50%)	Analyze	Understand				
	Sáiz	High (50%)	High	Medium				
A_2	Bloom	Analyze (0%)	Evaluate	Remember				
	Sáiz	High (50%)	High	Medium				
A_3	Bloom	Evaluate (50%)			Evaluate	Remember		
	Sáiz	High (50%)			High	Low		
A_4	Bloom	Analyze (50%)			Analyze	Remember		
	Sáiz	Medium (50%)			Medium	Low		
A_5	Bloom	Analyze (0%)					Evaluate	Understand
	Sáiz	High (0%)					Medium	Medium
A_6	Bloom	Evaluate (100%)					Evaluate	Evaluate
	Sáiz	High (50%)					Medium	High

Source: developed by author.

Assessment of the work performed by ChatGPT

As was previously mentioned, once their classification was completed, the experts were shown the results obtained with ChatGPT, asking them to rate (on a scale of 1 to 5) the classification and arguments provided by the chat. Table 13.2 shows an example of the classification and general arguments provided by ChatGPT.

Table 13.2. Example of classification and arguments provided by ChatGPT

Taxonomy	Assigned level	Arguments
Bloom (Anderson and Krathwohl's revision)	Analyze	The task requires the student to interpret the symbolism of a public act, understanding the meaning behind the selection of the date and the relationship to a historical event. This demands analytical skills to connect the visual information with previous knowledge of history and politics.
Model of Sáiz (2013)	High	The student must go beyond the mere identification of visual elements or textual comprehension (low and medium levels). They are expected to make a critical evaluation of the act in its historical and political context, which implies the creation of new information from a deep reflection on the symbology and intentionality behind the selection of the date.

The average rating is shown in Table 13.3. It is interesting to note that there is a high valuation in both dimensions. On the other hand, the experts based their ratings both on the classification and on the arguments provided by ChatGPT. Among their justifications (free text), based on identifying benefits and risks, the experts show interest and surprise in the classification made by ChatGPT, but they also state as a risk that teachers may slightly accept the results and rely too much on the tool. They make explicit reference to the fact that the tool can be wrong, and that some historical contexts may not be well understood, which can lead to errors in the results.

Table 13.3. Average ratings in the classification and arguments provided by ChatGPT *(Ai: Activity "i")*.

INDICATOR	A_1	A_2	A_3	A_4	A_5	A_6
Average rating in the classification	4	4	5	5	3	5
Average rating in the argument	5	4.5	5	5	4	5

Source: developed by author.

13.5. Conclusions and Projections

The achievement of the learning objectives is of utmost relevance for the quality of the teaching that is delivered. In this sense, one

of the aspects to consider and review in teaching is the complexity of the didactic activities proposed to students, to ensure that they are appropriate to the expected learning, for which Bloom's taxonomy (Wang et al., 2021), or others specific to the subject, can be used. If the subject of History is intended to develop historical thinking, the most appropriate levels of complexity should be the highest (Sáiz, 2014; Palacios, 2019; Bramann, 2021). The use of AI as a tool for the task of classifying activities can be valuable and contribute greatly to quality teaching.

The results of our experiments indicate an initial finding: the difficulty of assigning a specific cognitive level to didactic activities, despite the fact that these were selected directly from school textbooks and are well defined. This complexity is evidenced by the lack of consensus among experts, even using a commonly employed taxonomy such as Bloom's. This finding suggests the importance of expanding research to other areas of knowledge to better understand these challenges. A second significant finding of our study is that, for each didactic activity (except for Activity 5), the ChatGPT's results match at least the classification provided by one of the experts, when considering the taxonomies individually. This coincidence indicates that the ChatGPT's classifications are in line with the experts' evaluations, showing relevant consistency. This result suggests that the tool, in terms of classification, does not present significant deviations or obvious errors that could be considered as "hallucinations" in its responses. The third finding of our study is the experts' favorable assessment towards the use of ChatGPT in the assigned task, both in terms of classification and argumentation. However, experts also warn about the need for caution when using this tool. They highlight the importance of avoiding overconfidence of teachers and recall that, like any automated tool, ChatGPT is not error-free (Hashem et al., 2024; Kwan, 2024; Gill et al., 2024), e.g., in challenging contexts such as assigning a level of cognitive complexity where there are various reasons that can make this task difficult. For instance, in some cases, an activity can be associated with more than one level of Bloom's taxonomy (Rawat et al., 2023) or the amount of time and resources associated with this task when it is performed considering a big amount of activities (Rawat et al., 2023; Wang et al., 2021), which is why it is necessary to work on an automated tool to support it. This balance

between recognition of its usefulness and awareness of its limitations is crucial for its effective application in educational contexts.

The current results lead us to plan future research in two key directions: to extend the empirical evidence with a wider range of didactic activities and to diversify the areas of knowledge studied. In the medium term, our goal is to develop a software tool based on recent advances in AI. This tool will support decision making in assigning levels of cognitive complexity to didactic activities and will be aimed primarily at trainee and novice teachers as initial users.

References

Bramann, C. (2021). *Fostering Historical Thinking with Textbooks. A case study of tasks in Austrian history textbooks* (conference). International Conference on Textbooks and Educational Media, Berlin, Germany. https://doi.org/10.1007/978-3-030-80346-9_31

Bolaño-García, M., & Duarte-Acosta, N. (2024). Una revisión sistemática del uso de la inteligencia artificial en la educación. *Revista Colombiana de Cirugía, 39*(1), 51-63. https://doi.org/10.30944/20117582.2365

Förster, C., & Rojas-Barahona, C. (2017). Aprendizaje y evaluación: lo que no se evalúa, no se aprende. In C. Förster (Ed.). *El poder de la evaluación en el aula. Mejores decisiones para promover aprendizajes* (pp. 43-74). Universidad Católica de Chile.

Gill, S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Kumar, A., Stankovski, V., Abraham, A., Ghosh, S., Lutfiyya, H., Kanhere, S., Bahsoon, R., Rana, O., Dustdar, S., Sakellariou, R. ..., & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots. *Internet of Things and Cyber-Physical Systems, 4*, 19-23. https://doi.org/10.1016/j.iotcps.2023.06.002

Gómez, C. (2014). Pensamiento histórico y contenidos disciplinares en los libros de texto. Un análisis exploratorio de la Edad Moderna en 2.º de la ESO. *Ensayos. Revista de la Facultad de Educación de Albacete, 29*(1), 131-158. https://revista.uclm.es/index.php/ensayos/article/view/498

Gómez, C., & Miralles, P. (2016). Développement et évaluation des compétences historiques dans les manuels scolaires. Une étude

comparative France-Espagne. *Spirale. Revue de Recherches en Education*, *2*(58), 53-66. https://www.cairn.info/revue-spirale-revue-de-recherches-en-education-2016-2-page-53.htm

Hashem, R., Ali, N., El Zein, F., Fidalgo, P., & Abu Khurma, O. (2024). AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. *Research and Practice in Technology Enhanced Learning*, *19*. https://doi.org/10.58459/rptel.2024.19023

Henríquez, R., Carmona, A., Quinteros, A., & Garrido, M. (2018). *Leer y escribir para aprender Historia. Secuencias para la enseñanza y el aprendizaje del pensamiento histórico.* Universidad Católica de Chile.

Jiménez, S., Juárez-Ramírez, R., Castillo, V., & Tapia, J. (2018). *Affective Feedback in Intelligent Tutoring Systems. A practical approach.* Springer. https://doi.org/10.1007/978-3-319-93197-5

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A. Seidel, T. ..., & Kasneci, G. (2023). ChatGPT for Good? On opportunities and challenges of large language models for education. *Journal of Psychology and Education*, *103*. https://doi.org/10.35542/osf.io/5er8f

Kitson, A., Steward, S., & Husbands, C. (2015). *Didáctica de la Historia en Secundaria Obligatoria y Bachillerato. Comprender el pasado.* Morata.

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research*, *19*(2). https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

Kwan, C. (2024). *Exploring ChatGPT-Generated Assessment Scripts of Probability and Engineering Statistics from Bloom's Taxonomy* (conference). International Conference on Technology in Education, Hong Kong, China. https://doi.org/10.1007/978-981-99-8255-4_24

Martínez, M., & Gómez, C. (2018). Cognitive level and thinking historically competencies in history textbooks from Spain and England: A comparative study. *Revista de Educación*, *379*, 136-159. https://doi.org/10.4438/1988-592X-RE-2017-379-364

Palacios, N. (2019). Teaching violence, drug trafficking and armed conflict in Colombian schools: Are history textbook deficient? *Issues in Educational Research*, *29*(3), 899-922. https://www.iier.org.au/iier29/palacios.pdf

Rawat, A., Kumar, S., & Samant, S. (2023). *A Systematic Review of Question Classification Techniques Based on Bloom's Taxonomy* (conference). International Conference on Computing Communication and Networking Technologies, Delhi, India. https://doi.org/10.1109/ICCCNT56998.2023.10308403

Sáez-Rosenkranz, I. (2017). La enseñanza de la historia en los libros de texto de educación básica en Chile. *Revista Enseñanza de las Ciencias Sociales*, *16*, 27-40. https://doi.org/10.1344/ECCSS2017.16.3

Sáiz, J. (2013). Alfabetización histórica y competencias básicas en libros de texto de historia y en aprendizajes de estudiantes. *Revista Didáctica de las Ciencias Experimentales y Sociales*, *27*, 43-66. http://dx.doi.org/10.7203/dces.27.2648

Sáiz, J. (2014). Fuentes históricas y libros de texto en secundaria: Una oportunidad perdida para enseñar competencias de pensamiento histórico. *Ensayos. Revista de la Facultad de Educación de Albacete*, *29*(1), 83-99. https://revista.uclm.es/index.php/ensayos/article/view/503/458

Seixas, P., & Morton, T. (2013). *The Big Six Historical Thinking Concepts*. Nelson Education. https://bibliotecadigital.mineduc.cl/handle/20.500.12365/17630

Wang, Z., Manning, K., Mallick, D., & Baraniuk, R. (2021). *Towards Blooms Taxonomy Classification Without Labels* (conference). Artificial Intelligence in Education: 22nd International Conference, Utrecht, Netherlands. https://doi.org/10.1007/978-3-030-78292-4_35

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