Measuring cognitive task demands using dual task methodology, subjective self-ratings, and expert judgments: A validation study

Abstract

This study explored the usefulness of dual-task methodology, self-ratings, and expert judgements in assessing task-generated cognitive demands as a way to provide validity evidence for manipulations of task complexity. The participants were 96 students and 61 ESL teachers. The students, 48 English native speakers and 48 ESL speakers, carried out simple and complex versions of three oral tasks – a picture narrative, a map task, and a decision-making task. Half of the students completed the tasks under a dual task condition. The remaining half performed the tasks under a single task condition without a secondary task. Participants in the single condition were asked to rate their perceived mental effort and task difficulty. The ESL teachers provided expert judgments of anticipated mental effort and task difficulty along with explanations for their ratings via an online questionnaire. As predicted, the more complex task versions were found and judged to pose greater cognitive effort.

Introduction

The role of pedagogic tasks has received increasing attention from instructed second language acquisition (SLA) researchers over the past two decades. This growing interest has partly been inspired by theoretical proposals suggesting that tasks have the capacity to generate optimal conditions for the cognitive and social processes hypothesized to facilitate SLA, and partly by the rising popularity of task-based language teaching (TBLT) in a variety of educational settings. A large part of the existing empirical research has set out to test the predictions of two task-based models: Skehan's (1998, 2009) Limited Capacity Model and Robinson's (2001a, 2007, 2011) Cognition Hypothesis. These cognitive-interactionist models are primarily concerned with how manipulating task factors may affect second language (L2) learning opportunities, performance and development. Thus far, the bulk of empirical studies evaluating these models have focused on the impact of task complexity, i.e., the cognitive-attentional demands posed by the tasks, on the linguistic complexity, accuracy, and fluency (CAF) of L2 performance.

Recently, concerns have been raised regarding the adequacy of the research methods utilized in researching the relationship between task complexity and linguistic performance, questioning whether CAF (Housen & Kuiken, 2009) and cognitive task complexity (Norris & Ortega, 2003; Norris, 2010; Révész, 2014) are appropriately operationalized and measured. In response, researchers have begun to dedicate enhanced effort to identifying and exploring methodological procedures that permit valid examinations of the predictions of the Cognition Hypothesis, the Limited Capacity Model, and other theoretical proposals (e.g., Levelt, 1989) relevant to explaining links between task factors and SLA. By now, some useful guidelines have been produced on how to arrive at valid representations of the multidimensional and dynamic nature of CAF constructs (e.g., Norris & Ortega, 2009), to which researcher increasingly adhere. Recommendations have also been put forward for evaluating the validity

of cognitive task complexity manipulations (Révész, 2014). So far, however, only a few studies have made an explicit attempt to provide independent validity evidence for operationalizations of task complexity (Baralt, 2009; Malicka & Levkina, 2012; Révész, Sachs, & Hama, 2014; Sasayama, 2013), beyond the use of post-task questionnaires to confirm learners' perceptions of task difficulty (e.g., Gilabert, 2006, 2007; Gilabert, Barón, & Llanes, 2009, Gilabert, Barón, & Levkina, 2011; Levkina & Gilabert, 2012; Gilabert & Barón, 2013; Levkina & Gilabert, 2014; Kim, 2009; Kim & Tracy-Ventura, 2011; Michel, 2011; Révész, 2009, 2011; Robinson, 2001b, 2007).

The aim of this study was to help address this gap by utilising three methods – dual task methodology, self-ratings, and expert judgments – in an attempt to assess task-generated cognitive demands and thereby explore the relative efficacy of each in tapping the cognitive complexity of tasks. Our study is among the first to triangulate data obtained through these three techniques, and it is unique in that it investigated the effects of task complexity manipulations on cognitive demands for three rather than a single task type.

Literature review

Key constructs and models

To begin with, it is useful to clarify some key concepts and theoretical models pertinent to the exploration of how task factors may influence L2 learners' cognitive operations. As mentioned above, the two frameworks that have informed much of the previous research on the links between task characteristics, cognitive processes, an d L2 performance are Skehan's Limited Capacity Model (1998, 2009) and Robinson's Cognition Hypothesis (2001a, 2007). In the Limited Capacity Model, the primary independent variable is task demands, which are defined as having three components: code complexity, cognitive complexity, and communicative stress. In his most recent formulation, Skehan (2009) invokes Levelt's model

of speech production to explain the effects of manipulating task demands on task processes and outcomes. It is suggested that task factors may exert more or less pressure at the speech production stages of conceptualisation and/or formulation, and the quality of the resulting linguistic performance will depend on the extent to which the cognitive demands imposed on conceptualisation and formulation processes can be handled against working memory or attentional limitations. The dependent variables in the model are specified in terms of the linguistic performance measures of complexity, accuracy, and fluency.

In the Cognition Hypothesis, the main independent variable is task complexity or the inherent cognitive demands of tasks, which is similar to Skehan's cognitive complexity. In explaining how task complexity may affect L2 outcomes, Robinson (2003) draws on Wickens' (2007) multiple-resources account of attention. Wickens assumes the existence of multiple cognitive resource pools, which differ along three dichotomous dimensions: processing stage (perception vs. response), modality (auditory/vocal vs. visual/manual), and processing code (verbal vs. spatial). Each dimension is responsible for a distinct aspect of task performance, and the complexity of a task is seen as a consequence of the interference (i.e., confusion and cross talk) between similar codes or the competition for the same types of codes (e.g., verbal or visual), rather than the result of limitations in attentional capacity as Skehan assumes. For instance, carrying out two auditory tasks (e.g. listening to music and to an interlocutor) is more likely to result in interference and/or competition than performing a visual and an auditory task simultaneously (e.g., watching TV). Building on this and Levelt's (1989) speech production model, Robinson (e.g., 2001a, 2003) claims that making tasks more cognitively complex along certain task dimensions will impact, in predictable ways, processes such as conceptualisation and formulation of speech, allocation of attentional and memory resources to input and output, and depth of processing. The dependent variables that these processes are expected to affect include quality of linguistic performance defined in

terms of general and specific indices of CAF, patterns of uptake and interaction, retention of input, and automatization. Although the two models predict partly different linguistic outcomes for particular task dimensions, the major difference between them lies in the explanatory processes invoked.

It is important to point out here that the aim of this study was not to assess the predictions of these two task frameworks. Rather, our intention was to explore and identify methodological procedures that may assist, in future TBLT investigations including but not limited to tests of the Limited Capacity and Cognition Hypotheses, in adequately operationalizing cognitive task demands and tapping task-generated cognitive processes.

Issues in measuring task-generated cognitive demands and processes

A methodological shortcoming in many existing studies of task complexity lies in the fact that researchers have rarely sought separate evidence for the validity of their task complexity manipulations (see, however, Baralt, 2009; Malicka & Levkina, 2012; Révész, Sachs, & Hama, 2014; Sasayama, 2013). Operationalizing task complexity has typically involved creating a simple and complex version of a given task along a dimension which was assumed to influence the cognitive demands entailed in performing it. For example, a task may have been designed to lead to more or less reasoning, the expectation being that the resulting complex/simple task version would exert more/less pressure on cognitive operations. Until recently, it mostly remained unassessed whether this intended impact was in fact achieved. It was usually presumed rather than proved that the task version constructed to be more complex was truly more cognitively challenging (Norris, 2010; Révész, 2014). However, as Norris and Ortega (2003) rightly emphasise, when testing theoretical constructs, it is vital that they are measured clearly and independently. As regards task complexity, this means that researchers need to establish independently whether their task complexity

manipulations have indeed resulted in the anticipated changes in cognitive demands (Norris & Ortega, 2003; Norris, 2010; Révész, 2014).

It is worth noting that cognitive demands or cognitive load is a multi-dimensional construct, which reflects the interaction between task-generated processing demands and learner abilities (Bachman, 2002; Paas & van Merriënboer, 1994). In other words, any measurement of cognitive load will be affected by the extent of challenge learners encounter during task completion depending on their individual characteristics, thereby capturing certain aspects of what Robinson (2001a) terms as task difficulty (see below, however, for a discussion of potential differences between perceived cognitive load/mental effort and task difficulty). Thus, cognitive load assessment does not permit pure measurement of task complexity. Instead, it enables gauging the amount of cognitive effort that was induced by a certain task complexity manipulation.

Drawing on research from applied cognitive science, Révész (2014) outlines several types of methods which may enable researchers to measure the amount of cognitive load or mental effort generated by intended changes in task complexity. From among these, this study employed two methods, dual task methodology and self-ratings. In addition, we obtained expert judgements to gain further insights about the success of our task complexity manipulations. This selection of methods was motivated by our intention to use a combination of different types of techniques, empirical (dual-task methodology and self-ratings) versus analytical (expert opinions) and subjective (self-ratings, expert opinions) versus objective (dual-task methodology) (Paas & van Merriënboer's, 1994; Paas, Tuovinen, Tabbers, & van Gerven, 2003). Now we turn to a description of these techniques and a brief review of instructed SLA research that has so far utilised them.

Dual task methodology as a measure of cognitive load

Dual task methodology involves carrying out a secondary task parallel with the primary task. Secondary tasks normally take the form of simple activities that require continual attention, such as detecting a simple visual (Cierniak, Scheiter, & Gerjets, 2009) or auditory stimulus (Brünken, Seufert, & Paas, 2004). The underlying rationale of the procedure is that the amount of cognitive load imposed by the primary task is reflected in the performance on the secondary task. Secondary task performance is typically assessed in terms of reaction time and accuracy, slower or more inaccurate performance being taken to indicate more consumption of cognitive resources by the primary task. Dual task methodology is considered in cognitive psychology as a reliable and sensitive methodology, and it has the advantage over self-report measures that it constitutes a concurrent measure of processing load.

DeKeyser (1997) was one of the first studies to employ dual-task methodology in SLA research. In a longitudinal study of automatization, DeKeyser employed both single-task and dual-task conditions to assess participants' developing knowledge of explicitly taught morphosyntactic rules. As part of the dual-task procedure, participants were presented with a number between 100 and 1,000 on the computer screen before a picture related to the target linguistic item appeared. They additionally heard beeps at irregular intervals while they were subsequently working on production and comprehension items targeting the morphosyntactic rules. The secondary task involved memorising the number, counting the beeps, and calculating the difference between the original number and the number of beeps. DeKeyser predicted that, as the morphosyntactic rules are increasingly automatized, there will be less interference from the secondary task. He observed, however, a faster and less gradual decline in the difference between the single and dual conditions than expected, probably, he speculates, because the secondary task was not sufficiently demanding.

More recently, Declerck and Kormos (2012) also used dual-task methodology to examine encoding mechanisms and lexical selection processes in L2 speech production. The participants performed a parallel finger-tapping task while carrying out an oral production task. The finger-tapping task asked participants to press one out of ten keys every second as randomly as possible. The authors found that participants' finger-tapping behaviour became less random when they also had to perform a speaking task simultaneously. Thus, in this research, quality of performance on the finger-tapping secondary task was affected by the presence of the primary task to the predicted degree.

Finally, of particular relevance here are two recent studies of task complexity by Révész, Sachs, and Hama (2014) and Sasayama (2013). Révész et al. employed dual-task methodology, besides eye-tracking and expert judgements, to measure levels of cognitive load posed by the simple versus complex versions of two computer-delivered experimental tasks. The experimental tasks, which served as the primary tasks, required participants to choose between two past events to produce a past counter factual statement orally about a famous person's life. The events were presented on the computer screen by the means of corresponding pictures and key phrases. The secondary task involved, following Cierniak et al. (2009), the color of the computer background screen changing to red or green at random intervals. The participants' task was to respond as quickly and as accurately as possible to changes to one color (red) while ignoring changes to the other color (green). Although no difference emerged in secondary-task reaction times depending on intended task complexity, participants, as anticipated, were less accurate on the secondary task when carrying out the complex version of the primary task. These results for accuracy were found to be well aligned with the data obtained from expert judgments (see below) and eye-tracking, attesting to the validity of the task complexity manipulation.

Sasayama also triangulated results from dual-task methodology with two other techniques (time estimation and self-ratings) to assess cognitive load on a series of computer-delivered narrative tasks. Four versions of a narrative task were designed involving an increasing number of elements, with the intention to generate increasing cognitive effort on the part of the participants. As part of the secondary task, participants were asked to react as quickly and accurately as possible to changes in the colour of the capital letter A, which was projected above the picture story eliciting the narrative on the computer screen. Specifically, participants were instructed to respond when the color of the letter changed from black to red but ignore changes from red to black. Sasayama found increased reaction times, as predicted, on the most complex narrative as compared to the least complex narrative, but contrary to expectations, no effects were observed for accuracy. Importantly, the results for reaction times were largely parallel to those obtained via time estimation and self-ratings, although dual-task methodology appeared a less sensitive technique than self-ratings and time estimation because it was not able to distinguish the intermediate-complexity versions from the most simple and complex narratives. As Révész et al. (2014), the author took the corresponding patterns on the three measures as evidence for the validity of the task manipulation.

The results of Sasayama (2013) and Révész et al (2014) suggest that dual-task methodology is a viable way to tap task-generated cognitive load. Nevertheless, given the small amount of research available and the conflicting findings for accuracy and reaction times, more research is warranted evaluating the sensitivity of dual-task methodology in detecting differential task-induced cognitive load.

Self-rating scales as measures of cognitive load

So far, the most common method to determine cognitive load in educational psychology (Brünken, Seufert, & Paas, 2010) as well as in SLA has been the use of subjective self-rating scales or self-report questionnaires. The use of self-rating scales in assessing cognitive load builds on the assumption that people can assign a numerical value to the amount of mental effort they spent during a particular cognitive activity. Indeed, it has been demonstrated that ratings of perceived mental effort can provide valid and reliable measures of cognitive load (e.g., Paas, 1992).

Self-rating scales of cognitive load usually involve requesting learners to judge their perceived cognitive load in response to a semantic differential scale worded as "This task required ... mental effort," with options ranging from "no" to "extreme" on a 7- to 9-item scale. In addition, learners are frequently asked to provide ratings of variables such as perceived task difficulty and anxiety, resulting in a multi-dimensional scale. Importantly, while perceived mental effort or cognitive load often correlates highly with these factors, it is not isomorphic with them (Brünken et al., 2010). To give an example, an expert might rate both a simple and a complex version of an activity as easy (simple math problem: 2+2; more complex math problem: 11*11) on a 9-point scale, resulting in little or no difference in task difficulty ratings. Still, the more complex version might be evaluated by the same expert as requiring the investment of somewhat greater cognitive effort than the simple version on the same scale. It is not surprising, therefore, that unidimensional scales, which solely assess perceived mental effort in the absence of scales tapping related but distinct constructs, have been shown to provide valid and reliable measurement of cognitive load (Paas & van Merriënboer, 1994*)*.

Self-rating scales have been employed in a growing number of L2 task complexity studies (Baralt, 2013; Gilabert, 2005, 2006, 2007: Gilabert, 2006, 2007; Gilabert et al., 2009,

Gilabert et al. 2011; Levkina & Gilabert, 2012; Gilabert & Barón, 2013; Levkina & Gilabert, 2014); Kim, 2009; Kim & Tracy-Ventura, 2011; Malicka & Levkina, 2012; Michel, 2011; Révész, 2009, 2011; Robinson, 2001b, 2007; Sasayama, 2013), most of which have adopted or adapted a multi-dimensional scale introduced by Robinson (2001b). Robinson's original questionnaire included items tapping perceptions of task difficulty, stress, ability to complete the task successfully, interest in the task, and task motivation; but had no items tapping perceived mental effort. Among existing task complexity studies, so far only Sasayama's (2013) previously mentioned study has utilised both a mental effort and task difficulty scale. She found that, among the four task versions she created, the one intended to be the most complex was indeed rated as more difficult and requiring more mental effort than two of the three versions that were designed to be of lower complexity. Thus, in Sasayama's research, the two scales have yielded convergent findings, which patterned, for the most part, with those obtained from the dual-task method. Clearly, however, more research is needed to test the utility of these scales in tapping cognitive load.

Expert judgments as a way to evaluate expected cognitive load

An analytic, subjective method to obtain information about task-generated cognitive demands is to ask experts to judge the mental effort that a task complexity manipulation is expected to create. In the area of language testing, expert judgments are often utilised to analyse the content and difficulty of test items as part of test development and/or validation processes. In a large-scale study of task-based performance assessment, Brown, Hudson, Norris and Bonk (2002) also used expert judgments as a means of eliciting evaluations of the cognitive operations anticipated to be induced by task manipulations. Inspired by Skehan's task framework, cognitive operations were categorised in terms of two characteristics: (a) the extent to which the input/output needed to be organised or reorganised as part of the task and

(b) the degree to which the examinee was required to search for the information which had to be utilised to complete the task. The raters were experienced ESL teachers, who were familiar with the relevant test-taker population. Altogether they evaluated 103 tasks based on the task prompts and descriptions of task realia/materials. They were requested to assign a plus or minus for several task factors including input/output organisation and input availability based on whether they thought that a particular task was likely to be above or below the average ability of the learner population with respect to the task factor in question. Interrater agreement was found to be in the moderate range for both cognitive factors (input/output organisation: .75; input availability: .62).

Within the area of instructed SLA, expert judgments have also been employed to assess the expected mental effort presented by task complexity manipulations. As mentioned above, Révész et al. (2014) utilised expert judgments to measure levels of cognitive load posed by the simple versus complex versions of the experimental tasks. Two applied linguistics doctoral students with a background in task-based language learning and teaching were asked to evaluate the items in both experimental tasks in terms of task complexity on a 5-point Likert scale, with higher ratings indicating greater complexity. The versions designed to be more complex were rated as higher in complexity by the two raters. Interrater reliability was adequate (ρ = .75; p < .01). Triangulated by data from eye-tracking and dual-task methodology (see above), the authors interpreted these results as suggesting that their task complexity manipulation was successful. Although these results suggest that expert judgements provide a valid assessment of task-generated cognitive load, further research is needed to confirm the validity of this measure.

Research Questions

In light of the above, the following research questions and hypotheses were formulated:

1. Do task manipulations designed to result in different levels of cognitive complexity do indeed lead to different levels of cognitive load as measured by dual task methodology?
2. Do task manipulations designed to result in different levels of cognitive complexity do indeed lead to different levels of perceived cognitive load as measured by self-ratings?
3. Are task manipulations designed to result in different levels of cognitive complexity judged by experts as resulting in different levels of cognitive load?

On the basis of previous research in the areas of applied cognitive psychology and second language acquisition, it was hypothesized that task versions designed to be more complex would result in slower reaction times and lower accuracy rates on a visual secondary task, would be perceived as requiring higher mental effort by participants, and would be judged as generating greater cognitive load by experts.

Methodology

Design

The participants were 96 students and 61 English as a Second Language (ESL) teachers. Out of the 96 students, 48 were native speakers (NS) of English and 48 ESL speakers. All 96 students carried out simple and complex versions of three oral tasks – a picture narrative, a map task, and a decision-making task. Task type and task complexity were counterbalanced across the participants. Half of the NSs (n=24) and half of the ESL speakers (n=24) completed the tasks under a dual task condition. The remaining half performed the tasks under a single task condition without an added secondary task. Participants in the single condition were asked to rate their perceived mental effort and task difficulty on a 9-point Likert scale immediately after completing a task version. The ESL teachers provided expert judgments of anticipated mental effort and task difficulty and explanations for their ratings via an online questionnaire.

Participants

The 48 NS student participants were enrolled at a UK university, whereas 24 ESL students were studying at a university in Germany and 24 in Spain. The German and Spanish participants were equally distributed across the dual and single groups. Except for first language background, the participants had similar demographic characteristics across the three contexts. The mean age was 22.86 (SD = 3.50) with a range of 18 to 35 (NS: M=23.08, SD=3.73; German ESL: M=22.88, SD=3.59; Spanish ESL: M=21.61, SD=3.12), and the majority of the participants were female (NS: 83.3%; German ESL: 66.7%; Spanish ESL: 66.7%). The proficiency levels of the ESL learners were in the B1-B2 bands according to the Common European Framework of Reference (CEFR), as determined by the Oxford Placement Test (Dave, 2004) which was administered to the ESL participants. For the purposes of this study, the German and Spanish ESL learners were treated as one group, given that L1 background was not relevant to the focus of the research.

The age of the participating ESL teachers spanned from 22 to 67 with a mean of 38.20 (SD = 10.80). Most of the teachers were female (68.9%). All teachers held at least an English language teaching certificate, and almost half had a master's degree in TESOL or applied linguistics (47.5%). The majority were native speakers of English (65.6%), and a smaller number of the teachers had German (7.8%), Catalan (6.5%), Spanish (6.5%), or other first language backgrounds (13.6%). The teachers' experience varied widely, ranging from 1 to 30 years of English language teaching (Median = 10, Mean = 10.33, SD = 8.36). On a 5-point Likert scale, the teachers' mean rating of their knowledge of TBLT and task complexity was 3.59 (SD=1.23) and 2.93 (SD= 1.44) respectively, with higher ratings indicating greater familiarity.

Instruments and Procedures

Tasks and task complexity manipulations

In order to increase the probability of detecting differences in task-generated cognitive load between the simple and complex versions of our experimental tasks, we decided to use task versions that have been found to differ in task difficulty based on self-ratings in previous research.

Inspired by Robinson (2007), the narrative tasks were adopted from the picture arrangement subtest of the Wechsler Adult Intelligence Scale, Third Revised version (WAIS-III). We selected Story 3 (LAUNDRY), one of the narratives with lesser complexity, and Story 11 (SHARK), the most complex narrative. The simple narrative required participants to describe a sequence of routine activities, which were completed by a single character when doing his laundry. This task version was expected to impose little cognitive load on the participants as it depicted simple and familiar events organized in a highly predictable sequence. The complex narrative, on the other hand, involved narrating an unusual story of a surfer playing a prank on a crowded beach. It was hypothesised that this task version would pose more cognitive demands, since the story entailed an unpredictable set of unfamiliar events and required reasoning about the psychological state of the main character and his perceptions about other people's beliefs and intentions. Unlike in the original Wechsler test, participants were not asked to put the stories in order, they were presented with the pictures in the correct sequence and asked to narrate the story.

The map tasks were adapted from Gilabert and colleagues' work (Gilabert, 2007, Gilabert et al., 2009). They required participants to leave a voice mail message giving detailed directions on how to get to the city center in order to arrange some errands. In the simple version of the task, participants were asked to give directions to a news stand to buy a newspaper, to the post office to post a letter, and to a flower shop to purchase a bunch of

flowers before going back to a subway station. To accomplish the task, participants had to describe a path along a single lateral axis (i.e., left, right, and straight) and involving unique reference points which were easy to distinguish. In the complex version, the participants' task was to explain where to pick up a dog from the veterinary surgeon and to purchase dog food from a department store before returning to a subway station. Compared to the simple version, the complex map task entailed many and similar points of reference. Also, participants needed to describe more complicated directions, moving not only along the lateral (i.e. left, right, straight) but also the vertical (i.e., up, down) and sagittal axes (i.e., front, back), creating greater demands on landmark identification and path selection.

The decision-making tasks were also adapted from Gilabert and colleagues' work (Gilabert, 2007; Gilabert et al., 2009). Participants were asked to imagine that they volunteered for the university's fire emergency team, and the tasks involved explaining which actions they would take and in what order to save as many people as possible in case of an emergency. They were presented with a drawing of a building which was on fire and had a number of people trapped who needed to be rescued. In the simple version, people in the building were similar and were not part of vulnerable groups, there was sufficient equipment available (i.e., three fire trucks and a helicopter), and participants only had to take into account a few unrelated factors (i.e., people in safe places, static nature of fire, smoke moving away from the building). In contrast, the complex version required participants to rescue several vulnerable individuals (e.g., an elderly man, a pregnant woman with children, and a severely injured person). There were also fewer resources available (i.e., a single fire truck), and participants needed to consider closely related and dynamic factors generating different levels of danger when reaching decisions (e.g., fire moving towards the people, smoke blowing into the building). As a result, the complex version was expected to impose more reasoning demands on the speakers.

Dual-task methodology

In the current study, the simple and complex versions of the narrative, map, and decision-making tasks served as the primary task. The secondary task was adopted from Révész et al. (2014). The color of the computer background screen changed to red or green for 250ms at random during each interval of 2500ms. The participants were asked to react as fast and as accurately as possible to changes to green while ignoring changes to red. Accuracy and reaction times were obtained by E-Prime 2.0 (Schneider, Eschman, & Zuccolotto, 2002).

In the practice phase, participants first performed a practice task under the single condition without the added visual stimulus, similar to the participants in the single group. Then, baseline reaction times and accuracy rates were additionally collected. In the baseline phase, the participants saw the same visual prompts as under the single condition. Unlike the single group, however, they only performed the secondary task, that is, responded to color changes but did not need to speak. In the final stage, participants carried out the primary task and secondary task simultaneously, performing the practice task while also responding to color changes.

Next, as part of the actual dual-task experiment, participants performed the simple and complex versions of the narrative, map, and decision making tasks (primary tasks) while responding to color changes (secondary task). The order of tasks (narrative, map, vs. decision-making) and task-complexity versions (simple vs. complex) was counterbalanced across participants.

Self-rating scales

The perception questionnaire included six computer-delivered statements that the single groups participants needed to judge on a 9-point Likert scale after completing a task version. Out of the six items, only two are relevant to and discussed in the present article. These assessed participants' perceptions of (a) the mental effort required by the task and (b) overall task difficulty. The questionnaire items were administered to the participants in English, but care was taken to word the items in simple language.

Expert judgments

The expert judgments were obtained via an online questionnaire. For each task type, the simple and complex task versions were presented on one page. Before rating a particular task version, the teachers were asked to read the task instructions and consider the relevant visual prompts. Next, they assessed on a 9-point Likert scale (a) the overall difficulty of the task version and (b) the mental effort required by it. Once the teachers have completed the Likert-scale items for both the simple and complex task versions, they were asked to explain, in response to an open-ended item, why they gave the same or different ratings for mental effort/task difficulty for the two tasks. The aim of the open-ended items was to investigate the rationale for the teachers' judgements. The order of the simple and complex versions was randomised across the three task types.

Data Collection

The student participants took part in one individual session. First, informed consent was obtained, followed by the administration of a paper and pencil background questionnaire. The ESL participants were also asked to complete the paper and pencil version of the Oxford Placement Test (Dave, 2004). The rest of the experiment was delivered via the

psycholinguistic software E-Prime (Schneider, Eschman, & Zuccolotto, 2002). First, participants were familiarised with the task instructions and experimental procedures in a practice phase. This involved reading the task instructions, listening to a sample practice task performance, performing the practice task, and completing the task perception questionnaire. During the practice phase, participants were encouraged to ask any questions they had regarding the procedures. Next, participants moved on to completing the six experimental tasks. Participants had 30 seconds planning time to look at the prompt for each task before they were asked to start speaking.

The teachers were invited to participate in the task complexity questionnaire by e-mail and social media. The message specified the purpose of the study, the types of questions included, and the time commitment involved in completing the questionnaire (15 minutes). The questionnaire was administered via SurveyMonkey.

Data analyses

Statistical analyses

First, descriptive statistics were calculated for the accuracy and reaction time data, the self-ratings, and the expert judgements. To examine the effects of task complexity on the dual-task data and self-ratings, a series of ANOVAs was conducted. Dependent samples t-tests were utilised to compare the teachers' expert judgements of mental effort and task difficulty across the simple and complex task versions. Standard diagnostic procedures were used to ensure the appropriateness of all statistical models. The alpha level was set at $p < .05$, and we employed partial eta-squared (η_p^2) and Cohen's *d* to measure effect sizes. Following Cohen (1988), η_p^2 values of .01, .06, and .14 and *d* -values of .20, .50, and .80 were considered small, medium, and large. Prior to submitting the data to these statistical procedures, we performed statistical power analysis for all tests using GPower 3.1 (Faul,

Erdfelder, Lang, & Buchner, 2007). The sample sizes were found to be adequate to detect medium effect sizes for all factors of interest with an $\alpha = .05$ and power = .80.

Analysis of open-ended questionnaire items

The analysis of the open-ended questionnaire items included four phases. First, all three researchers reviewed the teachers' comments individually and identified emergent categories by annotating the data. Intercoder percentage agreement for category identification was found to be high for each pair of the three coders across the three task types (.80-. 94). Second, the coding categories were finalised through discussion among the researchers. Third, first author coded all of the teacher comments by annotating the data using the agreed coding scheme, and the second and third author both blind-coded a different and non-overlapping half of the dataset. In this way, all the data were double-coded. Intercoder agreement was high for both pairs of coders for all three tasks (narrative: coder 1-coder 2: .89, coder 1-coder 3: .94; map: coder 1-coder 2: .88, coder 1-coder 3: .97; decision-making: coder 1-coder 2: .84 , coder 1-coder 3: 85). The coding of the first author was included in further analyses. Finally, a frequency count of all the annotations for each task version was obtained by adding up the annotations falling into a specific category.

Results

Dual task methodology

Table 1 and 2 summarize the descriptive statistics for accuracy and reaction time on the secondary task under the baseline, simple, and complex task conditions for the three tasks after outliers were removed. For each task, outliers were identified considering the difference scores in accuracy and reaction time between each pairing of task conditions (simple-

complex, baseline-simple, and baseline-complex). Outliers were defined as having difference score values more than three standard deviations away from the mean.

To determine whether participants' accuracy in reacting to the correct color differed across the baseline, simple, and complex task conditions, a separate mixed-model ANOVA was conducted for each task using task condition (baseline, simple, complex) as the within-subjects variable and NS/ESL status as the between-subjects factor. Each of the three analyses yielded a significant and large effect for task condition, narrative: $F(2,88)=8.29$, $p=.001$, $\eta_p^2=.16$, map: $F(2,88)=29.81$, $p<.001$, $\eta_p^2=.40$, decision-making: $F(2,84)=21.30$, $p<.001$, $\eta_p^2=.34$. However, no significant effects were detected for NS/ESL status, narrative: $F(1,44)=1.40$, $p=.243$, $\eta_p^2=.03$, map: $F(1,44)=1.74$, $p=.194$, $\eta_p^2=.04$, decision-making: $F(1,42)=.78$, $p=.381$, $\eta_p^2=.02$, or the interaction, narrative: $F(2,88)=.11$, $p=.896$, $\eta_p^2=.02$, map: $F(2,88)=.82$, $p=.445$, $\eta_p^2=.02$, decision-making: $F(2,84)=.48$, $p=.621$, $\eta_p^2=.01$.

Next, a series of post-hoc mixed-model ANOVAs was carried out. In each analysis, the within-subjects factor was defined as a pairing of the three task conditions (simple-complex, baseline-simple, baseline-complex), and NS/ESL status was kept as the between-subjects factor. As expected, on each task participants produced significantly higher mean accuracy rates on the secondary task when carrying out the simple, as compared to the complex, versions of the primary tasks, narrative: $F(1,44)=5.41$, $p=.025$, $\eta_p^2=.11$, map: $F(1,44)=5.70$, $p=.021$, $\eta_p^2=.12$, decision-making: $F(1,42)=9.21$, $p=.004$, $\eta_p^2=.18$. The effect sizes had medium to large values. Participants also performed the secondary task significantly more accurately under the baseline condition than under either the simple dual-task condition, narrative: $F(1,44)=4.27$, $p=.045$, $\eta_p^2=.09$, map: $F(1,44)=22.96$, $p<.001$, $\eta_p^2=.34$, decision-making: $F(1,42)=18.44$, $p<.001$, $\eta_p^2=.31$, or than under the complex dual-task condition, narrative: $F(1,44)=13.77$, $p=.001$, $\eta_p^2=.24$, map: $F(1,44)=56.71$, $p=.001$, $\eta_p^2=.56$, decision-making: $F(1,42)=27.41$, $p<.001$, $\eta_p^2=.40$. The effect sizes were, again, in the medium to large

range. These results for the baseline-simple and baseline-complex comparisons indicate that, for accuracy, the secondary task, in combination with the primary tasks, constituted a sufficiently sensitive measure of cognitive load.

INSERT TABLE 1 AROUND HERE

For reaction times, the same type of overall mixed-model ANOVA were carried out. Task condition (baseline, simple, or complex) emerged as a significant and strong predictor of secondary task performance on each of the three tasks, narrative: $F(2,90)=224.92$, $p<.001$, $\eta_p^2=.83$, map: $F(2,90)=205.24$, $p<.001$, $\eta_p^2=.82$, decision-making: $F(2,90)=213.53$, $p<.001$, $\eta_p^2=.83$. For the narrative task, NS/ESL status, $F(1,45)=4.13$, $p=.048$, $\eta_p^2=.08$, and the interaction, $F(2,90)=11.17$, $p<.001$, $\eta_p^2=.20$, were also identified as significant but considerably weaker predictors of reaction times. On the map and decision-making tasks, however, no significant effects were found for either NS/ESL status, map: $F(1,45)=.29$, $p=.594$, $\eta_p^2=<.01$, decision-making: $F(1,45)=.25$, $p=.619$, $\eta_p^2<.01$, or the interaction, map: $F(2,90)=2.58$, $p=.082$, $\eta_p^2=.05$, decision-making: $F(2,90)=2.68$, $p=.074$, $\eta_p^2=.06$.

In contrast to our expectations, post-hoc mixed-model ANOVAs detected no significant effect for task complexity for any of the three tasks, narrative: $F(1,45)=2.69$, $p=.108$, $\eta_p^2=.06$, map: $F(1,45)=.216$, $p=.644$, $\eta_p^2<.01$, decision-making: $F(1,45)=.50$, $p=.484$, $\eta_p^2=.01$. Large and significant differences, however, were found between the baseline and simple conditions, narrative: $F(1,45)=360.53$, $p<.001$, $\eta_p^2=.89$, map: $F(1,45)=291.02$, $p<.001$, $\eta_p^2=.87$, decision-making: $F(1,45)=376.10$, $p<.001$, $\eta_p^2=.89$, and the baseline and complex conditions, narrative: $F(1,45)=366.05$, $p<.001$, $\eta_p^2=.89$, map: $F(1,45)=447.11$, $p<.001$, $\eta_p^2=.91$, decision-making: $F(1,45)=301.75$, $p<.001$, $\eta_p^2=.87$. Notably, NS/ESL status only emerged as a significant predictor on the narrative when the simple and complex dual conditions were

compared, that is, native speakers reacted significantly faster to the color changes than ESL speakers, baseline-simple: $F(1,45)=3.63$, $p=.063$, $\eta_p^2=.08$; baseline-complex: $F(1,45)=.53$, $p=.470$, $\eta_p^2=.01$; simple-complex: $F(1,45)=9.16$, $p=.004$, $\eta_p^2=.17$.

In sum, the reaction time data offer no evidence in support of the validity of our task complexity manipulations. Participants, however, achieved higher mean accuracy rates on the secondary task when carrying out the simpler versions of the primary tasks. This suggests that, as intended, the complex task versions generated greater cognitive load than the simple task versions.

INSERT TABLE 2 AROUND HERE

Self-ratings

Table 3 presents the descriptive statistics for the self-ratings of perceived mental effort and task difficulty for the simple and complex tasks versions of the three tasks. For each task, the data were first detected for outliers, defined as values more than three standard deviations away from the mean. No outliers were identified using these criteria. Spearman rank correlations revealed strong but not perfect correlations between perceived mental effort and difficulty, suggesting that the two scales were assessing related but not identical constructs (simple narrative: $\rho = .66$, $p < .001$, CI: [.43, .87]; complex narrative: $\rho = .69$, $p < .001$, CI: [.43, .88]; simple map: $\rho = .72$, $p < .001$, CI: [.47, .89]; complex map: $\rho = .79$, $p < .001$, CI: [.60, .93]; simple decision-making: $\rho = .79$, $p < .001$, CI: [.64, .87]; complex decision-making: $\rho = .68$, $p < .001$, CI: [.43, .87]).

To examine whether participants' self-ratings of mental effort were different for the simple and complex task conditions, separate mixed-model ANOVAs were carried out for each task using task complexity as the within-subjects variable and NS/ESL status as the

between-subjects factor. Each of the three analyses found that, in line with the intended task manipulations, participants rated the more complex task versions as requiring significantly more mental effort, narrative: $F(1,46)=5.28$, $p=.026$, $\eta_p^2=.10$, map: $F(1,46) = 13.52$, $p = .001$, $\eta_p^2=.23$, decision-making: $F(1,46)=7.26$, $p=.010$, $\eta_p^2=.14$, with effect sizes ranging from medium to large. No significant effects were found for NS/ESL status, narrative: $F(1,46)=.25$, $p=.618$, $\eta_p^2<.01$, map: $F(1,46)=.32$, $p=.572$, $\eta_p^2<.01$, decision-making: $F(1,46)=2.16$, $p=.148$, $\eta_p^2=.05$, or the interaction, narrative: $F(1,46)=.19$, $p=.669$, $\eta_p^2<.01$, map: $F(1,46)=.20$, $p=.661$, $\eta_p^2<.01$, decision-making: $F(1,46)=.40$, $p=.529$, $\eta_p^2<.01$.

A series of the same type of mixed-model ANOVAs was also conducted to assess the effects of task complexity on the task difficulty ratings across the three tasks. As predicted, participants perceived the more complex versions of the map task, $F(1,46)=17.75$, $p<.001$, $\eta_p^2=.28$, and decision-making task, $F(1,46)=17.25$, $p<.001$, $\eta_p^2=.27$, as more difficult. The effect sizes were also found to be large for both analyses. Contrary to expectations, however, there was no significant difference in task difficulty ratings between the simple and complex versions of the narrative, $F(1,46)=.90$, $p=.348$, $\eta_p^2=.02$. None of the analyses found a significant effect for NS/ESL status, narrative: $F(1,46)<.01$, $p=.957$, $\eta_p^2=<.01$, map: $F(1,46)=3.65$, $p=.062$, $\eta_p^2=.07$, decision-making: $F(1,46)=.26$, $p=.615$, $\eta_p^2<.01$, or the interaction, narrative: $F(1,46)=1.92$, $p=.172$, $\eta_p^2=.04$, map: $F(1,46)=2.11$, $p=.153$, $\eta_p^2=.04$, decision-making: $F(1,46)=.89$, $p=.350$, $\eta_p^2=.02$.

Taken together, the results for self-ratings confirm the validity of our task manipulations. As anticipated, the more complex task versions were rated as involving more mental effort and posing more difficulty. The only exception to this pattern was the lack of difference in participants' ratings of task difficulty between the simple and complex narratives.

INSERT TABLE 3 AROUND HERE

Expert judgments

The descriptive statistics for the expert judgments of mental effort and task difficulty are presented in Table 4. No outliers were detected utilising the same criteria as for the self-ratings. Spearman rank correlations, which were computed to assess the strength of the relationships between the experts judgments of required mental effort and task difficulty, yielded very high coefficients (simple narrative: ρ=.84, p < .001, CI: [.70,.92]; complex narrative: ρ=.92, p <.001, CI: [.85,.96]; simple map: ρ=.96 , p<.001, CI: [.93,.99]; complex map: ρ=.82, p<.001, CI: [.63,.95]; simple decision-making: ρ=.89, p< .001, CI: [.63,.95]; complex decision-making: ρ=.89, p<.001, CI: [.81,.94]). This could be interpreted as suggesting that the teachers perceived the two constructs as closely linked.

To assess whether the teachers judged the simple and complex task conditions as involving more mental effort and/or posing more difficulty, a series of dependent samples t-tests were conducted. In each of the six analyses, task complexity served as the independent variable and the ratings of mental effort and task difficulty as the dependent variables. The t-tests confirmed that, across all three task types, the more complex task version was judged by the teachers as necessitating significantly and considerably more mental effort, narrative: t(60)=6.49, p<.001, d=.76, map: t(60)=6.58, p<.001, d=.92, decision-making: t(60)=7.04, p<.001, d=1.01, and presenting greater difficulty, narrative: t(60)=5.25, p<.001, d=.68, map: t(60)=7.07, p<.001, d=.97, decision-making: t(60)=6.22, p<.001, d=.91. Effect sizes were high overall. In other words, the expert judgments indicate that our task complexity manipulations were successful.

INSERT TABLE 4 AROUND HERE

Tables 5-7 present the list of categories that emerged from the content analysis of the open-ended questionnaire comments. Examples and frequency counts for each category are provided separately for the three task types. Given that the teachers did not make a distinction between task difficulty and anticipated mental effort in their comments, the constructs are not distinguished in the remainder of the section.

For the narrative, the content analysis of the open-ended questionnaire item generated 103 annotations, none to five annotations per teacher. As shown in Table 5, teachers most frequently (85%) explained the differential difficulty between the two narratives by the varied amount of reasoning required. A number of more specific reasoning-related subcategories also emerged from the teachers' comments. The teachers thought that the narratives differed in the extent to which they required interpretation (26%), intentional reasoning (8%), creativity (7%) and causal reasoning (3%), whether their storyline was clear (25%) and predictable (8%), and whether they were more or less open-ended (8%). Number of elements in the narratives was the second most frequently mentioned factor determining task difficulty. Nearly half of the teachers (43%) referred to this general factor. Several teachers further specified that the rationale for their decision lay in the varied number of characters (18%), actions (16%), and places (5%). Differential linguistic demands was the third most frequently reported factor, with more than a quarter of the teachers (26%) making a reference to it. Some also mentioned complexity of lexis (21%) or morphosyntax (2%) as potential variables distinguishing the narratives. Four additional categories emerged from the content analysis (story not interesting, lack of clarity in pictures, sense of irony involved, culture-specific info). These were raised by only a small number of the participating teachers (2-7%).

Turning to the map task, the coding of the open-ended responses resulted in 76 annotations, ranging from none to four annotations per teacher. As Table 6 demonstrates, the

most oft-cited factor (69%) differentiating the two versions of the map task was the clarity of the visual information presented. Some teachers further explained that the difference in their difficulty/mental effort rating was a reflection of the extent of clarity in the maps (28%) and the three- versus two-dimensional nature of the diagrams (25%). The second and third most frequently mentioned factors were the number of elements and the difficulty of the route in the two map task versions. These were referred to by considerably fewer teachers (28% and 20%, respectively) than the clarity of the visual prompt. Varied linguistic demands was found to be the next most frequent category, with slightly more than ten percent of the participants (11%) raising it as a characteristic explaining task difficulty. The remaining two factors, difficulty of task instructions and cultural familiarity, were mentioned by only one teacher each (2%).

Finally, the coding of the comments on the decision-making task generated from none to four annotations per teacher, 79 annotations in total (see Table 7). Teachers most often (39%) attributed the difference in difficulty between the decision-making tasks to the differential reasoning demands posed by them. Several subcategories were also identified based on the teachers' comments, including how difficult decisions had to be made (10%), how much creativity was required (7%), how much need there was to prioritise (2%), and the extent to which a value judgment was necessary (3%). Number of elements was the second most frequently (33%) mentioned factor, with two subcategories emerging from the data. Several participating teachers referred to varied number of people (18%) and varied number of actions (3%) when explaining their difficulty/mental effort ratings. The next most oft-cited factor was difficulty involved in finding a solution. Availability of resources (10%) and wind direction (7%) were identified as subcategories associated with this general factor. The rest of the categories – level of danger, linguistic demands, and clarity of visual input – were

reported as perceived sources of difficulty by relatively few teachers (13%, 11%, 10% respectively).

In summary, the expert teacher judgments were found to be well aligned with our intended task manipulations. The task versions that were designed to be more complex were perceived by the teachers as more difficult and requiring more mental effort. The teachers' qualitative comments further revealed that what the teachers saw as potential sources of difficulty were in line with the rationale underlying our original task manipulations.

INSERT TABLES 5-7 AROUND HERE

Discussion

Our first research question asked whether task manipulations designed to result in different levels of cognitive complexity do indeed lead to different levels of cognitive load as measured by dual task methodology. As discussed earlier, the assumption underlying the dual-task technique is that an increased amount of cognitive load imposed by the primary task will be reflected in slower reaction times and lower accuracy rates on participants' performance on the secondary task. In light of this rationale, the results from the dual-task experiment provided a partially affirmative answer to our research question. As anticipated, participants achieved higher mean accuracy rates on the visual secondary task (responding to color changes) when carrying out the simple as compared to the complex versions of the three types of primary tasks, suggesting that the more complex versions of the primary task indeed posed greater cognitive demands. Contrary to expectations, however, no difference was observed in secondary task reaction times depending on the task complexity manipulation for any of the primary tasks. Notably, the same trends were observed in Révész et al.'s (2014) study, where an identical secondary task was utilised in combination with a

different type of primary task, also a computer-delivered oral task. Taken together the results of this research and those of Révész et al. (2014), it appears that secondary task accuracy rates are more sensitive measures of cognitive load than secondary task reaction times when the dual task condition is operationalized as this particular type of colour changing task.

A possible explanation for the lack of secondary-task effects for reaction times in our and Révész et al.'s (2014) study may lie in the nature of the secondary task employed. Stronger secondary-task effects might have emerged if there had been more competition present for attentional resources between the primary tasks and the secondary task. As explained earlier, Wickens (2007), in his multiple-resource framework, posits that the relative ease or difficulty of a task will depend on the extent of interference and/or competition within three dichotomous dimensions: processing stage (perception vs. response), modality (auditory/vocal vs. visual/manual), and processing code (verbal vs. spatial). In the current study, the primary task required participants to engage in both perceiving and responding to input (processing stage), rely on vocal as well as visual channels (modality), and deal with both verbal and spatial information (processing code). While the secondary task also called for perception (noticing color changes) as well as provision of response (reacting to certain color changes), it only involved the use of the visual/manual modality and the spatial code. Perhaps activities like recall of letters or words at certain intervals (e.g., Chandler & Sweller, 1996) would have led to clearer and more pronounced secondary-task effects, since this type of activities, similar to the primary task, would also have entailed the use of the auditory/vocal channel and the processing of verbal material, thereby resulting in greater interference along the modality and code of processing dimensions. In light of the fact that Sasayama (2013) did not find simultaneous effects for accuracy and reaction times either, it would also appear that simply presenting verbal material (like the letter A in her study)

without requiring the processing and vocalisation of that material via the verbal channel (e.g., by recall) may be insufficient to create the required level of interference.

The second research question addressed the issue of whether task manipulations designed to result in different level of cognitive complexity would be perceived as posing differential cognitive demands as measured by self-ratings. The results for the self-ratings revealed that, in line with our task complexity manipulation, participants rated the more complex task versions as requiring more mental effort. The same patterns were found for task difficulty on the map and decision-making tasks, but, contrary to our prediction, no significant difference emerged between the participants' ratings of task difficulty on the narratives. The absence of a complete correspondence between the results for the two scales was also evident in the strong but not perfect correlations between the task difficulty and mental effort self-ratings. It is not unexpected that the results for the two scales do not entirely overlap. Findings from previous research in educational psychology (Brünken et al., 2010) suggest that task difficulty and mental effort are related but not isomorphic constructs. As discussed earlier, task difficulty ratings appear to depend more on learner's level of expertise or prior domain-specific knowledge than judgments of mental effort. In light of this, it might have been due to the participants' greater familiarity with the narrative task type that they overall perceived the difference in task difficulty between the simple and complex narratives considerably smaller than the difference in mental effort between the two narrative versions, unlike for the map and decision-making tasks where there was somewhat greater correspondence between students' ratings on the two Likert scales. This possibility is supported by the lower overall difficulty ratings awarded by participants for the narrative as compared to the map and decision-making tasks (see Table 3).

It is also worth highlighting that the results obtained for the mental effort scale are more closely aligned with the findings of the dual task experiment and the expert judgments. This

raises the issue of whether ratings of mental effort may provide a more valid measurement of task-generated cognitive demands than ratings of task difficulty. The validity of the mental effort scale as a measure of cognitive load is supported by research in cognitive psychology. Researchers have consistently shown (e.g., Paas & van Merriënboer, 1994) that rating scales eliciting judgements of mental effort could successfully differentiate between different levels of mental effort when participants were asked to provide ratings for instructional conditions that were expected to affect cognitive load. More research is needed to clarify this issue.

The third research question was concerned with whether our intended task complexity manipulations were reflected in experts' judgments of required mental effort and task difficulty. The teachers, who served as expert judges, rated the complex version of each of the three task types as involving both more effort and difficulty. The correlations were also found to be very high between the mental effort and task difficulty ratings, indicating that, unlike the student participants, the teachers considered the two constructs to be closely linked. The open-ended questions revealed that the rationales for the teachers' judgments largely overlapped with the rationales for our task complexity manipulations. Following Robinson (2007), we hypothesised that that the complex narrative would impose greater reasoning demands on learners than the simple version of the task. Likewise, the teachers referred to differential reasoning demands most frequently (85%), when accounting for the difference in difficulty between the simple and complex narrative. As for the map task, we anticipated that the more complex version would be more cognitively taxing since it would require reference to a greater number (Robinson, 2001a) and more similar reference points and description of a more complicated route involving directions not only along the lateral but also to the vertical and sagittal axes. In line with this, the teachers most often mentioned the clarity of the visual information presented as a factor determining task difficulty (69%), making specific reference to the clarity of the map (28%) and the two versus three

dimensional nature of the diagrams (25%). The number of elements (28%) and the difficulty of the route (20%) in the maps were also identified as determinants of expected mental effort expanded by the teachers. Turning to the decision-making tasks, the more complex version was intended to create a greater need to engage in reasoning on the part of the participants by requiring decisions about vulnerable people under more difficult conditions with fewer resources available. The expert comments reflected this rationale, the teachers most often referred to reasoning demands (39%) and number of elements (33%) in describing sources of difficulty. In sum, the data obtained from the expert judgments, both quantitative and qualitative, appear to confirm the validity of our task complexity manipulations.

Conclusion

In response to recent calls to increase the internal validity of research designs testing cognitive-interactionist models of task-based learning (Robinson, 2001a; Skehan, 1998, 2009), the aim of this study was to explore methods that researchers can use to establish the validity of cognitive task complexity manipulations, that is, provide independent evidence that tasks that were designed to be more complex do, in fact, place greater cognitive demands on research participants. In particular, this study intended to examine the extent to which three methods - dual-task methodology, subjective self-ratings, and expert judgments - was useful for this purpose. The novelty of our study lay in that we investigated whether these three methods would lead to parallel results for three task types – a narrative, a map, and a decision-making task. While some researchers have already begun to utilise subjective and/or objective measures of task-generated cognitive load to validate manipulations of task complexity, studies have so far tended to focus on a single task type, thus not permitting generalisations across different task types.

Our study was successful in providing confirmation of the validity of our task complexity manipulations, since the data collected largely converged on the finding that the task versions we considered to be more complex required greater cognitive effort. For all three task types, both native and non-native speakers achieved lower accuracy rates on a secondary task when carrying the complex versions of the primary tasks, both native and non-native speakers perceived the high-complexity task versions as more cognitively demanding and most of the time more difficult, and experts judged the complex tasks to demand more mental effort and to be more difficult. Only the secondary reaction times and student difficulty self-ratings for the narratives yielded no significant difference between the simple and complex task versions, suggesting that these indices were not sufficiently sensitive measures of cognitive load in this context.

One finding that is particularly worth highlighting is that our objective measure (accuracy rates) and subjective measures (self-ratings and expert judgments) yielded parallel results. This suggests that subjective ratings and judgments constitute potentially valid research tools to estimate task-generated cognitive load. This result, if replicated in further research, also has important practical implications. Given that subjective measures are easy to administer, non-intrusive, inexpensive, and have high face validity with both teachers and second language learners; they potentially lend themselves well to use by second language practitioners when making task-grading and sequencing decisions in task-based and task-supported instructional contexts.

In addition to replicating this research using the same three methods, in future research it would also be worthwhile to triangulate the results obtained here with data gathered via other research techniques. One limitation of the present study is that a single objective measure, one type of dual-task methodology was employed. It would be interesting to explore the utility of additional objective measures in assessing cognitive load, such as different

forms of dual-task methodology and physiological measures, including the assessment of heart rate variability (Fredericks et al., 2005), neuro-imaging techniques (Murata, 2005), and the recording of eye movements (Révész et al., 2014). Another limitation concerns the lack of detailed information about students' cognitive processes during task performance. In future research, introspective methodology could prove useful in gaining more in-depth insights into students' thought processes and thereby lead to a more thorough understanding of what specific task parameters are responsible for causing increased mental effort during task work. Finally, given that the present research only focused on native speakers and L2 users with intermediate proficiency level and they carried out a limited number of task types, future studies are needed to examine whether the results found here would transfer to other proficiency levels and task types.

References

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453–476.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition, 35*, 689–725.

Brown, J. D., Hudson, T., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessments*. Technical Report No. 24. University of Hawaii, Second Language Teaching & CurriculumCenter.

Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 32*, 115–132.

Brünken, R., Seufert, T., Paas, F. (2010). Measuring cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 181–202). Cambridge: Cambridge University Press.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 1–20.

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25*, 315–324.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York: Academic Press.

Dave, A. (2004). *Oxford Placement Test 1.* Oxford: Oxford University Press.

Declerck, M.**,** & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. *Bilingualism: Language and Cognition, 15*, 782–796.

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19*, 195–221.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fredericks, T. K., Choi, S. D., Hart, J., Butt, S. E., & Mital, A. (2005). An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. *International Journal of Industrial Ergonomics, 35*, 1097–1107.

Gilabert, R., & Barón, J. (2013). The impact of increasing task complexity on L2 pragmatic moves. In A. Mackey & K. McDonough (Eds.), *Second language interaction in diverse educational settings* (pp. 45–69*)*. Amsterdam: John Benjamins.

Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 105–140). Amsterdam: John Benjamins.

Gilabert, R., Barón, J., & Llanes, À. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics in Language Teaching, 47*, 367–395.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching, 45*, 215–240.

Gilabert, R. (2006). The simultaneous manipulation of task complexity along planning time and (/- here-and-now/): Effects on L2 oral production; investigating tasks in formal language learning; second language acquisition. In M.P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Clevedon, England: Multilingual Matters.

Housen, A., & Kuiken, F. (2009) (Ed.). Special issue: Complexity, accuracy and fluency in second language acquisition research. *Applied Linguistics, 30*, 461–601.

Kim, Y. (2009). The effects of task complexity on learner-learner interaction. *System*, *37*, 254–268.

Kim, Y., & Tracy-Ventura, N. (2011). Task complexity, language anxiety, and the development of the simple past. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 287–306). Amsterdam: Benjamins.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levkina, M., & Gilabert, R. (2014). Task sequencing in the L2 development of spatial expressions. In M. Baralt, R. Gilabert, & P. Robinson (Eds.), *Task sequencing and instructed second language learning.* New York: Bloomsbury.

Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In A. Housen, I. Vedder, & F. Kuiken (Eds.), *Dimensions of L2 performance and proficiency investigating complexity, accuracy, and fluency in SLA* (pp. 171–198). Amsterdam: John Benjamins.

Malicka, A. & Levkina, M. (2012). *Measuring task complexity: does L2 proficiency matter?* In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 43–66). Amsterdam: John Benjamins.

Michel, M. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: researching the Cognition Hypothesis of language learning and performance* (pp. 141–174). Amsterdam: John Benjamins.

Murata, A. (2005). An attempt to evaluate mental workload using wavelet transform of EEG. *Human Factors, 47*, 498–508.

Norris, J. M. (2010, September). *Understanding instructed SLA: Constructs, contexts, and consequences*. Plenary address delivered at the annual conference of the European Second Language Association (EUROSLA), Reggio Emilia, Italy.

Norris, J., & Ortega, L. (2003). Defining and measuring L2 acquisition. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Malden, MA: Blackwell.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84,* 429–434.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P.W.M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63–71.

Paas, F., & van Merriënboer, J. J. G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6,* 351–372.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition, 31,* 437–470.

Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *Modern Language Journal*, *95*, 168–181.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics, 35*, 87–92.

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning*, *64*, 615–650.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for investigating task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). New York: Cambridge University Press.

Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27–57.

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 631–678). Malden, MA: Blackwell.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45,* 193–213.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*, 1–36.

Sasayama, S. (October, 2013). *Is a 'complex' task really complex? Measuring task complexity independently from linguistic production*. Paper presented at the 5th Biennial International Conference on Task-Based Language Teaching, Banff, Alberta, Canada.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide.* Pittsburgh, PA: Psychology Software Tools, Inc.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*, 510–532.

Wickens, C. (2007). Attention to the second language. *International Review of Applied Linguistics, 45*, 177–191.

Table 1. Accuracy rates for the baseline and dual tasks on the narrative, map, and decision-making tasks

		Baseline			Simple			Complex		
	N	M	SD	95% CI	M	SD	95% CI	M	SD	95% CI
Narrative										
NS	22	.93	.07	[.89, .97]	.90	.09	[.86, .95]	.87	.08	[.82, .92]
ESL	24	.90	.10	[.87, .94]	.86	.12	[.82, .95]	.84	.14	[.80, .89]
Total	46	.92	.09	[.89, .94]	.88	.11	[.85, .91]	.86	.11	[.82, .89]
Map										
NS	24	.93	.07	[.90, .96]	.85	.11	[.81, .90]	.83	.09	[.79, .87]
ESL	22	.92	.09	[.89, .96]	.82	.11	[.78, .87]	.78	.12	[.74, .83]
Total	46	.93	.08	[.90, .95]	.84	.11	[.81, .87]	.81	.11	[.77, .84]
Decision-making										
NS	22	.93	.08	[.89, .97]	.87	.09	[.83, .91]	.82	.10	[.78, .87]
ESL	22	.90	.11	[.86, .94]	.84	.11	[.80, .88]	.82	.11	[.77, .87]
Total	44	.92	.09	[.89, .95]	.86	.10	[.83, .89]	.82	.11	[.79, .86]