The Effects of Reading Prompts and of Post-Reading Generative Learning Tasks on Multiple Document Integration: Evidence from Concept Network Analysis

Abstract

Learning from multiple documents is an essential ability in today's society. This experimental study used concept network analysis to consider how reading prompts and post-reading generative learning tasks alter students' document integration performance. Undergraduates (N = 119) read three documents about Alzheimer's disease with one of two reading prompts (integrative purpose vs. detailed purpose) and then after reading completed a generative learning task (concept mapping vs. summary writing). Three days later they completed a delayed writing task and an inference verification test. Participants' written texts were converted to concept networks to evaluate conceptual level integration, including the *quantity* of integration (measured by the proportion of integrative links), the *semantic quality* of integration (measured by the similarity of integrative links), and the *structural quality* of integration (measured by comparing graph centrality). Results showed that the integrative purpose relative to the detailed purpose enhanced the quantity of integration but not the semantic and structural quality. Further, concept mapping relative to summary writing significantly improved the structural quality of integration. In addition, this study provides a new approach for evaluating different aspects of integration and contributes to multiple document comprehension literature from the perspective of concept network analysis.

Keywords: multiple documents, integration, concept network analysis

Funding support: The National Science Foundation (Award Abstract # 2215807), Roy B. Clariana (PI)

1 Introduction

Comprehending multiple documents is an indispensable skill both in school and even more so in the real world. "Multiple text comprehension refers to the processes and behaviors whereby students make sense of complex topics or issues based on information presented not within a single source but rather across multiple texts" (List & Alexander, 2017, p. 143). Even though multiple text comprehension is considered to be more difficult than comprehending a single document, paradoxically, some research has shown that students learn better when information is presented as separate sources presented in a non-systematic order compared to the same information integrated into an internally coherent textbook-like chapter with introductory and transitional clauses between sources (Wiley & Voss, 1996).

When reading multiple documents, students must identify and integrate the relevant information that is usually complementary such as with domain normative science content or that may actually conflict across documents such as with historical documents (Perfetti et al., 1999; Rouet & Britt, 2011; Wineburg, 1991). Specifically, the integration process can be influenced by *internal resources*, such as prior knowledge (Stang Lund et al., 2019), prior beliefs (Bråten & Stromso, 2010; Wiley et al., 2020), and literacy skills (Florit, Cain, & Mason, 2020) and by *external resources* such as multimedia information (Salmerón, Sampietro, & Delgado, 2020), presentation format (Salmerón, Gil, & Bråten, 2018; van Meter & Cameron, 2018), and the "quality" of documents (Bråten, Strømsø, & Britt, 2009; Stadtler et al., 2013). Because external resources are easier to manipulate by researchers and instructors compared with internal

resources, research on the influence of external resources can lead to potential interventions to support multiple document integration.

There is a growing research body on multiple document comprehension from the perspective of readers' mental models as concept networks. For example, in Knowledge Revision with Multiple Documents (KReC-MD) theory extension of the KReC single document comprehension model, Butterfuss and Kendeou (2021) note that "Like other discourse frameworks reviewed here, a core assumption of KReC is that knowledge is organized in a network of interconnected nodes. Nodes consist of concepts or propositions, and links represent the relations among these concepts" (p. 1484). Lehmann, Pirnay-Dummer, and Schmidt-Borcherding (2020) note that multiple document integration is characterized as a "dynamic process of linking, connecting, distinguishing, organizing, and structuring […] patterns, templates, views, ideas, theories, and visualizations" (Linn, 2000, p. 783).

Lehmann et al. (2020) used concept network analysis to treat a student's learning product as a complex and comprehensive personally structured model to more fully describe multiple document learning outcomes. Pre-service teachers read three domain-specific textbook excerpts each about 2,000 words long and were asked to write essays that were later analyzed using T-MITOCAR automated structural and semantic measures. In line with prior research, results indicated integrative focus questions versus no questions were effective prompts for structural integration but not for semantic integration. Structural knowledge integration was measured as GAMMA of the essay networks (link density calculated as the quotient of links per concept

within a graph) while semantic knowledge integration was measured as BSM of the essay networks (a measure of correct propositions). Further, the results of the structural and semantic measures did not support previous findings on the efficacy of argument writing versus generic nonspecific writing for knowledge integration.

Writing and concept mapping are forms of knowledge structure organization and externalization that "... by itself can lead to a modification of a learner's current mental model, and hence, improve understanding" (Lehmann et al., 2020, p. 910). And regarding creating and using visual displays when reading, "Organization involves inferring relations between and among pieces of information in an instructional message" (p. 628, McCrudden & Rapp, 2017). In the current study, we aimed to investigate the effects of reading prompts (integrative prompts vs. detailed prompts) and post-reading generative learning tasks (concept mapping vs. summary writing) on the conceptual level quantity and quality of multiple document integration via concept network analysis. Data for analysis consisted of the *quantity of integration* measured as the proportion of integrative links in the essay networks, the *semantic quality of integration* of the essay networks relative to the integrated referent model of the three texts, and the *structural quality of integration* measured as the Freeman's graph centrality of the essay networks.

1.1 Reading prompts

Students probably read with a specific but idiosyncratic purpose. According to the Multiple-Document Task-based Relevance Assessment and Content Extraction (MD-TRACE) model (Rouet & Britt, 2011), students construct mental representations called *task models* based

on their reading purposes and use it to guide the integration process, such as identifying and linking purpose-relevant information across documents (Anmarkrud et al., 2013). The Reading as Problem Solving (RESOLV) model further proposes that multiple documents can be seen as a problem-solving situation with rich information, where students must continuously update their task models to reach their goals (Rouet et al., 2017). According to the Goal-Focusing Model of text processing (McCrudden & Schraw, 2007), readers will establish and then continually update *standards of relevance* to help them focus on their purposes by discriminating between high- and low-relevant information, and then more resources will be allocated to high-relevant information. In this study, integration-focused relevance might help the formation of *standards of relevance* (McCrudden et al., 2010; McCrudden & Schraw, 2010), which then results in more integrative links in their mental representations (i.e., a better quantity of integration).

Previous studies of single-document reading comprehension have shown that students' reading process can be directed by giving purpose-related instructions (e.g., for entertainment versus for study, Linderholm & van den Broek, 2002), such reading prompts influence students to adopt an intended reading strategy (Bannert, 2009). For example, van den Broek et al. (2001) reported that readers with a study goal produced more coherence-building inferences, whereas readers with an entertainment goal produced more associations and evaluations. Reading for study compared with reading for entertainment enables students to elicit richer knowledge and perform deeper understanding (Bohn-Gettler & Kendeou, 2014; Micai et al., 2021). These reading strategies would likely influence the construction of mental representations and

subsequent learning (Cerdán & Vidal-Abarca, 2008). Such reading purpose interventions are effective for students ranging from primary through secondary schools (Cheong et al., 2019; Swanborn & de Glopper, 2002) and college (Latini et al., 2019; Stadtler et al., 2014).

But it is uncertain to what extent different reading purposes influence integration performance. In previous studies, two approaches were widely used to estimate students' integration performance in their learning products. One approach counts the total frequency of integration (i.e., *quantity* of integration) by calculating the number of integrative statements (both correct and incorrect) in a student's writing (Lehmann, Rott, & Schmidt-Borcherding, 2019; List et al., 2019). Another approach counts the frequency of correct integration or the correctness of integration (i.e., *quality* of integration) (Cheong et al., 2019; Latini et al., 2019). But neither of these approaches distinguishes both measures of the quantity and quality of integration. As a consequence, it is still unclear whether reading purposes improve both the quantity and the quality of integration, or only one of these.

Considering that there are diverse gaps between students and domain experts (e.g., the level of prior knowledge), reading purposes may only promote students to process *more* integrative knowledge (i.e., the quantity), but not *better* integrative knowledge (i.e., the quality). In this research, we distinguished the measures of the quantity and the quality of integration via concept network analysis to investigate how reading purposes influence different aspects of integration.

1.2 Post-reading concept mapping and summary writing as generative learning tasks

Besides reading prompts, multiple document integration might also be affected by post-reading activities. Previous multiple document studies have mainly focused on the direct effect of post-reading tasks on immediate integration performance, and have sometimes used post-reading tasks as an integration measurement tool (Gil et al., 2010a; Le Bigot & Rouet, 2007; Primor et al., 2021). Whereas these findings enable researchers to understand how integration can be enhanced quickly and immediately, few studies have explored how post-reading tasks affect subsequent integration performance (e.g., in a delayed task three days later). Considering the complex processing needed for multiple document integration, knowledge consolidation is probably important, and post-reading tasks may play the role of *generative learning* (Mayer, 1984; Wittrock, 1974) to facilitate the retention of integrative knowledge in long-term memory by providing the chance to re-construct it.

According to the Select-Organize-Integrate (SOI) model (Mayer, 2021; 2014), generative learning enables students to *select* knowledge from documents, *organize* it in the working memory, and *integrate* the constructed mental representation of documents with relevant prior knowledge from long-term memory. This generative process can be completed in either a verbal task such as writing (Doctorow, Wittrock, & Marks, 1978) or a task that consists of both verbal and visual information such as concept mapping (Schwamborn et al., 2010).

Both summary writing and argumentative writing tasks have been shown to support multiple document comprehension/integration, but Gil et al. (2010a) note "However, only a few

studies exist that directly compare the effects of summary and argument tasks on multiple-documents comprehension and integration" (p. 158) Argumentative writing has been shown in to be superior to summary writing with history documents (Wiley & Voss, 1996, 1999; Bigot & Rouet, 2007), while summary writing has been shown in to be superior to argumentative writing in other cases (Gil et al. 2010a, 2010b; Lehmann et al., 2020). List et al. (2019) reported contrary to their expectations that writing a research report on overpopulation (a summary writing task) prompted a greater degree of knowledge integration in students' written responses compared to writing an argument about overpopulation.

Whereas both concept mapping and summary writing in single-document comprehension studies have been shown to be effective generative strategies and both have obtained similar effect sizes (Cohen's $d = 0.62$ for concept mapping and Cohen's $d = 0.50$ for summary writing, see Fiorella & Mayer, 2016), concept mapping and summary writing are obviously very different tasks. A visuospatial task, such as drawing or concept mapping, requires the reader to generate their understanding of the structural conceptual relationships in and among documents to build a holistic structure that illustrates how core concepts from different documents link to each other (Schneider, 2012). Also, the visuospatial nature of a concept map provides a concise conceptual placeholder for adding information from succeeding documents, while summary writing may be better for within-document but less effective at cross-document processing due to its linear-sequential nature (Leopold & Leutner, 2012). For example, List et al. (2019) reported that students written responses most commonly consisted of separate document representations, only

about a third wrote responses that reflected some knowledge integration and the amount of integration evidenced was quite limited ($M = 2.14$). Jonassen and Kim (2010) ask "Why are students so inept at constructing arguments?" and "Why do students argue with apparent blinders on?" then listed three likely causes: (1) teachers do not foster argumentation, (2) external pressures to cover content leaving no time for other skill development; and (3) deficient prior knowledge of the domain (i.e., argumentation requires a strong content foundation).

In contrast to summary writing which emphasizes verbal re-statement, concept mapping enables students to illustrate and visualize their understanding of documents by drawing a concept map that links core concepts from documents into a single structure (Author_03, 2006; Kim, 2012). Recent findings suggest that generative concept mapping might be effective for promoting multiple document integration and learning (Barzilai et al., 2023; List, 2019). However, few studies have compared *summary writing* which tends to recapitulate the individual documents consecutively with *concept mapping* which promotes integration by capturing a holistic structure across documents. For example, concept mapping as a form of note-taking enables students to continually add concepts and links to a developing map artifact (Gurlitt & Renkl, 2010; Schwendimann, 2014) while summary writing is linear-sequential due to its nature and so writing may be more likely fixed within one document (due to primacy or recency). Thus, although summary writing has been shown to be a little better than concept mapping for single document learning (Author_07, 2015; 2017), concept mapping may be better for improving the structural quality of integration across documents.

1.3 Concept network analysis

Concept network analysis was developed and widely applied in single document comprehension studies for the past two decades (Author_02, 2004; Ifenthaler, 2014; Kim & McCarthy, 2021) and has the potential to measure different aspects of integration. Although different theories have emphasized various dimensions of comprehension, most theories have assumed the existence of a connectionist representational mental architecture, with nodes representing concepts and edges representing the links among the concepts (McNamara & Magliano, 2009). By measuring the external re-representation of a mental representation extant in a production task such as writing, researchers can explore the characteristics of the nature and quality of inner mental representation using concept network analysis (Jonassen et al., 1993; Kintsch, 1988). This analysis can be performed automatically and effectively by computer and show stable internal consistency reliability and acceptable validity (Author_03, 2006; Kim, 2012).

In multiple document learning, links between concepts (i.e., edges in a concept network) derived from different documents can be used as the indicator of conceptual level integrative knowledge, while links that consist of concepts derived from a single document refer to intra-text knowledge. Considering the difference in total link numbers among concept networks, the *quantity* of integration can be calculated as the proportion of integrative links by dividing the number of integrative links by the total number of links in a concept network.

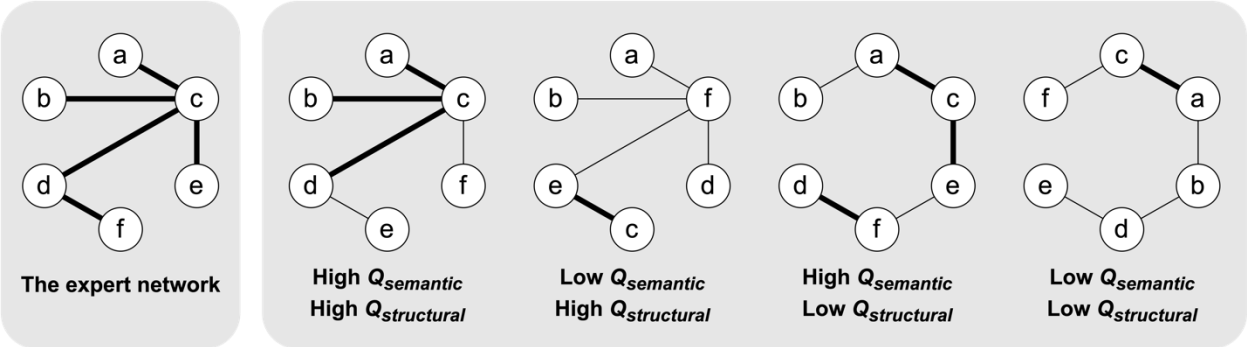

Figure 1. An example expert network (left side) and four example concept networks (right side) with the same *quantity* of integration (i.e., five integrative links) but differ in either the *semantic quality* (see $Q_{semantic}$) or the *structural quality* of integration (see $Q_{structural}$). The nodes represent concepts from different documents and the links represent relationships among them. The bold lines indicate links that are consistent with the expert network.

Note that the quantity of integration is not equal to the quality. As shown in Figure 1, even if students obtain the same quantity of integration in their mental representations (represented by concept networks), the semantic and structural quality (network form) of their integrative knowledge can be different. Previous studies have shown that it is difficult for a novice to perform expert-level strategies to process cross-document semantic statements (von der Muhlen et al., 2016; Wineburg, 1991). Recent studies also show that only a few novices elaborated on integrative statements in their summaries and even then the extent of integration is limited (Du & List, 2020; List & Du, 2021). Hence, the *semantic quality* of integration can be measured by comparing the similarity of integrative links between a student's concept network and the expert network (Author_06, 2018; Ifenthaler, 2014).

Another aspect of integration quality that previous studies have scarcely explored is whether readers can construct an accurate (i.e., expert-like) and coherent single structure of the

documents they have read measured here as the *structural quality* of integration). By representing students' learning products as concept networks, their holistic structural level of integration can be calculated as network graph centrality (GC) (Author_04, 2015; Author_07, 2017; Freeman, 1978). This GC value can be used to compare concept networks descriptively and inferentially when the networks are about the same size and are sparsely connected (i.e., where the average ratio of links per concept is about 1:1 or is at least less than 2-to-1). The structural quality of integration can be measured by comparing the GC similarity between a student's concept network and the expert network (Kim & McCarthy, 2021).

1.4 Research purpose and hypotheses

Integrating knowledge from multiple documents in an expert-like way is a complex cognitive process. The current study aimed to explore how different reading prompts (integrative purpose vs. detail purpose) and post-reading generative learning tasks (concept mapping vs. summary writing) assist undergraduates with integrating knowledge from three documents on the topic of Alzheimer's disease. By using concept network analysis, we analyzed students' integration performance in delayed writing by estimating three integration indicators, including the *quantity* (measured by the proportion of integrative links), the *semantic quality* (measured by the similarity of integrative links), and the *structural quality* of the integration (measured by the similarity of graph centrality). As supplementary measurements, students' prior knowledge about the topic of Alzheimer's disease was assessed before the experiment, and their comprehension of integrative knowledge was measured after the experiment. Following the findings from Lehman

et al. (2020), we hypothesized that the integrative purpose prompts compared to the detailed purpose prompts improve the quantity of integration (H1a) as well as the semantic and structural quality of integration (H1b). We also hypothesized that the concept mapping task compared to the summary writing task enables students to establish better structural quality of integration (H2). Finally, we hypothesized that an "integrative purpose + concept mapping" condition (i.e., an interaction effect) enables students to perform better integration compared with other conditions (H3).

2 Methods

2.1 Participants

One hundred and thirty-six undergraduate students were randomly assigned to one of four conditions. These four conditions were: *integrative purpose* + *concept mapping*, *integrative purpose* + *summary writing*, *detailed purpose* + *concept mapping*, or *detailed purpose* + *summary writing*. Participants came from a wide range of majors, such as economics, computer science and biology. None of the participants majored in psychology or medicine or had engaged in study or work on Alzheimer's disease, and thus only non-experts participated in the experiment.

To ensure participants have not read similar documents before, participants were asked two additional questions as follows. The first question was "whether you had read materials with contents similar to the experimental documents", and if the answer was yes, they were required to answer the second question "how do you think the level of similarity between the contents"

(rate on a five-point scale, 1 = totally dissimilar, 5 = totally similar). Participants who answered "yes" to the first question were excluded (n = 8, rating of the content similarity: M = 2.63, SD = 0.52). In addition, based on the pilot study, participants who completed the first section tasks within 1 minute were deemed hasty and subsequently excluded from participating in the second session (n = 9). Thus, data from one hundred and nineteen university students were included in the statistical analysis (78 females, age: M = 20.63, SD = 1.97).

A prior power analysis was conducted via the *Webpower* package in R for sample size estimation (Zhang & Yuan, 2018). We focused on the main effect of prompts in a two-way ANOVA and this effect showed a large effect size in the literature (Bohn-Gettler & Kendeou, 2014; Cerdán & Vidal-Abarca, 2008; Lehmann et al., 2020). With Cohen's f = 0.40, α = 0.05, and power = 0.80, the minimum sample size needed is 64. Thus, the obtained sample size of 119 was more than adequate to test the hypotheses.

The data collection in the current research was reviewed and approved by the local university ethics committee. All participants provided their written consent before the investigation.

2.2 Materials

2.2.1 Multiple documents, core concepts and the expert network

Three documents on the topic of Alzheimer's disease were used as reading materials. These documents were derived from the website of the National Institute of Health (NIH, 2021). According to the description on the website, these documents were designed for high school,

undergraduate, and graduate students, and the contents were trustworthy. Documents were translated into Chinese by three graduate students; and then two domain experts, who were medical graduates, worked together to check the contents to determine the final version of documents, identify core concepts, and then write an expert integrative summary.

A separate independent sample (N = 36, 14 females; age; M = 23.36, SD = 5.27) was arranged to examine the content coherence ("How do you think the content of this document is coherent", 1 = very incoherent, 5 = very coherent) and readability ("How easy do you think the content of this document is to understand", 1 = very difficult, 5 = very easy). Repeated measures ANOVA showed no main effects of content coherence, $F_{(2,70)} = 1.83$, $p = 0.169$, $\eta^2 = 0.050$, or readability among three documents, $F_{(2,70)} = 0.36$, $p = 0.703$, $\eta^2 = 0.010$. The document characteristics were shown in Table 1.

Table 1. Summary of documents.

	Brain changes, symptoms and diagnosis of Alzheimer's disease	Alzheimer's disease genetics	Alzheimer's and Down Syndrome and related clinic studies
Word count	596	500	613
Core concepts count	14	14	13
Links count	24	25	21
Graph centrality (GC)	0.41	0.49	0.37
Content coherence	3.77 (0.62)	3.75 (0.69)	3.56 (0.74)
Readability	3.64 (0.90)	3.72 (0.88)	3.61 (0.93)

Thirty-nine core concepts were identified from the documents by the experts, with thirteen or fourteen from each document. These concepts were verbs, nouns, or nouns modified

by adjectives. Core concepts among documents were unique except for a common concept *Alzheimer's disease*. The experts then selected thirteen core concepts from the thirty-nine ones to write an integrative summary to reflect the gist of three documents. The integrative summary was converted to the concept network and used as the expert network (see Figure 2). Detailed information on concept network conversion is provided in section 2.4.

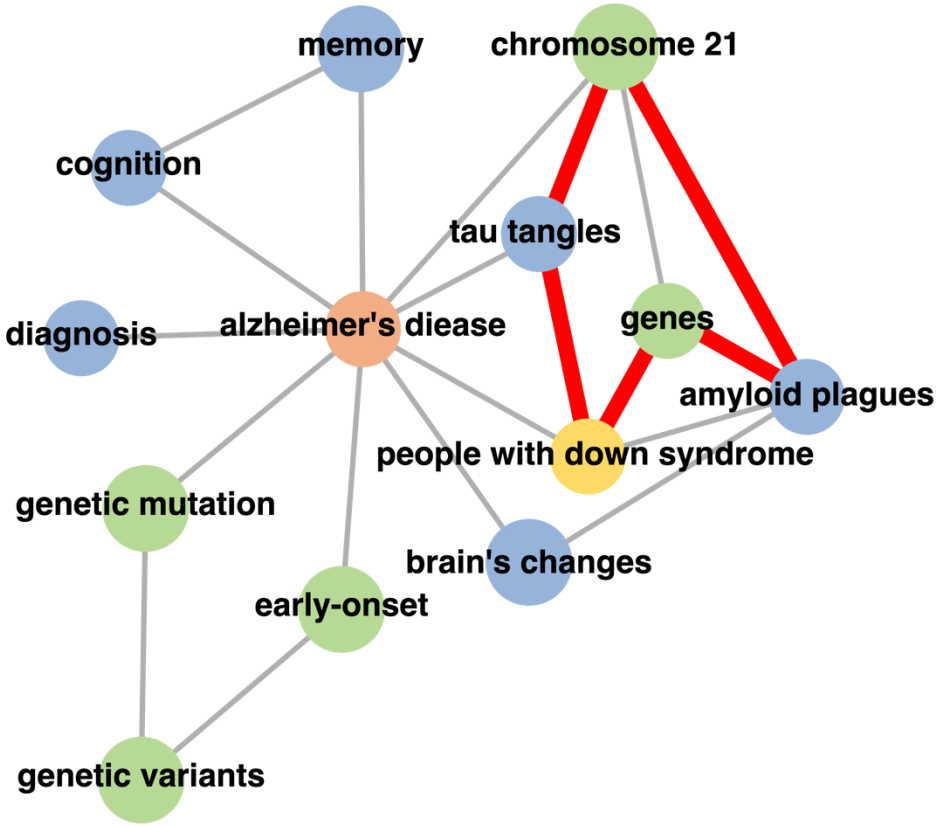

Figure 2. This network of the expert's integrative summary has a 0.58 GC value (concepts shown in English). Concepts from the three documents were represented as blue, green and yellow nodes, respectively. The orange node was the common central concept *Alzheimer's disease*. The bold red lines indicate integrative links.

2.2.2 Alzheimer's Disease Knowledge Scale

The well-established Alzheimer's Disease Knowledge Scale (ADKS) was used to measure topic-related prior knowledge (Carpenter et al., 2009). A separate independent sample (N = 60, 47 females; age: M = 23.32, SD = 2.47) was used to conduct the item analysis for the Chinese version of ADKS. The final scale contained sixteen items in a verification format (i.e., true/false) and Cronbach's α was 0.65. The total scores of ADKS were used as the participants' level of prior knowledge.

2.2.3 The reading prompts

Two reading prompts as focus questions were developed (see Table 2) to establish a specific reading purpose (Lehmann et al., 2019; McCrudden & Schraw, 2007). The integrative prompts were four questions to assist integration, while the detailed prompts focused on specific details such as the number of paragraphs, which would elicit a little less integrative information. The prompts were presented before and during the reading task as well as during the post-reading generative learning tasks.

Table 2. Reading prompts in the form of focus questions (shown in English).

Prompts for the integrative purpose	Prompts for the detailed purpose
(1) How can you relate these three documents to each other through the concept *Alzheimer's disease*?	(1) How many paragraphs are there in these three documents?
(2) Can you find concepts related to different documents?	(2) Are there any rarely-used words in these three documents that you don't know?
(3) Can you find statements in a document that can be used to explain the contents presented in the other documents?	(3) What do you think of the readability of the three documents?
(4) What overall conclusion can you derive from these three documents?	(4) What do you like about these three documents?

2.2.4 The reading task and post-reading generative learning tasks

The reading task and post-reading generative learning tasks (i.e., concept mapping and summary writing) were designed and conducted via *PsychoPy-3* (PsychoPy version: 2021.2.3; Python version: 3.6.8). The interface of the reading task referred to both *Read&Answer* software (Cerdán et al., 2018; Vidal-Abarca et al., 2011) and the multiple documents comprehension test software (Hahnel et al., 2019).

In the reading task, documents and instructions were placed on either the left or the right side (see the top of Figure 3). The contents of each document were masked in the default stage and could be accessed by moving the mouse cursor over each paragraph. Only one document was presented at a time and participants could switch documents by pressing buttons in the middle of the interface.

In the concept mapping task, participants needed to select thirteen core concepts to draw concept networks that reflected their mental representations (see the middle of Figure 3). There was a canvas surface and a concept list on the interface. Participants constructed their maps by dragging concepts from the list to the canvas and adding links between them. Removing or re-adding concepts or links was allowable. However, the concept *Alzheimer's disease* was compulsory to use and it would be presented on the canvas at the beginning of the task and was not allowed to be removed.

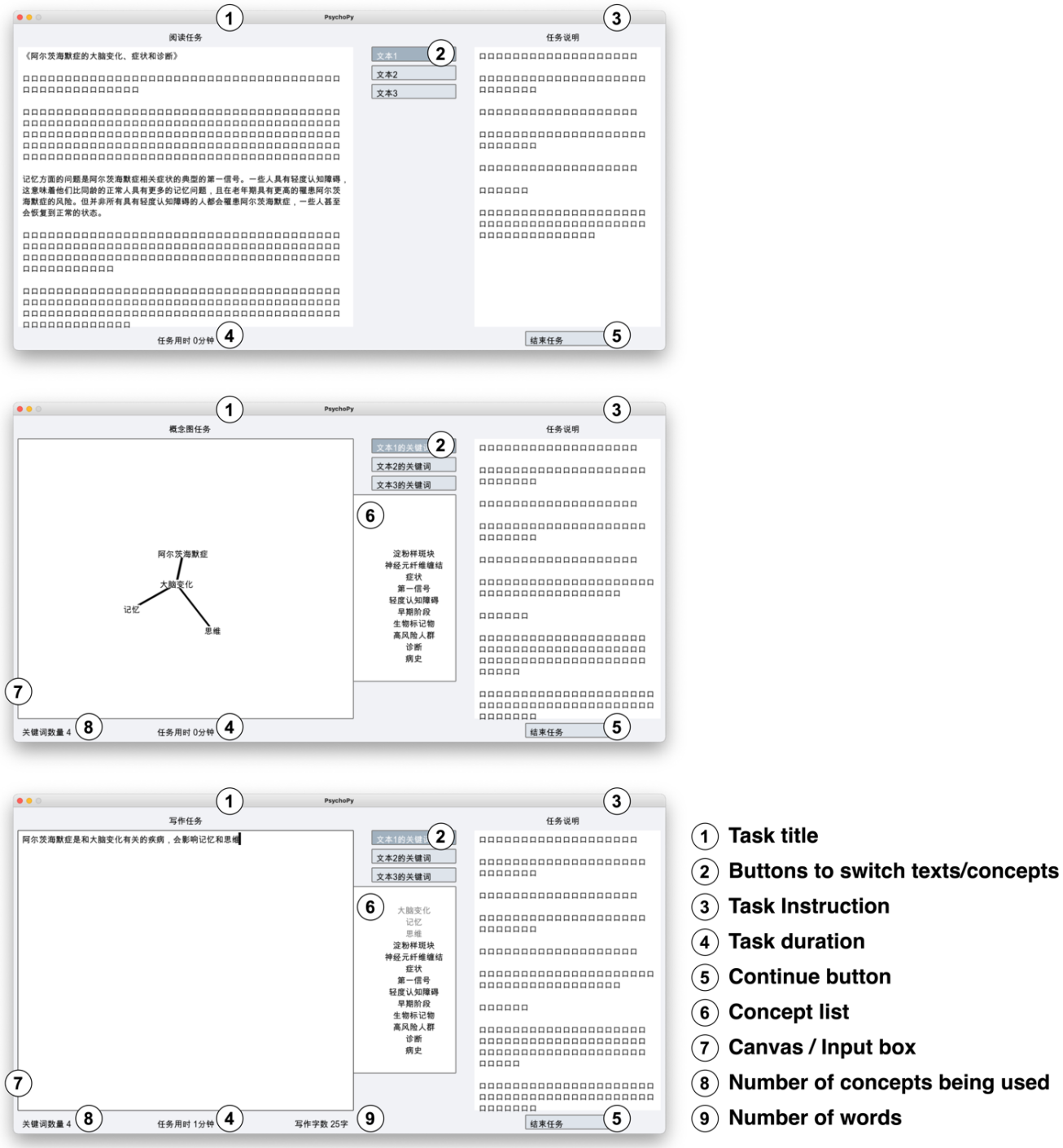

Figure 3. The user interfaces of the reading task (top), the concept mapping task (middle), and the writing task (bottom). The instructions for the tasks were shown in the appendix.

In the summary writing task, participants used thirteen core concepts in the concept list, including the compulsory concept *Alzheimer's disease,* to write summaries that reflected their mental representations (see the bottom of Figure 3). Several details were manipulated so that the summary writing task and the concept mapping task could be matched as much as possible. First, the compulsory concept *Alzheimer's disease* was on the input box at the beginning of the writing task. Second, to help students to use all the terms the colors of core concepts in the list would become gray when they were selected. Note that participants were informed that they can use the selected concepts repeatedly in their writing.

2.2.5 Inference verification test

The inference verification test (IVT) was used to measure comprehension and performance in recalling and recognizing integrative knowledge (Cerdán & Vidal-Abarca, 2008; Cheong et al., 2019; Lehmann et al., 2019; Royer et al., 1979). Similar to the manipulations in the literature (Mahlow et al., 2020; Schoor et al., 2020), we developed an IVT covering three theory-based cognitive requirements: *collaboration* (Wineburg, 1991), *integrated situation model*, and *intertext model* (Rouet & Britt, 2011). All items of the IVT were administered in a verification format (i.e., true/false). The same separate independent sample in the pilot study of ADKS (Chinese version) was used to conduct this item analysis ($N = 60$). The final scale contained eleven items (i.e., six for *collaboration*, three for *integrated situation model*, and two for *intertext model*), and Cronbach's α is 0.69. Because the item number and reliability of each

subtest question type were not sufficient, we only analyzed overall IVT performance. The total scores of IVT were calculated as the integrative knowledge comprehension.

2.3 Procedure

There were two sessions in this investigation (see Figure 4). In the first session, participants were randomly assigned to one of four groups after signing up for the experiment and accomplishing the ADKS prior knowledge test. They received either the integrative prompts or the detailed prompts and then completed the reading task.

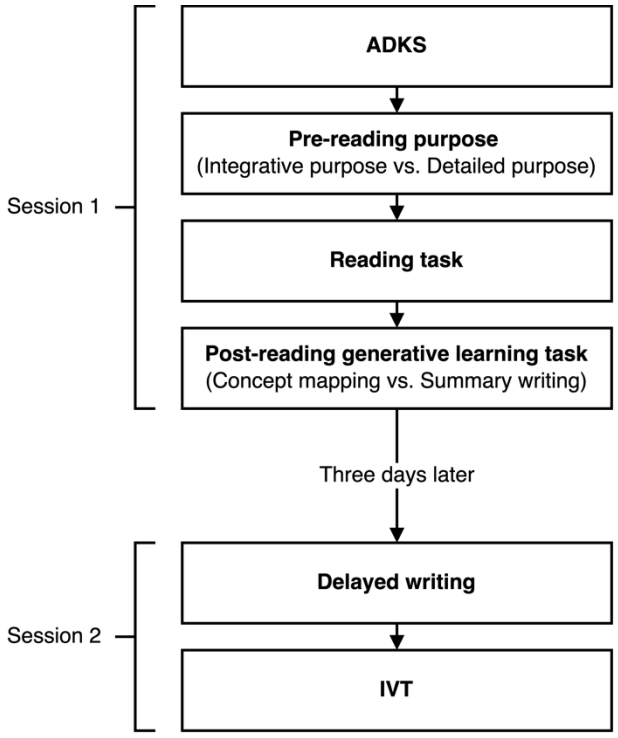

Figure 4. The experimental procedure.

Next, each participant was required to select thirteen of thirty-nine core concepts to construct the gist of three documents with either a concept map (i.e., the concept mapping task)

or a 300-word summary (i.e., the summary writing task). A brief introduction to concept mapping combined with an example was given before the concept mapping task so that participants could understand how a concept map is created. The first session took approximately 25 mins.

The second session was arranged to collect delayed measures three days after the first session (Kang et al., 2007; van Peppen et al., 2021). All participants wrote delayed essays that reflected the content of their generative products in the first session. At this stage, each participant received a specific concept list showing the thirteen selected concepts in the first session. Finally, participants completed the IVT, two additional questions about similar materials reading experience, and a demographic questionnaire.

2.4 Data analysis

Concept networks of the delayed writing were established using the analysis of lexical aggregates (ALA; Author_05, 2009) with the PFNet algorithm using r = infinity and $q = n - 1$ (Schvaneveldt et al., 1988; Schvaneveldt et al., 1989). In brief, in ALA, the co-occurrences of preselected concepts in a linear pass through the text are used to identify term-term associations, the resulting network edges then consist of links between these pairs of concepts. The quantity, the semantic quality, and the structural quality of integration for each participant's essay concept network were evaluated as follows.

The *quantity of integration* was measured by calculating the proportion of integrative links. For example, if a network contains 20 links in total and 5 are integrative links, the proposition of integrative links will be 5/20 = 0.25).

The *semantic quality* of the network was measured by calculating the similarity of integrative links (see Equation 1; Tversky, 1977). In Equation 1, A and B indicate the subset of integrative links from either a participant's concept network or the expert network, and the default value of α and β is 0.5, respectively (Pirnay-Dummer & Ifenthaler, 2011). When α = β = 0.5, this similarity is also called network overlap (Author_05, 2009).

$$S_{semantic} = f(A \cap B) / [\alpha * f(A - B) + \beta * f(B - A) + f(A \cap B)] \quad (1)$$

The *structural quality* was measured by calculating GC similarity. Graph centrality (GC) can be measured by calculating the degree centrality of each node in a concept network at first while accounting for total degrees of freedom (see Equation 2) and then aggregating total degree centralities to a single value (see Equation 3).

$$C_D(v) = \deg(v) / (n - 1) \quad (2)$$

$$C_D(G) = \sum^{v}_{i=1} [max(C_D(v_i)) - C_D(v_i)] / (n - 2) \quad (3)$$

Where v indicates a vertex (i.e., node), $max\ (C_D(v_i))$ indicates edge numbers of the node with the greatest centrality, and n indicates the total number of vertices in a concept network. Note that GC data are probably NOT additive, although Kim and McCarthy (2021) have shown the potential and utility of comparing the graph centrality of two networks to show how novice networks tend to become more like the expert. In their approach, GC similarity can be calculated

by comparing GC values between a participant's concept network and the expert network (see Equation 4; Kim & McCarthy, 2021).

$$S_{GC} = 1 - abs(C_D(G_1) - C_D(G_2)) / max(C_D(G_1), C_D(G_2)) \quad (4)$$

For example, given the GC of two networks as $C_D(G_1) = 0.20$ and $C_D(G_2) = 0.30$, then $S_{GC} = 1 -$ (absolute value of $(0.20 - 0.30) / 0.30$) $= 1 - (0.10 / 0.30) = 1 - 0.33$, thus $S_{GC} = 0.66$, larger S_{GC} means greater similarity to the expert referent network.

3 Results

Regarding pretreatment group equivalence, results of a one-way ANOVA of topic-related prior knowledge showed no significant difference in prior knowledge among the four groups, $F_{(3,115)} = 1.92$, $p = 0.13$, $\eta^2 = 0.05$, indicating appropriate random assignment to groups. The mean accuracy of ADKS was 0.65, $SD = 0.13$. In the primary analysis below, we first conducted ANCOVA with ADKS scores as the covariate but found that it did not influence the results, so we did not include the ADKS as a covariate in the following analyses.

3.1 Comprehension of integrative knowledge

Differences in integrative knowledge comprehension among groups were explored via a 2 (reading prompts) * 2 (post-reading generative learning tasks) ANOVA of the inference verification test (IVT) scores as the dependent variable. Results did not show any main effect of reading prompts, $F_{(1,115)} = 0.03$, $p = 0.86$, $\eta^2 = 0.01$, the main effect of post-reading tasks, $F_{(1,115)} = 0.11$, $p = 0.74$, $\eta^2 = 0.01$, nor the two-way interaction, $F_{(1,115)} = 2.94$, $p = 0.09$, $\eta^2 = 0.03$. This

indicated that participants in different groups perform equally in recalling and recognizing integrative knowledge (see Table 3).

Table 3. Descriptive statistics for the IVT scores.

Group	*M*	*SD*	*n*
Integrative purpose + Concept mapping	0.82	0.12	31
Integrative purpose + Summary writing	0.86	0.11	32
Detailed purpose + Concept mapping	0.85	0.11	27
Detailed purpose + Summary writing	0.82	0.15	29

3.2 The quantity and quality of integration in essay concept networks

We next analyzed the differences in the quantity, the semantic quality, and the structural quality of integration in participants' concept networks of their essays via 2 (reading prompts) * 2 (post-reading generative learning tasks) ANOVAs. For the semantic and structural quality of integration, results of Levene's test indicated unequal variances among groups and thus White-corrected *F* and *p* values as proposed by Long and Ervin (2000).

For the quantity of integration, results showed a significant main effect of reading prompts, $F_{(1,115)} = 8.21$, $p = 0.01$, $\eta^2 = 0.06$ (see Figure 5a), supporting H1a. Participants who received the integrative purpose ($M = 0.25$, $SD = 0.14$) preserved significantly more integrative links in their delayed writing than the ones who received the detailed purpose ($M = 0.17$, $SD = 0.15$). There was also a significant main effect of post-reading tasks, $F_{(1,115)} = 4.36$, $p = 0.04$, $\eta^2 = 0.03$. Participants in the concept mapping condition ($M = 0.24$, $SD = 0.15$) preserved more integrative links in their delayed writing than those in the summary writing condition ($M = 0.18$,

SD = 0.14). However, no interaction was found, $F_{(1,115)} = 0.01$, $p = 0.97$, $\eta^2 = 0.01$, thus H3 was not supported.

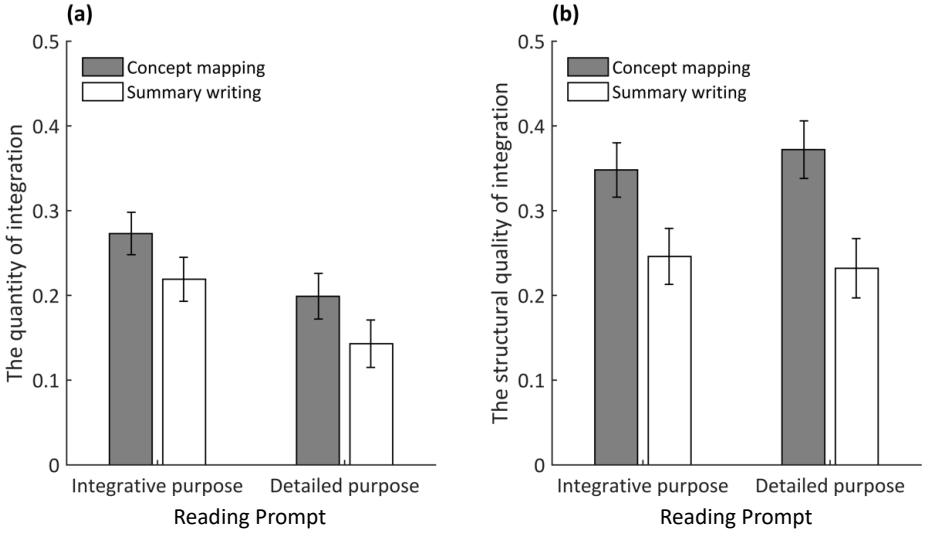

Figure 5. Results of the quantity and the structural quality of integration.

For the semantic quality of integration, results did not show a main effect of reading prompts, $F_{(1,115)} = 0.01$, $p = 0.95$, $\eta^2 = 0.01$, thus supporting H1b and neither the main effect of post-reading tasks, $F_{(1,115)} = 2.00$, $p = 0.16$, $\eta^2 = 0.01$ nor the two-way interaction were not significant, $F_{(1,115)} = 0.35$, $p = 0.55$, $\eta^2 = 0.01$, thus H3 was not supported.

For the structural quality of integration, results showed the main effect of reading prompts was not significant, $F_{(1,115)} = 0.13$, $p = 0.72$, $\eta^2 = 0.01$, thus supporting H1b, the two-way interaction effect was not significant, $F_{(1,115)} = 0.29$, $p = 0.59$, $\eta^2 = 0.01$, thus H3 was not supported. However, there was a significant main effect of post-reading tasks, $F_{(1,115)} = 6.76$, $p =$

0.01, $\eta^2 = 0.10$ (see Figure 5b) thus supporting H2. The results showed as expected that participants in the concept mapping condition ($M = 0.36$, $SD = 0.24$) had a significantly greater expert-like structure in their delayed writing compared with the summary writing condition ($M = 0.24$, $SD = 0.09$).

To summarize, these results indicated that both reading prompts and post-reading tasks affect the *quantity of integration* (supporting H1a), and post-reading tasks even affect the *structural quality of integration* (supporting H2). But it seems like these two factors independently influence integration (i.e., no interaction effect, did not support H3). However, neither reading prompts nor post-reading tasks affect the *semantic quality of integration*.

3.3 Average networks

Group-average networks that reduce individual idiosyncratic error have been shown to have probative value for *post-hoc* analysis (Author_01, 2022). We generated each group's average network by following three steps: First, each participant's concept network was represented as a 39*39 proximity matrix. Second, an average network with 39 concepts was generated for each group by averaging across the matrix cells and then generating a network of this averaged matrix, this group average network contained the links that were presented in most participants' networks (Schvaneveldt, Durso, & Dearholt, 1989). Finally, a final average network was attained based only retaining thirteen concepts with the greatest node centrality (links), while the other concepts and links were excluded.

As shown in Figure 6, the common concept *Alzheimer's disease* was kept central in all four average networks, indicating that all participants obtained the topic of three documents. In addition, the group-average networks also showed that most participants tended to process concepts in the first and the second document, which both described *basic domain knowledge*, and also combined a central concept from the third document (i.e., "people with Down syndrome"), which described the *application* of Alzheimer's disease studies in a specific population. This pattern was similar to the expert network (see Figure 2).

The average networks of both the *integrative purpose + concept mapping* group and *integrative purpose + summary writing* group retained similar integrative links (i.e., "people with Down syndrome - chromosome 21", and "chromosome - chromosome 21", see the bold red lines in Figure 6a and 6b), while there were no integrative links in the average networks of the *detailed purpose + concept mapping* group and *detailed purpose + summary writing* groups (see Figure 6c and 6d). The average networks of the two concept mapping groups (see Figure 6a and 6c) were more alike (63% overlap), suggesting the likely influence of concept mapping on delayed writing.

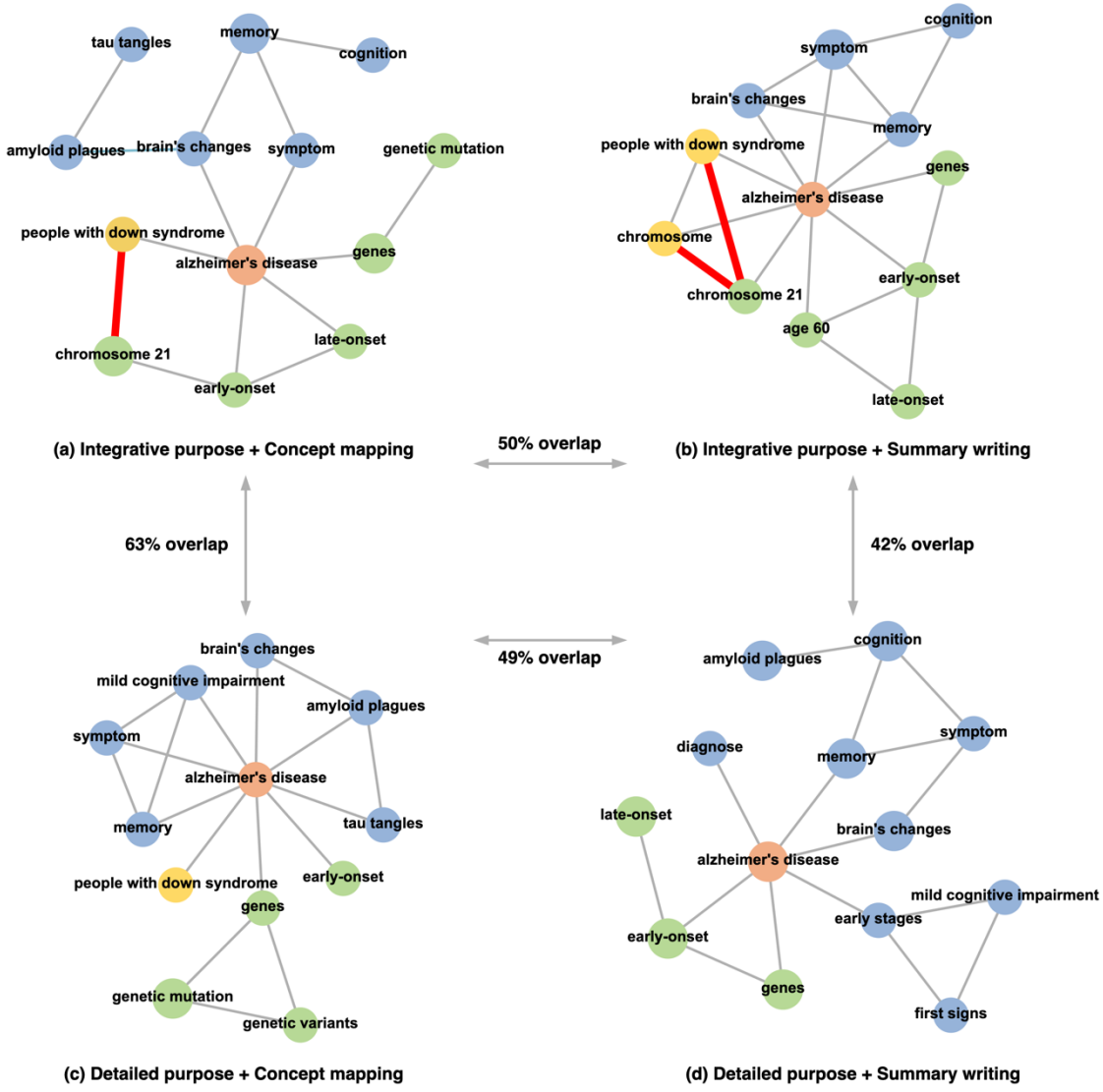

Figure 6. Average networks of the four groups. Concepts from the three documents were represented as blue, green, and yellow nodes, respectively. The orange node is the common central concept *Alzheimer's disease*. The bold red lines indicated integrative links.

4 Discussion

This experimental study examined how reading prompts and post-reading generative learning tasks alter readers' conceptual knowledge structure in multiple document integration. Regarding integrative knowledge comprehension under different purposes and tasks, there were

no significant results in the IVT scores, however, it has been suggested that the IVT scores as a *receptive measure* (Primor & Katzir, 2018) might not be sensitive enough to represent deep learning, rhetorical skills, or expressing ideas combined with prior knowledge. Meanwhile, delayed writing as an *expressive measure* before the IVT might enhance participants' comprehension and thus eliminate the gap among groups (Anderson & Thiede, 2008; Thiede & Anderson, 2003). Hence, participants' integration performance can be better interpreted based on the results of the delayed writing. By representing the delayed writing as concept networks, we found that the integrative purpose led to a stronger quantity of integration than the detailed purpose, with a median effect size ($\eta^2 = 0.064$). In addition, the post-reading concept mapping task helped improve the quantity and the structural quality of integration compared with the summary writing task. However, the semantic quality of integration under different manipulations was not significantly different.

We explain the results of the *quantity* of integration based on readers' goals and strategies to process purpose-relevant information. When reading multiple documents, readers will intuitively identify across-document information with different levels of relevance to their reading purposes (Anmarkrud et al., 2013). Previous studies have indicated individual differences in this ability (Vidal-Abarca, Mañá, & Gil, 2010), but individual differences might be compensated by providing appropriate reading prompts (Model of text processing, McCrudden & Schraw, 2007). In this investigation, integration-focused relevance might help the formation of *standards of relevance* and thus prompt participants to read across-document information longer

(McCrudden et al., 2010; McCrudden & Schraw, 2010), which then results in a better quantity of integration. The function of standards of relevance is in line with the MD-TRACE model, which functions to direct readers to select and retain the most important information (Britt & Rouet, 2020). In addition, our findings on the quantity of integration, which are based on the proportion of integrative links, are consistent with findings using human coding approaches (e.g., McCrudden et al., 2022; Primor et al., 2021).

How the *quality* of integration can be affected by pre- and post-reading factors was the second main research question in this study. Results showed that the quality of integration during multiple document learning cannot be improved by the integrative purpose. This implies that these reading prompts have a limited effect on eliminating the gap between novices and domain experts, even when it promotes more integration processing. On the other hand, results showed that concept mapping as post-reading generative learning improves the structural quality of integration compared with summary writing. Concept mapping as a graphic organizer can be considered a spatial strategy that helps students create a spatial mental arrangement of the core concepts (Fiorella & Mayer, 2017).

Cognitive load may be important in multiple document integration (Cerdán et al, 2018). We explain the effects of concept mapping in enhancing the structural quality of integration by its potential role of reducing cognitive load in multiple document learning. First, concept mapping may serve as a reduction strategy by only including the most important concepts in the maps (Hilbert & Renkl, 2008; O'donnell, Dansereau, & Hall, 2002). Such a strategy may

highlight macrostructural knowledge, which represents the top-level information in materials such as the global structure of documents (Kintsch, 1998). Second, concept mapping may also reduce cognitive load by enabling simultaneous processing from both verbal and visuospatial channels (dual coding). Unlike summary writing which only encodes document content in verbal form, concept mapping enables dual coding by presenting concepts in verbal form and relationships among concepts in visual form. As both verbal and visuospatial processing take up different resources of working memory (Baddeley, 1992), readers may process and store knowledge from multiple documents better when drawing concept maps. This could also explain why our results showed that concept mapping significantly improves the quantity of integration compared with summary writing.

Unexpectedly but acceptably, we did not find that the semantic quality of integration was affected by either reading prompts or post-reading generative learning tasks. We explain these findings based on the Knowledge Revision with Multiple Documents (KReC-MD) theory (Butterfuss & Kendeou, 2021), which emphasized the importance of updating knowledge to develop a well-constructed mental representation. Such knowledge revision may be performed by providing materials including external expert-level information such as an expert network or outline (Author_01, 2022). This assumption should be considered in future studies.

5 Limitations and implications for future research

There are several limitations and implications for future research. First, only summary writing was used in contrast to argument writing which has been shown in some cases to prompt

integration when reading multiple documents (Bigot & Rouet, 2007; Lehmann et al., 2019; Wiley & Voss, 1999). Note that previous research on any advantage of argumentative writing may be related to the content area (e.g., history content lends itself to argumentation) and also reader characteristics, for example, high prior knowledge benefits most from writing an argument (Gil et al., 2010b) and in fact many students are very weak at argumentative writing (Jonassen & Kim, 2010). Hence, it will be important for future research to compare how concept mapping, summary writing, and argumentative writing can differentially influence multiple document integration when controlling for prior knowledge.

Second, we used a detailed purpose prompt as a control condition to better understand the influence of integrative prompts on multiple document integration in essays. But a detailed purpose prompt might actually distract students from actual integration. Further study should include a "no specific purpose" condition or even a “no prompt” condition to further consider the influence of integrative reading prompts.

Third, the essay prompt included a list of concepts taken from the three texts and required participants to use a minimum number of these concepts in their essays. This manipulation aimed to ensure participants' networks are comparable with the expert network (see the description of GC in section 1.3) and to hopefully scaffold writing to improve essay quality. Further research should consider how different writing prompts like this can alter readers' integration (e.g., different number of required concepts, or mapping without a concept list).

Finally, we mainly focused on conceptual level integration (i.e., concept-concept links), but other integration indices such as source-source links (Mason et al., 2018) and links between documents and background knowledge (Anmarkrud et al., 2013) can also be important in multiple document comprehension and thus should be further explored in future research.

6 Conclusion

Overall, a concept network analysis was employed to describe how different integrative reading prompts and post-reading generative learning tasks influence the conceptual level integration of multiple documents in different aspects. The quantity of integration (measured by the proportion of integrative links), the semantic quality (measured by the similarity of integrative links) and the structural quality of integration (measured by the comparison of graph centrality) were evaluated. The findings showed that the quantity and the structural quality of integration could be enhanced by using integrative reading prompts and post-reading concept mapping, while the semantic quality of integration did not improve. In addition, the concept network analysis used here provides a new approach for evaluating mental representations of multiple documents, this approach can be used as a new complementary measure of multiple document comprehension in future investigations.

References

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*(1), 110-118. https://doi.org/10.1016/j.actpsy.2007.10.006

Anmarkrud, Ø., McCrudden, M. T., Bråten, I., & Strømsø, H. I. (2013). Task-oriented reading of multiple documents: Online comprehension processes and offline products. *Instructional Science*, *41*(5), 873-894. https://doi.org/10.1007/s11251-013-9263-8

Baddeley, A. (1992). Is working memory working? The Fifteenth Bartlett lecture. *Quarterly Journal of Experimental Psychology*, *44* (1), 1-31. https://doi.org/10.1080/14640749208401281

Bannert, M. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychologie, 23*(2), 139-145.

Barzilai, S., Tal-Savir, D., Abed, F., Hagani-Mor, S., & Zohar, A. (2023). Mapping multiple documents: From constructing multiple document models to argumentative writing. *Reading and Writing*, *36*, 809-847. https://doi.org/10.1007/s11145-021-10208-8

Bigot, L. L., & Rouet, J.-F. (2007). The impact of presentation format, task assignment, and prior knowledge on students' comprehension of multiple online documents. *Journal of Literacy Research*, *39*(4), 445-470. https://doi.org/10.1080/10862960701675317

Bohn-Gettler, C. M., & Kendeou, P. (2014). The interplay of reader goals, working memory, and text structure during reading. *Contemporary Educational Psychology*, *39*(3), 206-219. https://doi.org/10.1016/j.cedpsych.2014.05.003

Bråten, I., Britt, M. A., Stromso, H. I., & Rouet, J. F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist*, *46*(1), 48-70. https://doi.org/10.1080/00461520.2011.538647

Bråten, I., Strømsø, H. I. (2011). Measuring strategic processing when students read multiple texts. *Metacognition Learning*, *6*, 111-130. https://doi.org/10.1007/s11409-011-9075-7

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly*, *44*(1), 6-28. https://doi.org/10.1598/RRQ.44.1.1

Britt, M. A., & Rouet, J. (2020). Multiple document comprehension. In *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.867

Butterfuss, R., & Kendeou, P. (2021). KReC-MD: Knowledge revision with multiple documents. *Educational Psychology Review*, *33*(4), 1475-1497. https://doi.org/http://dx.doi.org/10.1007/s10648-021-09603-y

Carpenter, B. D., Balsis, S., Otilingam, P. G., Hanson, P. K., & Gatz, M. (2009). The Alzheimer's disease knowledge scale: Development and psychometric properties. *Gerontologist*, *49*(2), 236-247. https://doi.org/10.1093/geront/gnp023

Cerdán, R., Candel, C., & Leppink, J. (2018). Cognitive load and learning in the study of multiple documents. *Frontiers in Education*, *3*. https://doi.org/10.3389/feduc.2018.00059

Cerdán, R., & Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *Journal of Educational Psychology*, *100*(1), 209-222. https://doi.org/10.1037/0022-0663.100.1.209

Cheong, C. M., Zhu, X., Li, G. Y., & Wen, H. (2019). Effects of intertextual processing on L2 integrated writing. *Journal of Second Language Writing*, *44*, 63-75. https://doi.org/10.1016/j.jslw.2019.03.004

Doctorow, M., Wittrock, M. C., & Marks, C. (1978). Generative processes in reading comprehension. *Journal of Educational Psychology*, *70*(2), 109-118. https://doi.org/10.1037/0022-0663.70.2.109

Du, H., & List, A. (2020). Researching and writing based on multiple texts. *Learning and Instruction*, *66*, 101297. https://doi.org/10.1016/j.learninstruc.2019.101297

Florit, E., Cain, K., & Mason, L. (2020). Going beyond children's single-text comprehension: The role of fundamental and higher-level skills in 4th graders' multiple-document comprehension. *British Journal of Educational Psychology*, *90*(2), 449-472. https://doi.org/10.1111/bjep.12288

Fiorella, L., Mayer, R. E. (2016). Eight Ways to Promote Generative Learning. *Educational Psychology Review*, *28*, 717-741. https://doi.org/10.1007/s10648-015-9348-9

Fiorella, L., & Mayer, R. E. (2017). Spontaneous spatial strategy use in learning from scientific text. *Contemporary Educational Psychology*, *49*, 66-79. https://doi.org/10.1016/j.cedpsych.2017.01.002

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*, 215-239. https://doi.org/10.1016/0378-8733(78)90021-7

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. (2010a). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology*, *35*, 157-173. https://doi.org/10.1016/j.cedpsych.2009.11.002

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. (2010b). Understanding and integrating multiple science texts: summary tasks are sometimes better than argument tasks. *Reading Psychology, 31* (1), 30-68. https://doi.org/10.1080/02702710902733600

Gurlitt, J., & Renkl, A. (2010). Prior knowledge activation: How different concept maptasks lead to substantial differences in cognitive processes, learning outcomes, and perceived self-efficacy. *Instructional Science*, *38*(4), 417-433. https://doi.org/10.1007/s11251-008-9090-5

Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., & Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *British Journal of Educational Psychology, 89*(3), 524–537. https://doi.org/10.1111/bjep.12278

Hilbert, T. S., & Renkl, A. (2008). Concept mapping as a follow-up strategy to learning from texts: what characterizes good and poor mappers? *Instructional Science*, *36*, 53-73. https://doi.org/10.1007/s11251-007-9022-9

Ifenthaler, D. (2014). Toward automated computer-based visualization and assessment of team-based performance. *Journal of Educational Psychology*, *106*(3), 651-665. https://doi.org/10.1037/a0035505

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Lawrence Erlbaum Associates, Inc.

Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: design justifications and guidelines. *Educational Technology Research and Development, 58* (4), 439-457. https://www.jstor.org/stable/40929365

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558. https://doi.org/10.1080/09541440601056620

Kim, M. (2012). Cross-validation study of methods and technologies to assess mental models in a complex problem solving situation. *Computers in Human Behavior*, *28*(2), 703-717. https://doi.org/10.1016/j.chb.2011.11.018

Kim, M., & McCarthy, K. S. (2021). Using graph centrality as a global index to assess students' mental model structure development during summary writing. *Educational Technology Research and Development*, *69*(2), 971-1002. https://doi.org/10.1007/s11423-021-09942-1

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163-182. https://doi.org/1037/0033-295X.95.2.163

Latini, N., Bråten, I., Anmarkrud, Ø., & Salmerón, L. (2019). Investigating effects of reading medium and reading purpose on behavioral engagement and textual integration in a multiple text context. *Contemporary Educational Psychology*, *59*, 101797. https://doi.org/10.1016/j.cedpsych.2019.101797

Lehmann, T., Pirnay-Dummer, P., & Schmidt-Borcherding, F. (2020). Fostering integrated mental models of different professional knowledge domains: Instructional approaches and model-based analyses. *Educational Technology Research and Development*, *68*, 905-927. https://doi.org/10.1007/s11423-019-09704-0

Lehmann, T., Rott, B., & Schmidt-Borcherding, F. (2019). Promoting pre-service teachers' integration of professional knowledge: Effects of writing tasks and prompts on learning from multiple documents. *Instructional Science*, *47*, 99-126. https://doi.org/10.1007/s11251-018-9472-2

Leopold, C., & Leutner, D. (2012). Science text comprehension: Drawing, main idea selection, and summarizing as learning strategies. *Learning and Instruction*, *22*(1), 16-26. https://doi.org/10.1016/j.learninstruc.2011.05.005

Linn, M. C. (2000). Designing the knowledge integration environment. *International Journal of Science Education, 22*, 781–796. https://doi.org/10.1080/09500 69004 12275

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, *94*(4), 778-784. https://doi.org/10.1037/0022-0663.94.4.778

List, A. (2019). Drawing is integrating: An examination of students' graphic representations of multiple texts. *Reading Psychology*, *40*, 1-34. https://doi.org/10.1080/02702711.2019.1629517

List, A., & Alexander, P. A. (2017). Cognitive affective engagement model of multiple source use. *Educational Psychologist, 52*, 182-199. https://doi.org/10.1080/00461520.2017.1329014.

List, A., & Du, H. (2021). Reasoning beyond history: Examining students' strategy use when completing a multiple text task addressing a controversial topic in education. *Reading and Writing*, *34*(4), 1003-1048. https://doi.org/10.1007/s11145-020-10095-5

List, A., Du, H., Wang, Y., & Lee, H. Y. (2019). Toward a typology of integration: Examining the documents model framework. *Contemporary Educational Psychology*, *58*, 228-242. https://doi.org/10.1016/j.cedpsych.2019.03.003

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217-224. https://doi.org/10.1080/00031305.2000.10474549

Mahlow, N., Hahnel, C., Kroehne, U., Artelt, C., Goldhammer, F., & Schoor, C. (2020). More than (single) text comprehension? - on university students' understanding of multiple documents. *Frontiers in Psychology*, *11*, 562450. https://doi.org/10.3389/fpsyg.2020.562450

Mason, L., Scrimin, S., Zaccoletti, S., Tornatora, M. C., & Goetz, T. (2018). Webpage reading: Psychophysiological correlates of emotional arousal and regulation predict multiple-text comprehension. *Computers in Human Behavior*, *87*, 317-326. https://doi.org/https://doi.org/10.1016/j.chb.2018.05.020

Mayer, R. E. (1984). Aids to text comprehension. *Educational Psychologist, 19*, 30-42. https://doi.org/10.1080/00461528409529279

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). New York: Cambridge University Press.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (Second Editionth ed., pp. 43-71). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139547369

McCrudden, M. T., Kulikowich, J. M., Lyu, B., & Huynh, L. (2022). Promoting integration and learning from multiple complementary texts. *Journal of Educational Psychology*, *114* (8), 1832-1843. https://doi.org/10.1037/edu0000746

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology*, *35*(4), 229-241. https://doi.org/10.1016/j.cedpsych.2009.12.001

McCrudden, M. T., & Rapp, D. N. (2017). How visual displays affect cognitive processing. *Educational Psychology Review, 29*, 623-639.

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing.

Educational Psychology Review, *19*(2), 113-139. https://doi.org/10.1007/s10648-006-9010-7

McCrudden, M. T., & Schraw, G. (2010). The effects of relevance instructions and verbal ability on text processing. *Journal of Experimental Education*, *78*(1), 96-117. https://doi.org/10.1080/00220970903224529

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In *The psychology of learning and motivation* (pp. 297-384). https://doi.org/10.1016/s0079-7421(09)51009-2

Micai, M., Vulchanova, M., & Saldana, D. (2021). Reading goals and executive function in autism: An eye-tracking study. *Autism Research*, *14*(5), 1007-1024. https://doi.org/10.1002/aur.2447

National Institute of Health. (2021). *Alzheimer's and Aging Resources for Students*. Retrieved from https://www.nia.nih.gov/health/alzheimers-and-aging-resources-students

O'donnell, A. M., Dansereau, D. F., & Hall, R. H. (2002). Knowledge maps as scaffolds for cognitive processing. *Educational psychology review*, *14*, 71-86. https://doi.org/10.1023/A:1013132527007

Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99-122). Mahwah, NJ: Lawrence Erlbaum Associates.

Pirnay-Dummer, P., & Ifenthaler, D. (2011). Text-guided automated self assessment. In D. Ifenthaler, J. M. Spector, P. Isaias, & D. Sampson (Eds.), *Multiple perspectives on problem solving and learning in the digital age* (pp. 217-225). Springer New York.

Primor, L., & Katzir, T. (2018). Measuring multiple text integration: A review. *Frontiers in Psychology*, *9*, 2294. https://doi.org/10.3389/fpsyg.2018.02294

Primor, L., Yeari, M., & Katzir, T. (2021). Choosing the right question: The effect of different question types on multiple text integration. *Reading and Writing*, *34*(6), 1539-1567. https://doi.org/10.1007/s11145-021-10127-8

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). IAP Information Age Publishing.

Rouet, J.-F., Britt, M. A., & Durik, A. M. (2017). RESOLV: Readers' representation of reading contexts and tasks. *Educational Psychologist*, *52*(3), 200-215. https://doi.org/10.1080/00461520.2017.1329015

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior*, *11*(4), 355-363. https://doi.org/10.1080/10862967909547341

Salmerón, L., Gil, L., & Bråten, I. (2018). Effects of reading real versus print-out versions of multiple documents on students' sourcing and integrated understanding. *Contemporary Educational Psychology*, *52*, 25-35. https://doi.org/10.1016/j.cedpsych.2017.12.002

Salmerón, L., Sampietro, A., & Delgado, P. (2020). *Using Internet videos to learn about controversies: Evaluation and integration of multiple and multimodal documents by primary school students. Computers & Education, 148,* 103796. https://doi.org/10.1016/j.compedu.2019.103796

Schneider, M. (2012). Knowledge integration. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1684-1686). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_807

Schoor, C., Hahnel, C., Mahlow, N., Klagges, J., Kroehne, U., Goldhammer, F., & Artelt, C. (2020). Multiple document comprehension of university students. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, & C. Lautenbach (Eds.), *Student learning in german higher education: Innovative measurement approaches and research results* (pp. 221-240). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27886-1_11

Schvaneveldt, R. W., Dearholt, D. W., & Durso, F. T. (1988). Graph theoretic foundations of pathfinder networks. *Computers & Mathematics with Applications*, *15*(4), 337-345. https://doi.org/10.1016/0898-1221(88)90221-0

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. *Psychology of learning and motivation, 24*, 249-284. https://doi.org/10.1016/S0079-7421(08)60539-3

Schwamborn, A., Mayer, R. E., Thillmann, H., Leopold, C., & Leutner, D. (2010). Drawing as a generative activity and drawing as a prognostic activity. *Journal of Educational Psychology*, *102*(4), 872-879. https://doi.org/10.1037/a0019640

Schwendimann, B.A. (2014). Making Sense of Knowledge Integration Maps. In: Ifenthaler, D., Hanewald, R. (eds) *Digital Knowledge Maps in Education* (pp. 17-40). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3178-7_2

Stadtler, M., Scharrer, L., Brummernhenrich, B., & Bromme, R. (2013). Dealing with uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cognition and Instruction*, *31*(2), 130-150. https://doi.org/10.1080/07370008.2013.769996

Stadtler, M., Scharrer, L., Skodzik, T., & Bromme, R. (2014). Comprehending multiple documents on scientific controversies: Effects of reading goals and signaling rhetorical relationships. *Discourse Processes*, *51*(1-2), 93-116. https://doi.org/10.1080/0163853x.2013.855535

Stang Lund, E., Bråten, I., Brandmo, C., Brante, E. W., & Strømsø, H. I. (2019). Direct and indirect effects of textual and individual factors on source-content integration when reading about a socio-scientific issue. *Reading and Writing*, *32*(2), 335-356. https://doi.org/10.1007/s11145-018-9868-z

Swanborn, M. S. L., & de Glopper, K. (2002). Impact of reading purpose on incidental word learning from context. *Language Learning*, *52*(1), 95-117. https://doi.org/10.1111/1467-9922.00178

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84* (4), 327-352. https://doi.org/10.1037/0033-295X.84.4.327

van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition*, *29*(8), 1081-1087. https://doi.org/10.3758/BF03206376

van Meter, P. N., & Cameron, C. (2018). The effects of presentation format on multiple document notetaking. *Learning and instruction*, *57*, 47-56. https://doi.org/10.1016/j.learninstruc.2018.03.002

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*(2), 129-160. https://doi.org/10.1016/S0361-476X(02)00011-5

van Peppen, L. M., Verkoeijen, P., Heijltjes, A., Janssen, E., & van Gog, T. (2021). Repeated retrieval practice to foster students' critical thinking skills. *Collabra-Psychology*, *7*(1), 28881. https://doi.org/10.1525/collabra.28881

Vidal-Abarca, E., Mañá, A., & Gil, L. (2010). Individual differences for self-regulating task-oriented reading activities. *Journal of Educational Psychology*, *102*(4), 817–826. https://doi.org/10.1037/a0020062

Vidal-Abarca, E., Martinez, T., Salmerón, L., Cerdán, R., Gilabert, R., Gil, L., Mañá, A., Llorens, A. C., & Ferris, R. (2011). Recording online processes in task-oriented reading with read&answer. *Behavior Research Methods*, *43*(1), 179-192. https://doi.org/10.3758/s13428-010-0032-1

von der Muhlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student-scientist comparison. *Reading and Writing*, *29*(8), 1677-1698. https://doi.org/10.1007/s11145-015-9601-0

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. https://doi.org/10.1016/j.learninstruc.2019.101266

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, *91*(2), 301–311. https://doi.org/10.1037/0022-0663.91.2.301

Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*(1), 73-87. https://doi.org/10.1037/0022-0663.83.1.73

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist*, *11*, 87–95. https://doi.org/10.1080/00461520903433554

Zhang, Z., & Yuan, K.-H. (2018). *Practical statistical power analysis using Webpower and R*. ISDSA press.

Appendix The instructions for the reading task and post-reading generative learning tasks

Instructions for the reading task

During this task, you need to read documents with the following question: [the focus questions for the given reading purpose shown here]. You can click the buttons in the center to switch the present document. The content in the document is masked by default and you need to move the mouse cursor to each paragraph to show it. You can also move the mouse cursor to the right widget to check the instructions.

Instructions for the concept mapping task

During this task, you need to construct a concept map to summarize the documents when considering the following question: [the focus questions for the given reading purpose shown here]. You can click the buttons in the center to show concepts in each document. Note that you only can use thirteen concepts to construct your map, and the common concept *Alzheimer's disease* is compulsory to use. Drag: left click a concept to drag it from the concept list to the canvas to use it, you can also drag it back to the concept list to remove it. Link: right click two concepts successively to create a link between them, and do the same operation again if you want to delete the link.

Instructions for the summary writing task

During this task, you need to construct a summary to summarize the documents when considering the following question: [the focus questions for the given reading purpose shown here]. You can click the buttons in the center to show concepts in each document. Note that you

only can use thirteen concepts to construct your summary, and the common concept *Alzheimer's disease* is compulsory to use.

Blinded authors – do not delete, keep this page but remove it during submission

Author_01 – Chen, X., Li, Z., Wei, Z., & Clariana, R. B. (2022). The influence of the conceptual structure of external representations when relearning history content. *Educational Technology Research and Development, 71*, 415-439. https://doi.org/10.1007/s11423-022-10176-y

Author_02 – Clariana, R. B., & Koul, R. (2004). A computer-based approach for translating text into concept map-like representations. In A. J. Canas, J. D. Novak, and F. M. Gonzales, Eds., *Concept maps: theory, methodology, technology*, vol. 2, in the Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain, Sep 14-17, pp.131-134.

Author_03 – Clariana, R. B., Koul, R., & Salehi, R. (2006). The criterion-related validity of a computer-based approach for scoring concept maps. *International Journal of Instructional Media*, *33*, 317-326. https://www.learntechlib.org/p/64253

Author_04 – Clariana, R. B., Rysavy, M. D., & Taricani, E. (2015). Text signals influence team artifacts. *Educational Technology Research and Development*, *63*(1), 35-52. https://doi.org/10.1007/s11423-014-9362-5

Clariana, R. B., & Solnosky, R. (2023). Anchoring concepts influence essay conceptual structure and test performance. In D. G. Sampson, D. Ifenthaler & P. Isaías (Eds.), *Proceedings of the 20th international conference on Cognition and Exploratory Learning in the Digital Age – CELDA 2023* (pp. 241-248). IADIS Press. ISBN: 978-989-8704-52-8

Author_05 – Clariana, R. B., Wallace, P. E., & Godshalk, V. M. (2009). Deriving and measuring group knowledge structure from essays: The effects of anaphoric reference. *Educational Technology Research and Development*, *57*(6), 725-737. https://doi.org/10.1007/s11423-009-9115-z

Author_06 – Follmer, D. J., Fang, S.-Y., Clariana, R. B., Meyer, B. J. F., & Li, P. (2018). What predicts adult readers' understanding of stem texts? *Reading and Writing*, *31*(1), 185-214. https://doi.org/10.1007/s11145-017-9781-x

Author_07 – both papers below

Kim, K., & Clariana, R.B. (2015). Knowledge structure measures of reader's situation models across languages: Translation engenders richer structure. Technology, Knowledge and Learning, 20, 249-268.

Kim, K., & Clariana, R. B. (2017). Text signals influence second language expository text comprehension: Knowledge structure analysis. *Educational Technology Research and Development*, *65*(4), 909-930. https://doi.org/10.1007/s11423-016-9494-x