Research Report

Simulating Real-World Context in an Email Writing Task: Implications for Task-Based Language Assessment

ETS RR–23-05

John M. Norris
Shoko Sasayama
Michelle Kim

December 2023

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Simulating Real-World Context in an Email Writing Task: Implications for Task-Based Language Assessment

John M. Norris[1], Shoko Sasayama[2], & Michelle Kim[3]

1 Educational Testing Service, Tokyo, Japan
2 Waseda University, Tokyo, Japan
3 Educational Testing Service, Princeton, NJ, USA

Accomplishing a communication task in the real world requires the ability not only to do the task per se but also to manage aspects of the context in which it occurs. For this reason, simulations of target language use contexts have been incorporated into the design of communicative language tests as a way of enhancing the authenticity of assessment task performance. Although some contextual factors may increase extraneous cognitive load and distract learners from focusing on the task at hand (Sweller, 1994), they represent important design considerations in task-based language assessment (TBLA), where the purpose of assessment is to determine what second language (L2) learners can do with the target language in the real world. In that sense, the extraneous cognitive load might well be part of the construct we are interested in assessing. Accordingly, the current study simulated aspects of a real-world task performance context as part of an email writing task assessment. Simulated context was operationalized as (a) additional information about the task scenario, (b) a visual image to simulate the physical context, and (c) an audio to replicate the real-world experience. A total of 276 L2 English learners performed the email task, half with simulated context and the other half without it. Findings revealed that, when presented with simulated context, the tasks were perceived by participants to have induced more time pressure and to be more interesting. In terms of performance effects, the provision of simulated context negatively affected the syntactic complexity of participants' writing but positively affected their syntactic fluency. It also led to greater discrimination among learners at different proficiency levels on various measures of language performance. The paper concludes by highlighting implications for task design and validity evaluation, especially in TBLA.

Keywords cognitive load; email writing; L2 writing; simulated context; task-based language assessment

doi:10.1002/ets2.12366

Communicating in a second or foreign language in the real world is a challenging and complex endeavor. Successful communication depends, of course, on sufficient knowledge of the grammatical and lexical forms and rules required for making meaning in the given language, as well as the capacity to comprehend and express them in spoken and written modalities. It also depends on a degree of awareness and sensitivity to the interlocutors involved in the event and their purposes and interests in communicating, and on the language user's capacity to strategically marshal their linguistic and other cognitive resources to engage in meaning-making (Bachman & Palmer, 1996, 2010; Canale & Swain, 1980; Hymes, 1972; Purpura, 2014). These competences are important, but accomplishing language tasks in the real world often involves additional challenges that are not about the language learner's knowledge and abilities per se. Real language use, that is, language use that happens outside of the relatively sanitized confines of a standardized test, pages of the textbook, and physical, or virtual boundaries of the classroom, occurs in a variety of contexts that feature any of a variety of factors that may influence how, and how successfully, communication takes place. Many tasks take place in physical environments that are characterized by sights and sounds and actions that can support, augment, or distract from language use. Many tasks take place within a particular time frame, with real or presumed implications for the deployment of language to get things done in a timely and appropriate manner. Many tasks also involve other actors in the event, about whose motivations and characteristics the language user may know little or a lot, and with whom the language user eventually has to make meaningful communication happen. These, and other features of the real-world task context, play a real role in shaping what and how and how well language users communicate in the target language. As such, if a central goal of language testing is to assess how well learners can communicate in the real world, then it probably makes sense to expose learners to these challenging features of tasks as they occur in various communication contexts.

Corresponding author: J. Norris, E-mail: jnorris@etsjapan.jp

In the current study, we explore the effects of introducing several simulated aspects of context into a second-language task performance. The kinds of simulated features that we include are relatively low-threshold additions to the presentation of tasks delivered in a standard laptop computer environment, of the sort that language teachers or assessment developers could easily replicate for teaching and testing purposes. Our goal in the current study is to illuminate whether and how the introduction of simulated context influences second-language learner perceptions of the task as well as whether and how it affects language performance in accomplishing the task. Prior to introducing the study, we situate these ideas within relevant portions of the language assessment literature, and we consider key ideas from cognitive psychological research on the role of simulated context in human performance.

Literature Review

Simulating Context in Language Assessment

Simulating the context of communication has played an important role in language testing for some time now. Early calls for increasing the authenticity of language assessment (e.g., Morrow, 1978, 1979; Savignon, 1985) highlighted the simulation of communication tasks and language use settings as an important practice in developing language tests that could tell us something about learners' abilities to use the target language under real-world circumstances. Indeed, the need for authenticity in communicative language testing was adopted as an essential characteristic in conceptual frameworks for language test development and validation in the 1990s (e.g., Bachman, 1990; Bachman & Palmer, 1996). From this perspective, the authenticity of a language assessment task could be conceived of both in situational terms, referring to aspects of the physical and communicative environment, and in interactional terms, referring to aspects of the participants and their communicative intentions and behaviors (Lewkowicz, 2000). Language assessment tasks, then, called for comparison of their situational and interactional authenticity with the target language use tasks and settings to which they were intended to generalize or extrapolate (Bachman & Palmer, 1996) as one approach to understanding and delimiting the validity of interpretations based on language tests.

In practical terms, simulation of target language use tasks and their communicative contexts has taken multiple forms in language assessment, with particularly robust interest in the testing of oral proficiency and spoken interaction. Oral proficiency interviews of various kinds generally feature context simulation in the form of role-plays, where test takers are asked to assume the role of a hypothetical speaker interacting with a hypothetical audience, as well as in the form of task descriptions that present the test taker with scenarios in which they must speak (see, e.g., Isbell & Winke, 2019). The simulated oral proficiency interview (SOPI) is a good example. In the SOPI, context is simulated in several ways: (a) through written descriptions of a communication setting as well as an audience or interlocutor for distinct test tasks; (b) through aural prompt questions asked by a simulated interlocutor of a particular age, gender, and relationship with the test taker; (c) through a specified purpose for communication (e.g., giving directions, describing a process, making an argument); and (d) through the use of visual graphic elements (e.g., Stansfield & Kenyon, 1992). Similarly, tests of pragmatic and interactional competence have relied on simulations to impart critical aspects of the setting that have a bearing on how examinees perform (see Youn, 2020). For example, Hudson et al. (1992, 1995) developed a test of cross-cultural pragmatic competence in English that simulated diverse interlocutor relationships according to power, social distance, and degree of imposition of the given speech act (requests, apologies, refusals); these were then simulated in the form of relatively rich descriptions of scenarios that asked the test taker to imagine an encounter with a given interlocutor for a given purpose. These kinds of low-fidelity simulations, very common in language testing practice, depend heavily on the extent to which the test taker both understands the simulated context, as depicted or described, and accepts or "buys into" the simulation as a participating character. Whether and how test takers respond to these kinds of simulations, as well as what effect they have on language performance, should play a role in validity evaluation of language tests.

Simulation of target tasks and language use contexts with higher degrees of fidelity has received particular attention within the special and professional purposes language testing domains (e.g., Douglas, 2000, 2012; Jacoby & McNamara, 1999; Knoch & Macqueen, 2019; Wu & Stansfield, 2001) as well as in task-based language assessment (TBLA; see Norris & East, 2021) more generally. Here, higher fidelity simulation is important for several reasons. From the perspective of special purposes language testing, the context of communication is one of the defining characteristics that determine what kind of language use is essential for accomplishing tasks that typify the given domain (e.g., specific workplaces). For example, in testing the English ability of nurses who are seeking to work in Canada, the Canadian

English Benchmark Assessment for Nurses (CELBAN) test (The CELBAN Centre, 2018) presents candidates with a variety of scenarios specific to the nursing profession to test their speaking, listening, reading, and writing abilities. The listening section of the CELBAN provides an interesting example of simulated context: Listening input is given in the form of both video and audio recordings of interactions that take place (a) between a variety of individuals, such as nurses, patients, and family members and (b) in a variety of depicted settings, such as medical offices, hospitals, and patient homes.

In TBLA, the context of communication is essentially considered one dimension of the language ability construct (Norris, 2009; Norris et al., 1998; Robinson & Ross, 1996). That is, where the goal of assessment is to elicit students' abilities to engage in meaningful, real-world experiences so that they can demonstrate what they can do with the language, the contextual aspects of the communicative task must be present in the assessment (at least to some degree). Beyond just informing more accurate interpretations about task-specific language ability, TBLA also draws on tenets of the authentic assessment movement, where authentic assessment tasks intentionally wash back on what is learned and how it is taught (e.g., Wiggins & McTighe, 2005). According to McTighe and Willis (2019, p. 158), authenticity is determined by "A task or assessment that simulates or replicates important real-world challenges containing genuine goals or purposes, audiences, and constraints." Assessment tasks of this sort are particularly important where one goal of language testing is to guide language education toward more communicative language ability outcomes (see Norris & East, 2021). Simulations of tasks and their contexts, then, play a particularly important role in TBLAs in a variety of educational and workplace settings (Norris, 2016, 2018b), to both enhance construct validity and to encourage positive washback.

Simulations of a particularly high-fidelity variety are also becoming increasingly feasible in the rapidly developing application of technology mediation to language assessments and language education more generally (e.g., Blyth, 2018; Sydorenko et al., 2019; Wang et al., 2020). For one interesting example of a high-fidelity task-based assessment for special purposes language testing, Park (2018) described the simulation of a military air traffic control setting and associated tasks for the purpose of testing the English communication ability of air traffic controllers in a multilingual military base environment. Park developed interactive visual depictions of the air traffic control tower, the helicopter take-off and landing areas and military base, and various computer input screens to embed the test taker visually within the task context. He also audio recorded various types of pilot and other voice communications, along with authentic background sounds, as the listening test input. Putting these simulated aspects together within the computer-mediated second-life virtual immersive environment, he was able to expose test takers to a very high-fidelity simulation of the types of communication demands encountered while working as a military air traffic controller. Park also showed that the simulation was capable of eliciting a rich array of communication strategies and interactions with the simulated setting and input sources, leading to robust interpretations about test takers' abilities in accomplishing such tasks in the real world.

Although simulated context has featured substantially in speaking (e.g., Xi et al., 2021), and to some extent listening (e.g., Ockey & Wagner, 2018; Papageorgiou et al., 2021) assessment, comparatively little attention has been paid to the representation of communication context in standard approaches to L2 writing or reading assessment (see Cumming et al., 2021; Schedl et al., 2021). This reality may reflect the predominance of a handful of task types that have tended to typify (at least large-scale, standardized) literacy assessment, including the ubiquitous academic essay writing task and academic reading comprehension passage, neither of which has seemed to call for much contextualization of the communication event per se. Such trends notwithstanding, with the acknowledgment that both writing and reading competence involve a lot more than these selected genres (e.g., Anderson, 2015; Norris et al., 2021; Norris & Manchón, 2012), recent attention has shifted to the incorporation of a diversity of assessment tasks that occur in distinct communication environments (e.g., Cumming et al., 2021; Schedl et al., 2021).

In the specific case of L2 writing (Hyland, 2019), richer and more contextualized tasks are being explored to meet distinct assessment needs and in particular to enable more nuanced interpretations of learners' developing writing abilities. In recommending areas for further development and improved construct coverage of academic writing assessment, Cumming et al. (2021) argued for the expansion of test tasks to include practical-social writing, explanatory writing, transactional writing, and expressive writing. For example, in transactional writing they call out the need for assessment of common tasks like email writing, noting that "These types of written genres require writers to pay close attention to their intended audiences and purposes in a specific social context … " (p. 136).

In order to operationalize context for these kinds of task types — that is, to provide test takers with sufficient information about audiences, purposes, and contexts for communicating — test developers have begun to incorporate simulations into corresponding assessments. Returning to the specific case of email writing, several examples illustrate the need for and provision of simulated context. In order to assess different dimensions of L2 learners' pragmatic competence in crafting email messages, Youn (2014) set the stage for her writing (and pragmatics) assessments by providing test takers with rich descriptions of the audience and purpose for writing, the setting of the email task, graphic, and written input to incorporate or refer to, and realistic time limits for completing the tasks (see another example of assessing pragmatics and email writing in Haider, 2019). In a second example, Oliveri et al. (2021) described the design of formative assessment tasks that focus on effective workplace communication, specifically email writing for accomplishing various transactional goals and solving problems. Their email writing tasks similarly included information about audience, purpose, setting, and visual/graphical realia to situate the tasks in a meaningful context to which the test takers responded.

Simulating communication context, then, is one important technique employed by test developers to enhance the authenticity of language assessments, thereby improving the accuracy of interpretations about test takers' abilities to communicate effectively under real-world conditions, and potentially to wash back positively on language learning and teaching. Clearly, the more task based a language assessment is, the greater the call for contextualization given the situated nature of interpretations about task-specific abilities. The examples cited above indicate a variety of possibilities for lower to higher fidelity context simulations in language assessment. Generally speaking, these simulations have been included in assessments as aspects of the task design, but not in and of themselves as the object of validity research, beyond some examination of the perceived authenticity of test tasks from the perspective of the test taker or score users. Research in language testing has not specifically investigated the relative effect of aspects of simulated context versus absence of the same in terms of test-taker cognitive response or language performance. Prior to introducing a study with precisely this focus, we briefly consider relevant research on the potential impact of simulation from cognitive and performative perspectives.

The Impact of Simulated Context on Learner Response and Performance

Although it is well beyond the scope of the current paper to explore in any detail the relationship between simulations of various types of context and effects on learning or test taking, it is worth noting that there has been substantial interest in this relationship within several domains of educational research in particular. Recent meta-analyses have shown that simulation-based education that exposes learners to authentic experiences involving complex skills application has an overall large positive effect on learning outcomes in higher education settings (e.g., Chernikova et al., 2020). However, depending on the specific training or learning context and goals, the effectiveness of simulations may be complicated by factors such as the fidelity of authenticity to a given context, participants' perceptions of authenticity, and their interaction (e.g., for teacher training in virtual simulated classroom environments, see Howell & Mikeska, 2021). Similarly, with the rise of digital language learning (see Li & Lan, 2022), and the potential of virtual immersive environments in particular, much attention has been paid to the possibilities of high-fidelity simulations of authentic communication settings for fostering effective language learning (Blyth, 2018; Tai et al., 2022). Here, too, recent meta-analyses point to overall substantial positive effects of simulations on both linguistic gains and affective responses by learners (e.g., for three-dimensional (3D) immersive language learning, see Wang et al., 2020).

These findings regarding the general potential for simulations to impact learning in positive ways are encouraging. However, from a task-based language testing perspective (see, e.g., Norris et al., 1998; Skehan, 1998), rather than having a focus on overall learning outcomes, we are interested here in the specific effects that simulated communication context may have on L2 test-taker performance for a given task. That is, as we go about enhancing the representation of context in a language assessment task—so that we can better understand test takers' abilities to use language to communicate effectively in situ—what can we anticipate to be their affective and cognitive responses to the simulated context, and what might be the effects on their use of language as well as their overall success in accomplishing the communication task?

As a starting point, it might be helpful to think about tasks, both real-world and assessment tasks, from a cognitive perspective. In cognitive psychology, Sweller (1988, 1994, 2010) has proposed the *cognitive load theory* for learning tasks in particular, arguing that, given limited working memory capacity, the amount of information being processed at any one time—i.e., cognitive load—should be managed carefully in task design. Of direct relevance to our discussion here is that cognitive load can be of different types: (a) intrinsic, (b) germane, and (c) extraneous. Intrinsic cognitive load is inherent to the task and is induced by the types of knowledge and skills required to complete the task successfully. Germane load is treated as a desirable, beneficial type of cognitive load that directs learners' attention to the task at hand and thus facilitates their learning. Extraneous load, on the other hand, is generally portrayed to be an undesirable or even harmful type of cognitive load, and it is induced by how the task is presented to learners (e.g., task instructions are scattered and have to be compiled by the learner versus found in a single space; Ayres & Sweller, 2005), or by inclusion of excessive or complicated information (e.g., task input for a picture-based oral narrative task contains too many characters to decipher the storyline; Sasayama & Norris, 2019), or through any of a variety of performance conditions that may distract from the primary task.

Although extraneous cognitive load is generally treated as a phenomenon to be reduced or controlled from the perspective of learning-oriented cognitive load theory, returning to the importance of context in language assessment, it may have a clear purpose and role to play, especially in TBLA where the construct of interest is L2 learners' ability to perform communication tasks in the real world—a world that is arguably full of diverse sources of extraneous load. Within the proximate task-based language teaching literature, several prominent cognitive research agendas have been pursued for some time, inspired by the work of Skehan (2014) and Robinson (2011). Here too, sources of extraneous load have been treated as a feature of task conditions that, when reduced or controlled, can lead to improvements in linguistic performance. For example, ample research on the provision of planning time prior to task performance (i.e., leading to a reduction in the extraneous load associated with having to perform the task "cold" or spontaneously) has shown clear positive effects for the linguistic complexity and accuracy of communicative performances (see Sasayama & Norris, 2023).

Positive effects on performance notwithstanding, the reduction of extraneous load may not accurately depict the real-world conditions under which communication tasks are often realized. Thus, to gauge test takers' ability to deal with tasks in real life, it seems critical to *increase* or at least maintain, rather than decrease, extraneous load to replicate aspects of many real-world scenarios. What kinds of effects the presence of different real-world sources of extraneous load may have on communicative assessment task performance—whether deleterious or perhaps facilitative—remains open to investigation.

The potential effects of simulating aspects of the real-world context of a communication task may be realized both in terms of how test takers respond to and perceive the task conditions and associated sources of cognitive load as well as how test takers perform. Performance by L2 learners on open-ended constructed response language tasks is most typically captured in two primary ways. First, human judgments of performance, generally based on holistic rating scales that describe degrees of task accomplishment, provide an overall estimation of ability to perform the task (e.g., Kuiken & Vedder, 2017; Révész, 2014; Sasayama & Norris, 2023). In addition, more fine-grained aspects of linguistic performance may reveal important dimensions of language ability that are affected by task conditions, in particular, measures of the grammatical and lexical complexity, accuracy, and fluency of language use (the so-called CALF measures; see Housen et al., 2012; Norris & Ortega, 2009; Sasayama & Norris, 2023).

However, inferring effects of simulated context directly from patterns of language performance is at best indirect in that it assumes that context features actually trigger some kind of cognitive or affective response by the test takers in the first place. Determining whether the particular simulation affects learners in anticipated ways is, therefore, also important to measure (see Sasayama, 2016, for related arguments about task design features). Borrowing from the cognitive load research domain, one important aspect of test-taker response would entail their perceptions of the difficulty and mental effort realized in attempting to perform the target task (e.g., Kalyuga et al., 1999; Paas et al., 1994) as a means of detecting whether greater or lesser cognitive load is associated with the introduction of simulated context. In addition, related to the simulations explored in the current study, three other phenomena merit attention in test-taker responses. First, the extent to which pressure, especially time pressure, to perform the task is felt to a noticeable degree may have important effects on task outcomes (e.g., Phillips-Wren & Adya, 2020). Second, interest in the situation generated by the simulated context and the task itself may offer insights into test-taker engagement and motivation to fulfill the task expectations (e.g., Schraw & Lehman, 2001). Third, given the fundamental rationale for incorporating simulations of context into language assessment tasks in order to enhance the feel of authenticity and encourage test-taker buy-in to the task, it is important to gauge perceived authenticity as a potential factor contributing to performance effects (e.g., Lewkowicz, 2000).

Research Questions

The current study investigated the effects of simulating aspects of a real-world context on L2 English learners' performance in an email writing task as well as their perceptions of task difficulty, mental effort, time pressure, interest, and authenticity. Context was simulated through additional descriptions of the task scenario, inclusion of a visual representation of the performance location, and addition of background audio that played during the task performance.

The study was guided by the following research questions:

RQ1. Does simulated task context, in the form of (a) additional information about the scenario, (b) a visual to simulate the context, and (c) an audio (background noise) to replicate the real-world experience, have an effect on L2 English learners' perceptions of (a) task difficulty, (b) mental effort, (c) time pressure, (d) interest, and (e) authenticity?

RQ2. Does simulated task context have an effect on L2 English learners' task accomplishment (in the form of performance ratings) or linguistic performance (in the form of indices of complexity, accuracy, lexis, and fluency) on an email writing task?

RQ3. Does L2 proficiency mediate the relationship between task design (with or without simulated context) and L2 English learners' task performance?

RQ4. To what extent does the addition of simulated context affect discrimination among learners at distinct levels of proficiency based on performance ratings and linguistic measures?

Methods

Participants

A total of 276 English language learners (174 female, 102 male) participated in the study. Participants were recruited from English language programs within and outside of the United States to include learners with a variety of proficiency levels. Of those, 132 were studying English in the United States., and the others were studying English in Ecuador ($n = 119$), Mexico ($n = 1$), or Colombia ($n = 24$). The participants' ages ranged from 18 to 52, with an average age of 22.54 ($SD = 6.1$). Participants had various first languages (L1), but the majority were L1 speakers of Spanish, Chinese (Mandarin or Cantonese), Japanese, or Korean. They had studied English for 8.87 years ($SD = 5.65$) on average. For recruitment purposes, participants' proficiency levels were estimated based on (a) the English language courses in which they were enrolled at their universities, (b) standardized English language assessment scores (e.g., *TOEFL iBT*®), and (c) self-assessment of their proficiency levels. Based on these estimates, participants were divided into three proficiency groups (i.e., low, mid, high) prior to assignment to a control or experimental condition. To gauge the level of English language courses offered at different institutions, a site coordinator at each institution was asked to provide either TOEFL iBT ranges or the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) levels for each course. In addition, all students self-assessed their CEFR level in reference to an abbreviated summary of the six levels from A1 to C2. These different sources of information were aggregated to assign students to proficiency groupings, with differences decided in favor of English language course and the corresponding CEFR level. Table 1 presents the ranges of estimated CEFR levels and TOEFL iBT scores, and sample sizes, for each of the low, mid, and high proficiency groupings.

Materials and Instruments

Cognitive Laboratory Pilot Testing

To obtain trustworthy data, it was deemed critical to ensure accessibility of the study materials to all potential participants, and in particular low proficiency learners. Accordingly, prior to the main study, cognitive lab studies were conducted with representative English learners at various proficiency levels to confirm that participants were able to follow task instructions and engage with elements of the study as intended. The findings of the cognitive lab sessions suggested that participants, even the ones with quite low proficiency levels, were able to follow task instructions and attempt all tasks. However, questionnaires—when presented in English—were found to be challenging, especially for low proficiency learners, to understand and respond to in English. Thus, in the main study, all questionnaires were administered in participants' first languages (i.e., Spanish, Chinese, Japanese, and Korean).

Email Task

All participants completed the same email writing task. This task presented participants with an urgent issue to be dealt with using email that was thought to be familiar to university students. Specifically, participants were given a scenario in

Table 1 Estimated English Proficiency Levels of Low, Mid, and High Proficiency Groups

Proficiency level	CEFR level	TOEFL iBT range	*n*
Low	A1–A2	41 or below	71
Mid	B1	42–71	94
High	B2–C2	72–120	111

Note: Students assigned to the high group included those estimated at CEFR B2 and higher due to the interpretation of those proficiency levels as being adequate for entry into university study in the contexts under investigation.

which they were scheduled to give a presentation in an English class, but they would not be able to make it to class on time due to a delayed train (see Table A1 in appendix for the actual task scenario). Participants were asked to write an email to their professor to (a) apologize, (b) explain the situation, and (c) propose a solution to the problem. They were given 5 minutes to write their response.

Context

To investigate the potential role played by additional task context, half of the participants engaged in the email writing task with simulated context and the other half without it. In this study, *simulated context* was operationalized in three ways: (a) additional written information about the scenario, (b) a visual to represent the physical setting of the performance, and (c) an audio to enhance the sense of a real-world experience. The additional information was provided following presentation of the main task scenario, and it included information about the professor's personality ("Remember, Dr. Smith is very strict") and classroom policy ("He asked you to email him before class starts if you are going to be late for class") as well as timing of the event ("Your class starts in 5 minutes"). On the response page, participants in the simulated context condition were provided with a visual that simulated the scenario—an image of people on a train. In addition, an audio started playing automatically as soon as participants opened the response page. The audio consisted of recorded background noise that one would hear when taking a train, including the noise of a train moving and stopping, doors opening and closing, and spoken train conductor announcements. The purpose of these features of added task context was to simulate a more real-world experience of writing an email under authentic conditions.

Perception Questionnaires

After completing the email writing task, participants were given a questionnaire that asked about their perception of the task. The questions included (a) "How difficult was this particular task of writing an email message to your professor?", (b) "How much brainpower or effort did you use in writing the email?", (c) "How much time pressure did you feel while writing the email?", (d) "How interesting was this email task for you?", and (e) "To what extent was this email writing task similar to what you might do in real life?". Participants responded on a 7-point Likert scale, ranging from 1 (*Very easy*, *Very little*, *Not at all interesting*, or *Not at all similar*) to 7 (*Very difficult*, *A lot*, *Very interesting*, or *Very similar*). The perception questionnaires were administered in five languages, including English, Spanish, Chinese, Japanese, and Korean, to make sure that the participants—especially those with low levels of English proficiency—were able to view and answer the questions in their first language.

C-Test

A C-test (Norris, 2018a) was administered to all participants as a measure of their L2 proficiency. This C-test followed the standard design of deleting the second half of every second word in coherent, paragraph-length texts. The C-test consisted of two texts, and each had 20 blanks for a total of 40 blanks. Participants were given 7 minutes to complete each text. Cronbach's alpha reliability estimate for the 40-item C-test was .92.

Background Questionnaire

Participants filled out a background questionnaire to provide demographic information, including questions on gender, age, and duration of English language learning.

Table 2 Number of Participants in Each Condition and Their Average C-Test Scores

	N		Average C-test scores	
Proficiency level	Group A no context	Group B context	Group A no context	Group B context
Low	35	36	19.17	20.33
Mid	46	49	25.11	24.59
High	54	56	31.48	30.80

Procedures

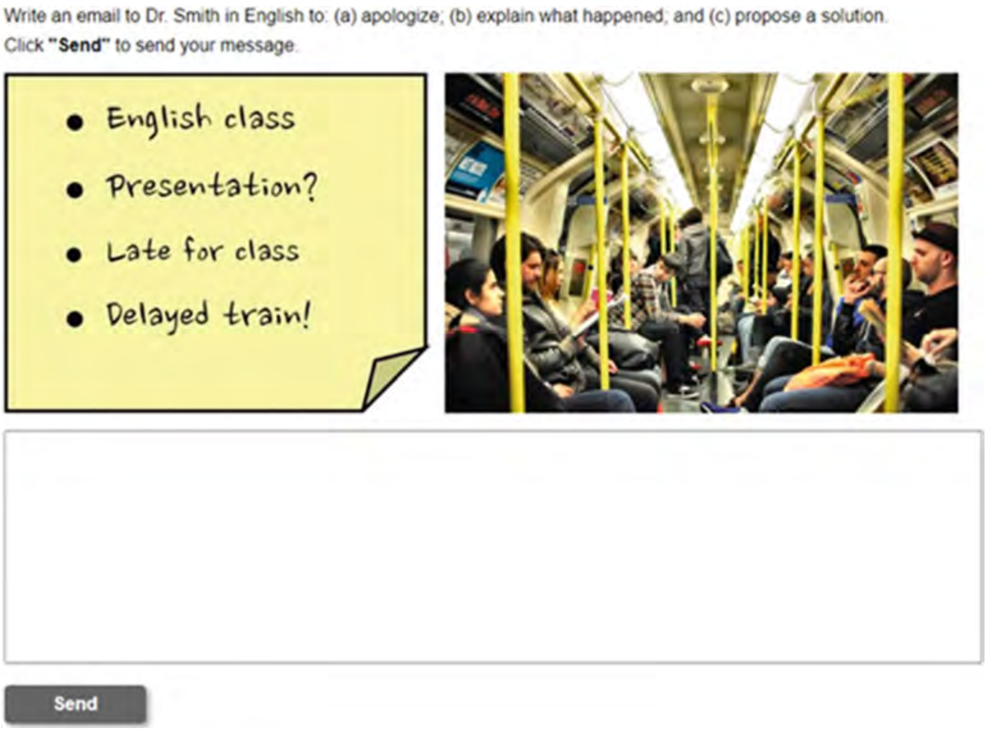
Participants completed an eligibility survey prior to participating in the experiment. The purpose of the eligibility survey was to collect information about the participant's English proficiency levels, including English course(s) in which they were enrolled, standardized English language assessment scores, and self-assessment of their English proficiency levels (A1 through C2 on the CEFR). To elicit self-assessment of their CEFR levels, participants were presented with the CEFR global scale (i.e., a short description of what learners at each proficiency level should be able to do; Council of Europe, 2001) and were asked to choose the level that best described their English ability. Based on the proficiency level information provided, participants were divided into high, mid, or low proficiency groups (see Table 1 and previous explanation).

Once participants' eligibility was confirmed, they were invited to participate in the study. Each participant took part in the study online at home, using their own computer. Participants were asked to complete all tasks alone in a quiet space. In the study session, each participant first completed the email writing task (with or without simulated context), then the perception questionnaire and the short background questionnaire, and finally the C-test. To ensure balanced representation of participants from each proficiency group in each of the two conditions, participants in each of the low, mid, and high proficiency groups were randomly assigned to either the *without simulated context* (Group A) or *with simulated context* (Group B) conditions. Table 2 shows the number of participants who were assigned to Group A versus Group B in each proficiency level, along with their average C-test scores.

Participants in Group A engaged in the email writing task without any additional context. In other words, all they saw was the basic description of the task scenario, without the additional information about the scenario, visual, or audio. They were given unlimited time to read the task scenario and were instructed to hit "Next" once they were ready. When they hit the next button, they were taken to the response page, and the scenario disappeared to prevent participants from copying and pasting directly from the task prompt when working on their response. Instead, they were presented with brief task instructions and a sticky note that summarized the main points of the scenario. The summary sticky note included the following bullet points: English class, presentation, late for class, and delayed train.

Participants in Group B, on the other hand, experienced the same task but with all additional simulated contextual features. As was the case in Group A, participants in Group B were given unlimited time to read the basic task scenario and were instructed to hit the next button to move on. However, for this group, once they hit the button, they were presented with the additional information about the scenario (i.e., Dr. Smith's personality, his class attendance policy, and timing of the event). On the next response page, participants were given the same brief task instructions and summary sticky note that were shown to Group A. In addition, they were presented with the image of people on a train as well as the background noise (see Figure 1). The background noise was programmed to play as soon as participants were taken to the response page and to play until the response time was up (5 minutes) or when they hit the "Send" button to indicate their completion of the task.

After the task, all participants were asked to answer the questionnaire to provide their perceptions of the email writing task that they had just completed. The questionnaire allowed participants to view the questions in English or their first language (Spanish, Chinese, Japanese, or Korean). Participants were required to respond to all questions before moving on to the background questionnaire. The questionnaire was then followed by the C-test. Participants were given 7 minutes to complete each of the two C-test texts.



Write an email to Dr. Smith in English to: (a) apologize; (b) explain what happened; and (c) propose a solution.
Click "**Send**" to send your message.

Figure 1 Response page for Group B.

Data Scoring, Coding, and Analysis

Task Accomplishment Ratings

Scoring rubrics were developed to assess the extent to which an email response was effective in terms of (a) language, (b) content, and (c) writing conventions (see Table A2 in appendix). A 6-point rating scale ranged from Score 5 (highest) to Score 0 (lowest). Descriptions of language, content, and writing conventions were developed iteratively to characterize performance holistically across the six levels. A fully successful Score 5 response was characterized as a response that (a) displayed a consistent facility in the use of the language, including the use of effective and accurate grammar, vocabulary, and pragmatics; (b) included all three required email components (apology, explanation, and solution) with sufficient elaboration to effectively accomplish the task; and (c) followed genre-appropriate writing conventions (appropriate opening and closing, professional title to address the professor). Lower scores were awarded as the effectiveness of the email decreased in terms of (a) language, (b) content, or (c) writing conventions and as the response became more difficult to interpret due to a lack of facility in the language use and/or information included in the response. A score of 0 was reserved for a response that was too short to judge its topic relevance (e.g., "dear dr smith this email is for"), rejected the task itself (e.g., "I don't know"), was not written in English, was entirely copied from the prompt, or consisted of arbitrary keystrokes. Blank responses were scored as a nonscorable rather than 0.

Using this scoring rubric, the email responses were double scored by a total of four trained raters. Raters first reviewed the rubrics with sample responses at each score band and then participated in a calibration session where they practiced scoring responses and discussed discrepancies. Each individual response was then scored by two raters. The scores between the two raters were averaged to come to the final task accomplishment score for each response. In rating responses, the raters were largely in agreement. On the 6-point scale, the ratings given by pairs of raters were either exactly the same (45%) or adjacent (i.e., within +/− one point) for 94% of the responses scored.

Linguistic Indices

Participants' email responses were also analyzed in terms of the linguistic indices commonly used to capture key features of L2 learners' written task performance (e.g., Norris & Ortega, 2009): (a) syntactic complexity, (b) accuracy, (c) lexis, (d) syntactic fluency, and (e) speed fluency. These measures were not intended to provide comprehensive coverage of any of the CALF constructs; instead, they were selected as representative or "marker" measures, so that the distinct dimensions of linguistic performance could be investigated relative to each other and to overall performance on the task (see related discussion in Sasayama & Norris, 2023). For syntactic complexity, accuracy, and syntactic fluency, the *T-unit* was chosen as the basic unit of analysis. Following Hunt (1970), T-unit was defined as "a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it" (p. 4). When coding for T-units, only the body of the message was taken into account; thus, opening and closing expressions (e.g., "Dear teacher," "Good morning Professor Smith," "sincerely," sorry," "thank you") were excluded from the total number of words counted.

Participants' global syntactic complexity was measured by the mean length of T-unit (MLT). Although syntactic complexity is a multidimensional phenomenon, the MLT has been shown consistently to reflect global linear development as the participants' L2 proficiency increases (e.g., Byrnes et al., 2010; Norris & Ortega, 2009). Given the wide range of proficiency levels of the participants in the study, the MLT was considered to be a trustworthy indicator of syntactic complexity for this study. To calculate the MLT, the total number of words produced in the email was divided by the total number of T-units produced by each participant.

Accuracy was analyzed in terms of global grammatical and lexical accuracy through the use of the error-free T-unit ratio (e.g., Crossley, 2020). Any T-unit that contained one or more grammatical or lexical errors was coded as an errorful T-unit. Lexical errors consisted of words that did not make sense within the given linguistic or semantic context. Grammatical errors consisted of morphological (e.g., subject-verb disagreement) or syntactic (e.g., inaccurate word order) inconsistencies that departed from standard written English. Spelling errors (as long as the meaning was clear) or mechanical errors were ignored; thus, if a T-unit contained these types of error(s) alone, it was coded as accurate. To calculate the overall ratio of error-free T-units, the number of error-free T-units was divided by the total number of T-units.

As a simple way to gauge L2-English learners' ability to produce a variety of vocabulary words, lexical diversity was measured by counting the sheer number of word types (i.e., unique, nonrepeated words) included in each email response. Following studies that utilized lexical diversity measures (e.g., Treffers-Daller et al., 2018), before counting the number of word types, spelling errors were corrected, so that different variations in spelling of the same word would not be counted as distinct word types. For example, if the word *presentation* was spelled correctly in one place and incorrectly (e.g., *pretentation*) in another, it was corrected to *presentation* so that it was considered as the same word type. In counting the total number of types, words that belong to the same lemma or word family but take on distinct morphological forms (e.g., go, goes, went, gone) were counted as separate types (Yu, 2010).

Syntactic fluency was defined as the total number of T-units produced in each response (e.g., Wigglesworth & Storch, 2009). This measure was thought to represent participants' ability to produce standalone syntactic structures (i.e., T-units) during pressured writing. Speed fluency, on the other hand, was operationalized as the number of words produced per second (e.g., Sasaki, 2000). The total number of words included in each participant's email response (including opening and closing for this measure) was counted, and it was divided by the time (in seconds) that each participant spent in writing their email response. The maximum time to respond to each task was 5 minutes (300 seconds).

Measures that required higher degrees of inference were dual coded by second coders. These measures included (a) identifications of T-units and (b) accuracy scoring. Two coders coded 50% of the data for these phenomena, and the simple agreement ratio for inter-rater reliability was found to be 96% for T-units and 92% for accuracy.

C-Test

The 40-item C-test was automatically scored using an exact-response approach for each blank. A participant was given one point for each blank where they were able to enter all of the correct missing letters. Each text included a total of 20 blanks, with 40 being the maximum score possible.

Perception Questionnaire

Participants' responses to the questions about (a) difficulty, (b) effort, (c) time pressure, (d) interest, and (e) authenticity were analyzed by calculating average ratings for each question for each group (i.e., Group A or B).

Statistical Analyses

In order to discern patterns between the conditions and measures in this data set, we focused on comparisons of mean values and 95% confidence intervals between the conditions on each measure (Norris, 2015). Prior to making these comparisons, we utilized inferential tests to examine whether condition, proficiency grouping, or measure effects were robust enough to detect differences beyond reasonable levels of error. Using SPSS v. 27, Bonferroni-adjusted factorial and one-way analysis of variance (ANOVA) calculations were conducted, with the conditions and proficiency groupings serving as the independent variables and the five participant perception ratings, or the four linguistic indices plus performance scores, serving as dependent variable measures. The overall alpha level was set at $p < .05$ prior to Bonferroni adjustment within each analysis. Note that multivariate analyses were not employed due to the highly disparate nature of the dependent variable measures, including linguistic indices on very distinct scales, performance ratings, and perception questionnaire Likert-scale ratings. Where appropriate, subsequent univariate analyses were conducted for each measure, followed by graphical and descriptive statistical comparisons between the conditions and proficiency groupings on each measure. In order to examine capacity of the tasks to discriminate among learners at different proficiency levels, Pearson correlation coefficients were also calculated to compare linguistic indices and performance scores with C-test scores.

Results

This section presents results of the study in the following order. First, the comparability of the two experimental groups is confirmed through examination of proficiency test scores. Next, the possible effects of simulated context on participants' subjective perceptions are analyzed based on Likert-scale ratings of several perception variables for the email writing task. Then, possible effects of simulated context, as well as proficiency level, on participants' email writing performances are examined through between-groups comparisons of several measured variables. Finally, the relationship between participants' proficiency and performance under the two context conditions is examined through correlations.

As described in the methodology, a key assumption for this study was that participants were equivalently distributed into the two experimental groups. To examine this assumption, average L2 English proficiency scores on the C-test were calculated for each group overall and by a priori proficiency level designation (low, mid, high). As shown in Table 3, the two groups exhibited virtually identical means and standard deviations for C-test scores overall, suggesting very similar distributions of proficiency in each. A univariate ANOVA, [$F[1,266] = 0.05$, $p = .826$]. provided additional support of this interpretation, showing no statistically significant difference in mean C-test scores between the two conditions.[1] At the a priori proficiency group level, minimal differences can be seen in the mean C-test scores for the two experimental conditions at each level, though these small differences are outweighed by the substantial and statistically significant differences between the low, mid, and high levels overall ($F[2,266] = 49.76$, $p = .001$). Given these observed patterns, assumptions of comparability between the two groups were deemed to have been met, and further analyses were undertaken.

Prior to examining potential effects on email writing performance, it was important to understand whether the addition of simulated contextual enhancements was registered by participants and in turn whether it affected participants' perceptions of performing the email writing task. Average Likert-scale ratings (1–7 points) for each group were calculated for five different perception variables (see Table 4). Group B participants, who were exposed to simulated context, expressed on average greater difficulty, effort, time pressure, and interest in comparison with participants in Group A, who were assigned to the nonenhanced context condition. Interestingly, time pressure showed the largest difference (approximately 3/4 of a scale point). Although very slightly higher authenticity was perceived by Group A, the ratings for authenticity were the highest of all the variables for both groups, suggesting that all participants found the task to be quite authentic. A

Table 3 Means (Standard Deviations) for C-Test Scores by Group

L2 Proficiency		
Level	Group A No context	Group B Context
Low	19.62 (7.08)	21.35 (7.51)
Mid	25.11 (7.96)	24.48 (9.04)
High	31.58 (5.72)	31.07 (6.34)
Total	26.29 (8.38)	26.21 (8.64)

Table 4 Participants' Perceptions of Performing the Email Writing Task

	Email writing task	
Perception	Group A no context	Group B context
Difficulty	3.23 (1.50)	**3.36 (1.60)**
Effort	3.82 (1.56)	**4.14 (1.73)**
Time pressure	3.60 (1.90)	**4.34 (1.92)**
Interest	4.78 (1.65)	**5.18 (1.66)**
Authenticity	**5.71 (1.48)**	5.66 (1.54)

Note: Bolded figures represent the higher mean ratings for each variable on each task.

series of univariate ANOVAs, with Bonferroni-adjusted *p* values, was conducted on the perception measures, with findings indicating statistically significant differences for time pressure (F[1,266] = 11.78, p = .001) and interest (F[1,266] = 4.00, p = .047), but not for the other measures: difficulty (F[1,266] = 0.59, p = .445), effort (F[1,266] = 2.37, p = .125), authenticity (F[1,266] = 0.00, p = .990).

The differences in participants' perceptions of performing the email writing tasks under enhanced simulated versus nonenhanced conditions suggest that the provision of simulated context was, at a minimum, registered in how participants subjectively experienced the task. In particular, the simulated context seemed to induce a sense of time pressure in completing the task, even though the available time was identical for both groups. It is also worth emphasizing that the simulated context group found the task somewhat more interesting. Whether such differences also impacted task performance was the focus of additional analyses.

In order to examine whether the addition of simulated contextual enhancements affected participants' email writing performances, outcomes on six performance measures were compared between the two groups on the email writing task. Table 5 displays descriptive statistics on the six measures for each proficiency level and in each context condition. An initial comparison of the two context conditions on total scores for each measure indicated no substantial mean differences except for Syntactic Complexity and Syntactic Fluency, and comparisons within each condition across the three proficiency levels showed generally consistent increases on most measures from low to high proficiency. A series of factorial ANOVAs was run to further investigate the effects of both task context and proficiency level on participant performance, with *p* values Bonferroni-adjusted due to repeated testing. In terms of the effect of context condition, statistically significant effects were found for Syntactic Complexity (F[1,266] = 4.48, p = .035) and for Syntactic Fluency (F[1,266] = 4.18, p = .042), but not for the other performance measures: Task Accomplishment (F[1,266] = 0.06, p = .807), Grammatical Accuracy (F(1,266) = 0.02, p = .883), Lexical Variety (F[1,266] = 0.45, p = .504), and Speed Fluency (F[1,266] = 2.56, p = .111). In terms of the effect of proficiency levels, statistically significant differences were found for Task Accomplishment (F[2,266] = 29.51, p = .000), Syntactic Complexity (F[2,266] = 8.46, p = .000), Grammatical Accuracy (F[2,266] = 3.52, p = .031), and Lexical Variety (F[2,266] = 15.69, p = .000). However, neither Speed Fluency (F[2,266] = 0.62, p = .538) nor Syntactic Fluency (F[2,266] = 1.17, p = .313) showed statistically significant differences by proficiency level.

Examining the Syntactic Complexity and Syntactic Fluency measures in more detail, several patterns of difference are noteworthy. First, Group B, which experienced the simulated context condition, produced writing that was noticeably *less* syntactically complex than Group A, on average .73 words per T-unit less (see Figure 2). In addition, although participants at all proficiency levels produced less complex writing in the simulated context condition, participants in the low proficiency level produced dramatically less complex writing than their counterparts in the nonenhanced context condition, with T-units that were on average 1.5 words shorter in length.

Second, turning to syntactic fluency (see Figure 3), the pattern of difference between the two groups was the opposite. Thus, Group B participants (who experienced the simulated context condition) produced noticeably *more* T-units than did Group A participants. Furthermore, although participants at all proficiency levels produced greater numbers of T-units in the simulated context condition, participants in the low proficiency level wrote considerably more T-units than did their counterparts in the nonenhanced context condition, on the order of greater than one T-unit more per email (1.08, 0.47, 0.14).

A final analysis examined the extent to which email writing performance under the two conditions was more or less related to participants' L2 English proficiency. Pearson correlations were calculated for each of the six performance

Table 5 Means (Standard Deviations) for Performance Measures on the Email Writing Task

Task	Proficiency level	Group A no context	Group B context
Task accomplishment (performance rating)	Low	2.66 (0.82)	2.82 (0.75)
	Mid	3.20 (0.81)	3.05 (1.00)
	High	3.63 (0.69)	3.69 (0.69)
	Total	3.23 (0.85)	3.24 (0.90)
Syntactic complexity (words per T-unit)	Low	10.30 (4.87)	8.79 (2.17)
	Mid	10.39 (2.64)	9.90 (3.02)
	High	11.55 (2.59)	11.17 (2.76)
	Total	10.83 (3.36)	10.10 (2.80)
Grammatical accuracy (% error-free T-units)	Low	0.48 (0.27)	0.50 (0.30)
	Mid	0.66 (1.13)	0.56 (0.28)
	High	0.68 (0.27)	0.73 (0.24)
	Total	0.62 (0.70)	0.61 (0.29)
Lexical variety (word types)	Low	37.12 (12.27)	41.22 (12.26)
	Mid	44.76 (11.35)	45.04 (14.60)
	High	50.60 (13.40)	49.35 (11.50)
	Total	45.14 (13.45)	45.70 (13.17)
Syntactic fluency (number of T-units)	Low	4.97 (2.32)	6.05 (2.57)
	Mid	5.74 (2.02)	6.21 (2.37)
	High	5.92 (2.03)	6.06 (2.16)
	Total	5.62 (2.12)	6.11 (2.33)
Speed fluency (words per second)	Low	0.27 (0.23)	0.35 (0.33)
	Mid	0.29 (0.15)	0.31 (0.15)
	High	0.32 (0.12)	0.33 (0.12)
	Total	0.30 (0.16)	0.33 (0.20)

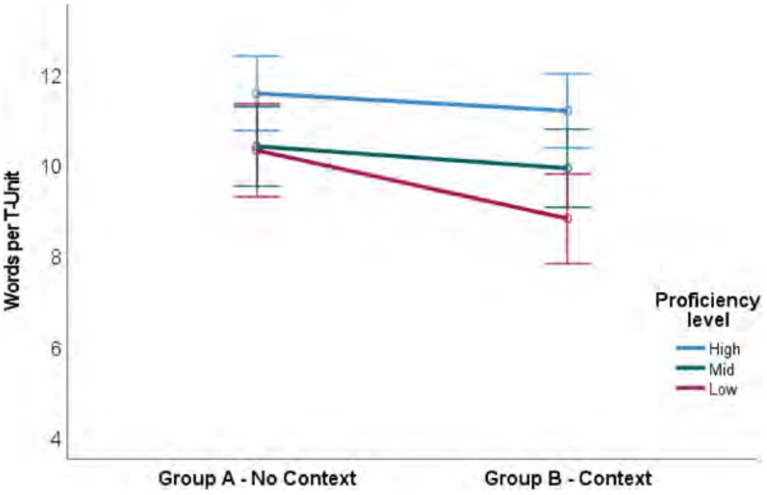


Figure 2 Syntactic complexity measures for two groups.

measures in relation to scores on the C-test criterion measure, and the results were compared between the two groups (see Table 6). Several patterns of relationship are noteworthy. For five of the six measures, correlations were highest in Group B, the simulated context condition, and for the sixth measure of Speed Fluency, correlations were nearly identical and extremely small. The strongest correlation was found between C-test scores and Task Accomplishment, suggesting that rated performances on the email writing task overall are meaningfully related to global language proficiency. The largest differences in strength of correlation—in favor of the simulated context condition—were found for measures of Grammatical Accuracy and Syntactic Complexity. On the whole, these patterns of relationship suggest that email writing performance under the simulated context condition is somewhat more closely associated with participants' language proficiency than performance in the nonenhanced context condition, although to greater or lesser degrees depending on the different measures.

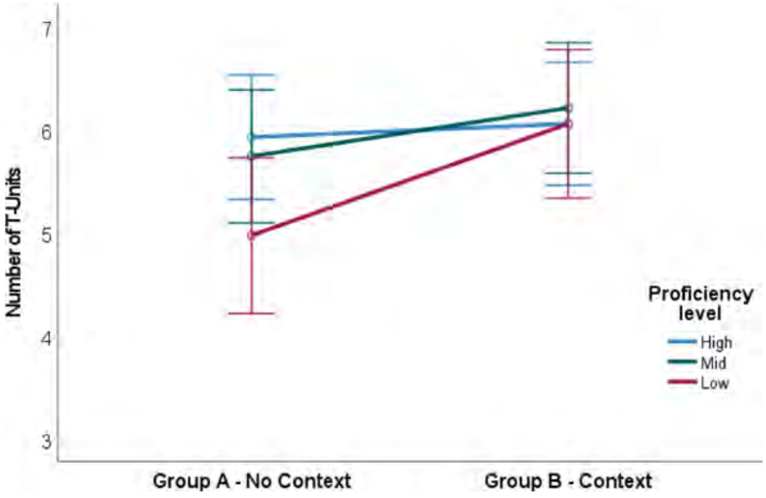


Figure 3 Syntactic fluency measures for two groups.

Table 6 Pearson Correlations Between L2 Proficiency and Task Performance Measures

	L2 Proficiency	
Performance measure	Group A no context	Group B context
Task accomplishment	0.59	0.60
Syntactic complexity	0.10	0.17
Grammatical accuracy	0.19	0.45
Lexical variety	0.44	0.47
Syntactic fluency	0.27	0.31
Speed fluency	0.07	0.06

Note: Bolded figures represent the higher correlations for each variable on each condition.

Discussion

Findings from the current study indicate that the addition of simulated context—in the form of extra instructions about the situation, as well as visual and audio enhancements during the task performance—has a salient impact on several dimensions of participants' perceptions about the task as well as aspects of their performance on the task. In addition, differences were observed in the relationship between linguistic and holistic measures of task performance and a criterion measure of English language proficiency, with slightly stronger relationships observed under the simulated context task condition. In the following, these findings are discussed from the perspective of TBLA and cognitive load theory.

Participant Perceptions

A key question for task-based assessments that seek to enhance test authenticity by simulating aspects of the communication context is whether test takers buy into and respond to the simulations. The current study revealed that participants clearly reacted to the simulated task context, although in different ways for distinct features. In general, it is noteworthy that participants in both conditions found the email writing task to be quite authentic to very similar degrees. This finding, and the lack of difference between conditions, may have been due to the orientation of the perception question, which focused on whether the task was similar to something the test taker would do in real life. Here, then, it may have been the case that writing an email, the core primary task, was deemed to be similar to daily life activities of most participants, regardless of the contextual features of the performance situation. That is, the authenticity of the task may have overridden or outweighed the relative situational authenticity of simulated context or the lack thereof.

Turning to indicators of cognitive load, although not statistically significant, participants perceived both the difficulty and mental effort required in the task to be higher for the simulated context condition. These perceptual measures are often used in research into differences in the cognitive load or complexity of tasks, but, as Sasayama (2016) has pointed out, they may only be sensitive to very substantial and obvious differences in task complexity (e.g., a task with only one character versus nine characters). It may be, then, that the degrees of difficulty and effort realized by the participants working under the simulated context condition were not particularly substantially greater, although perhaps noticeable to at least some extent. Here again, it is possible that the general challenge of writing the email—with its inherent cognitive load—superseded effects of simulated context on the extraneous cognitive load realized during performance.

Focusing on time pressure and interest, a more salient pattern of difference emerged in relation to the context condition of performance. Here, the findings suggest that additional contextual factors, such as extra information about the scenario, as well as the addition of visual and audio elements to replicate the real-world experience, effect an actual sense of urgency, anxiety, or stress, while generating a noticeably higher degree of engagement with the task. Some of the comments made by Group B participants—who experienced the task with simulated context—in explaining their high ratings on time pressure further underscore this interpretation:

- … the instruction which said "Dr. Smith is VERY STRICT" made me anxious and stressed. (U305149)
- The simulated train background noise increased my pressure. (N315237)

Group B participants similarly attributed their high ratings on perceived interest to the simulated real-world experience, as illustrated in these examples:

- It was so interesting because it tried to be like a real situation and the sound give the ambient to make real. (C209761)
- It was interesting for me to write as I actually thought that I'm on a subway and I'll miss my class. (N107171)
- Very cool, I felt that I was on a real train during the task. (N313219)

The various contextual enhancements, then, seemed to induce a genuinely interesting and "real" task performance experience, accompanied by a clear perception of enhanced pressure. These differential perceptions between the two conditions provide key sources of evidence that the contextual enhancements have an effect on the mental state of the participants—that is, at a minimum the designed differences were registered in the reactions of the participants to the task performance situation in the form of perceived pressure and interest, and potential effects of a smaller degree were found for measures of cognitive load. The patterns also reveal a potential tension in how participants responded to the simulated context enhancements. On the one hand, time (and other sorts of) pressure to perform would almost certainly be interpreted as a source of extraneous cognitive load and generally presumed to be detrimental to performance. On the other hand, interest in the task and performance situation would be interpreted as a source of germane cognitive load, helping learners to focus on performing the task at hand. Arguably, the combination of pressure and interest leads to specific performance outcomes, as we consider next.

Task Performance

Turning to the performance measures, linguistically, the simulated context positively affected syntactic fluency and negatively affected syntactic complexity, whereas accuracy, lexical variety, and speed fluency were not affected to a meaningful extent. Simulated context, which resulted in increased perceived time pressure, seems to have encouraged participants to produce more but shorter and less complex T-units quickly to get the required meaning across. This finding is in line with Skehan's (1996) prediction on the effect of increased time pressure (or communicative stress in general); tasks with increased time pressure tend to push learners to prioritize meaning conveyance over form and to focus on fluency rather than linguistic complexity or accuracy.

Interestingly, however, the increased sense of time pressure did not affect speed fluency in comparisons between the two groups. In other words, participants produced approximately the same number of words per second on average (0.30 words by Group A and 0.33 words by Group B), regardless of the context conditions. A similar pattern was observed in Johnson's (2017) meta-analysis of 20 studies that investigated the effect of task design manipulations (i.e., cognitive task complexity) on L2 learners' written performance. He found that cognitive task complexity generally has no discernible effects on writing fluency when it is measured by speed fluency (or repair fluency in a few studies that operationalized it).[2] Last, it is worth noting that the effects of simulated context on both syntactic fluency and syntactic complexity were exaggerated among the lower proficiency learners in comparison with the higher proficiency learners. For the lowest proficiency group, learners at the A1–A2 CEFR range produced in excess of one T-unit more under the simulated context condition, and their syntactic complexity was gauged to be on average 1.5 words shorter per T-unit. By comparison, learners at the B2 – C2 CEFR range exhibited much smaller differences between the two conditions. This finding supports the notion that aspects of simulated context may have heightened influence on learners who are still developing their L2 abilities, whereas learners who have achieved greater mastery of the target language are able to handle the communicative context more readily.

Turning to lexical variety and grammatical accuracy, it is worth highlighting that the core primary task of writing an email to the professor was the same for the two performance conditions, and this consistent basic requirement may explain why these linguistic phenomena were not particularly affected by the enhanced context condition. That is, the participants in both conditions needed to use whatever linguistic resources (vocabulary, grammar) they had at their disposal to attempt to get across the required meaning-making aspects of the task. Although the ways in which they put those resources together syntactically were affected by the simulation-induced pressure, leading to an apparent trade-off effect between syntactic fluency and complexity, the basic building blocks of vocabulary and grammar rules at their disposal and relevant to meeting the demands of the task were not. It is additionally of interest that the lexical and accuracy measures were not influenced by the apparent extraneous load exerted by the simulated context enhancements; that is, the effects were contained to distinct dimensions of syntactic performance.

Finally, looking at the holistic task accomplishment measure, participants who experienced the task without context (Group A) and with simulated context (Group B) were awarded essentially the same average task accomplishment scores (3.23 for Group A and 3.24 for Group B out of 5). In general, the email writing task was relatively easy for all participants, with even the low proficiency group receiving scores that indicated partially successful responses, regardless of the context conditions. When comparing participants' performances under the two conditions, in relation with the minimal differences in perceived task difficulty and mental effort exerted, it may be that the simulated context did not pose noticeably higher overall cognitive load and thus did not take substantial attentional resources away from completing the task (or from deploying vocabulary and grammatical resources to similar degrees in doing so; see above). As a result, the overall level of load was manageable for many of the participants in Group B and did not lead to deteriorated task performance when measured holistically in terms of accomplishment.

Discrimination and Learner Proficiency

An important attribute of any language test is its capacity to discriminate among learners at different levels of the ability being assessed. Of interest in the current study was the extent to which additional simulated context within a TBLA task would affect this discrimination capacity. Examining correlations with a holistic measure of written L2 English proficiency, in the form of C-test scores, most measures of participants' CALF performances showed slightly stronger relationships under the simulated context condition. Although holistic performance ratings correlated the strongest and very similarly with C-test scores for both conditions, measures of grammatical accuracy and syntactic complexity yielded the largest differences in strength of correlation between the two groups, in favor of the simulated context condition. It is clear from the performances of the three proficiency groupings within each condition that the simulated context condition spread learners out to a greater extent than did the nonenhanced context condition. This pattern is most apparent on the grammatical accuracy and syntactic complexity measures, and of course that dispersion of measurement values is necessary for higher correlations. By contrast, dispersion of measures for speed fluency, for example, was much more truncated between the proficiency levels, leading to much lower correlations with C-test scores under both conditions. On the syntactic complexity measure, however, the participants in the nonenhanced context condition scored similarly across the different proficiency levels, with a mean difference between the high and the low groups of 1.25 words per T-unit. On the other hand, the participants who experienced the task with the simulated context scored more differentially across the proficiency groups, on the order of 2.38 words per T-unit difference between the low and high groups. Similar patterns were observed for the grammatical accuracy measure. As noted above, the simulated context may have been acting as a realistic stressor on the developing language abilities of the lower proficiency learners, while allowing the higher proficiency learners to exhibit their abilities to handle realistic communication tasks. This differential effect of simulated context may point to a closer alignment between language performance on an authentic communication task, versus a stripped-down version of the same, and a global language proficiency construct.

Limitations and Directions for Future Research

Several factors limit generalizations based on the current study. First, it is an empirical question whether the findings of this study could be extended to other tasks or other types and degrees of simulated contextual factors. Here, only a handful of easily operationalized aspects of the context (representation of the interlocutor and his motivation, visual and audio depiction of the task performance location) were included with a single email writing task, whereas a number of other such simulations would be possible with a variety of other task types (e.g., of higher fidelity in particular, taking advantage of technological innovations, immersive environments, and the like). The field would benefit from future studies that investigate the actual effects of additional task context implemented in relation to different real-world task types (e.g., ordering food, making an appointment, giving a presentation) to better understand how extraneous cognitive load affects L2 learners' assessment task perceptions and performances. Second, although of sufficient size and variability in English L2 proficiency, the participant sample was one of convenience. Several idiosyncratic characteristics were not controlled for, such as first language, age, language learning experiences, and reason for participating, and these may have influenced how the participants engaged in the various research activities and resulting patterns. Third, individual participants completed the entire experiment in an unsupervised, self-access format. Although efforts were made to ensure engagement and completion of all steps in the study (e.g., instructions at the beginning of the study, checking of completion), there was no way to control participants' actual efforts to pay attention, try their best, or provide honest, accurate answers to perception questionnaires. Finally, the approaches to measuring both participant perceptions as well as performances were selective and arguably represented incomplete depictions of the phenomena under investigation. Thus, the perception questionnaire items might have been augmented usefully with other ways of inquiring into participant responses to the task and simulated conditions (e.g., retrospective interviews), which might have allowed for much deeper insights into how and why they were influenced by different factors or not. Similarly, a host of other measures of linguistic performances might have been included to get at a more multifaceted and comprehensive picture of how learners deploy their second-language resources. These emendations would be beneficially revisited in future research to cast light on a more complete and in-depth understanding of the relationship between task design, including the simulation of context, and L2 task perception and performance.

Conclusion

This study investigated the potential effects of introducing low-threshold simulations of a few real-world dimensions of communicative task context into an email writing assessment task for L2 learners of English. Findings indicated that the simulated context enhancements registered with participants, especially by introducing a sense of time pressure to complete the task and by generating a heightened degree of interest in performing the task. These perceptual responses, however, did not lead to differences in performance on holistic task accomplishment ratings, nor on measures of lexical complexity, grammatical accuracy, or speed fluency. The simulated context enhancements were associated with clear differences in two dimensions of the syntactic production of learners' writing: greater syntactic fluency coupled with diminished syntactic complexity. Learners in the simulated context condition wrote noticeably more syntactic units, though of an overall less complex nature, arguably in response to the perceived time pressure exerted by the simulation. These effects were also observed to be considerably exaggerated for the lower proficiency learners in comparison to more advanced learners of English. Furthermore, measures of linguistic performance under the simulated context condition correlated consistently, though slightly, higher with a written measure of global L2 proficiency (a C-test), suggesting a higher degree of proficiency-related discrimination in association with the addition of simulated context.

From a TBLA perspective, the addition of a few simulated contextual features may have resulted in a more realistic task performance experience for the language learner participants. Although the task of writing an email to a professor was, overall, deemed highly authentic by both groups of participants, those exposed to simulated context features generally seemed to accept and respond to the real-world situation, and it generated heightened interest as well as specific performance effects. Of particular importance for an assessment, the simulated context version of the task proved to be more discriminating between learners at distinct proficiency levels, suggesting that adding aspects of the real world—including those that might be associated with extraneous cognitive load—may in fact contribute to improved measurement performance for these kinds of task-specific language tests. Future research into the simulation of context with the goal of enhancing language test authenticity, as well as expanding construct coverage, would do well to investigate similar phenomena across a variety of L2 task types and communication situations.

Acknowledgments

We would like to acknowledge the important contributions made to this study by the site coordinators and participating English learners, as well as by our ETS colleagues Michael Ecker, Jeremy Lee, Jung Aa Moon, Renka Ohta, and Michael Suhan.

Notes

[1] Null hypothesis significance tests are not designed to prove a lack of difference (Norris, 2015); however, their use can provide additional information regarding the degree of certainty in observations related to mean difference. Given the sizeable samples in the current study, the significance test here serves to support (not prove) the interpretation of no meaningful difference between the two groups in terms of their English proficiency scores.

[2] Note that three comparisons made in the meta-analysis between the simple [+Here-and-Now] condition and the complex [−Here-and-Now] condition yielded an average effect size of 0.44 for fluency in favor of the simple condition.

References

Anderson, N. J. (2015). Academic reading expectations and challenges. In N. W. Evans, N. J. Anderson, & W. G. Eggington (Eds.), *ESL readers and writers in higher education: Understanding challenges, providing support* (pp. 95–109). Routledge.

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135–146). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.009

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press. https://doi.org/10.7916/salt.v10i2.1430

Blyth, C. (2018). Immersive technologies and language learning. *Foreign Language Annals*, *51*(1), 225–232. https://doi.org/10.1111/flan.12327

Byrnes, H., Maxim, H., & Norris, J. M. (2010). Realizing advanced FL writing development in collegiate education: Curricular design, pedagogy, assessment [monograph]. *Modern Language Journal*, *94* [Issue supplement].

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research*, *90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Cumming, A., Cho, Y., Burstein, J., Everson, P., & Kantor R. (2021). Assessing academic writing. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 107–151). Routledge. https://doi.org/10.4324/9781351142403-4

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732911

Douglas, D. (2012). ESP and assessment. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 367–383). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118339855.ch19

Haider, I. (2019). Cyberpragmatics: Assessing interlanguage pragmatics through interactive email communication. In S. Papageorgiou & K. M. Bailey (Eds.), *Global perspectives on language assessment* (pp. 152–168). Routledge. https://doi.org/10.4324/9780429437922-11

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32

Howell, H., & Mikeska, J. N. (2021) Approximations of practice as a framework for understanding authenticity in simulations of teaching. *Journal of Research on Technology in Education*, *53*(1), 8–20. https://doi.org/10.1080/15391523.2020.1809033

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. University of Hawai‘i Press.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. University of Hawai‘i Press.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*, *35*(1), 1–67. https://doi.org/10.2307/1165818

Hyland, K. (2019). *Second language writing*. Cambridge University Press.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–285). Penguin Books.

Isbell, D., & Winke, P. (2019). ACTFL Oral proficiency interview–computer (OPIc). *Language Testing*, *36*(3), 467–477. https://doi.org/10.1177/0265532219828253

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, *18*(3), 213–241. https://doi.org/10.1016/S0889-4906(97)00053-7

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, *37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*(4), 351–371. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Knoch, U., & Macqueen, S. (2019). *Assessing English for professional purposes*. Routledge.

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, *17*(1), 43–64. https://doi.org/10.1177/026553220001700102

Li, P., & Lan, Y. J. (2022). Digital language learning (DLL): Insights from behavior, cognition, and the brain. *Bilingualism: Language and Cognition*, *25*(3), 361–378. https://doi.org/10.1017/S1366728921000353

McTighe, J., & Willis, J. (2019). *Upgrade your teaching: Understanding by design meets neuroscience*. ASCD.

Morrow, K. (1978). *Techniques for evaluation for a notional syllabus*. Royal Society of Arts.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit, & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–157). Oxford University Press.

Norris, J. M. (2009). Task-based teaching and testing. In M. Long and C. Doughty (Eds.), *The handbook of language teaching* (pp. 578–594). Blackwell. https://doi.org/10.1002/9781444315783.ch30

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and some solutions. In J. M. Norris, S. Ross, & R. Schoonen (Eds.), *Improving and extending quantitative reasoning in second language research* (pp. 95–124). Wiley-Blackwell.

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244. https://doi.org/10.1017/S0267190516000027

Norris, J. M. (2018a). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7–33). Peter Lang.

Norris, J. M. (2018b). Task-based language assessment: Aligning designs with intended uses and consequences. *JLTA Journal*, *21*, 3–20. https://doi.org/10.20622/jltajournal.21.0_3

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessment*. University of Hawai‘i Press.

Norris, J. M., Davis, J. McE., & Xi, X. (2021). Framing the assessment of academic English for admissions decisions. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 1–21). Routledge. https://doi.org/10.4324/9781351142403-1

Norris, J. M., & East, M. (2021). Task-based language assessment. In M. J. Ahmadian & M. Long (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 507–528). Cambridge University Press. https://doi.org/10.1017/9781108868327.029

Norris, J. M., & Manchón, R. M. (2012). Investigating L2 writing development from multiple perspectives: Issues in theory and research. In R. M. Manchón (Ed.), *L2 writing development: Multiple perspectives* (pp. 221–244). deGruyter. https://doi.org/10.1515/9781934078303

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity* (Vol. *50*). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.50

Oliveri, M. E., Mislevy, R. J., & Slomp, D. H. (2021). Principled development of workplace English communication part 1: A sociocognitive framework. *The Journal of Writing Analytics*, *5*, 34–70. https://doi.org/10.37514/JWA-J.2021.5.1.02

Paas, F., van Merrienboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79*(1), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

Papageorgiou, S., Schmidgall, J., Harding, L., Nissan, S., & French, R. (2021). Assessing academic listening. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 61–106). Routledge. https://doi.org/10.4324/9781351142403-3

Park, M. (2018). Innovative assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. *ReCALL*, *30*(2), 196–213. https://doi.org/10.1017/S0958344017000362

Phillips-Wren, G., & Adya, M. (2020). Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems*, *29* (sup1), 213–225. https://doi.org/10.1080/12460125.2020.1768680

Purpura, J. E. (2014). Cognition and language assessment. In A. J. Kunan (Ed.), *The companion to language assessment* (pp. 1452–1476). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla150

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, *35*(1), 87–92. https://doi.org/10.1093/applin/amt039

Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. John Benjamins Publishing Company. https://doi.org/10.1075/tblt.2

Robinson, P., & Ross, S. (1996). The development of task-based assessment for academic purposes programs. *Applied Linguistics*, *17*(4), 455–476 https://doi.org/10.1093/applin/17.4.455.

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, *9*(3), 259–291. https://doi.org/10.1016/S1060-3743(00)00028-X

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, *100*(1), 231–254. https://doi.org/10.1111/modl.12313

Sasayama, S., & Norris, J. M. (2019). Unravelling cognitive task complexity: Learning from learners' perspectives on task characteristics and L2 performance. In Z. Wen & M. J. Ahmadian (Eds.), *Researching L2 task performance and pedagogy: In honor of Peter Skehan* (pp. 95–132). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.13.06sas

Sasayama, S., & Norris, J. M. (2023). Designing speaking tasks for different assessment goals: The complex relationship between cognitive task complexity, language performance, and task accomplishment. *TASK: Journal on Task-Based Language Teaching and Learning*, *2*(2), 184–217.

Savignon, S. J. (1985). Evaluation of communicative competence: The ACTFL provisional proficiency guidelines. *The Modern Language Journal*, *69*(2), 129–134. https://doi.org/10.1111/j.1540-4781.1985.tb01928.x

Schedl, M., O'Reilly, T., Grabe, W., & Schoonen, R. (2021). Assessing academic reading. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 22–60). Routledge. https://doi.org/10.4324/9781351142403-2

Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for further research. *Educational Psychology Review*, *13*, 23–52. https://doi.org/10.1023/A:1009004801455

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38–62. https://doi.org/10.1093/applin/17.1.38

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press. https://doi.org/10.5070/L4111005027, *11*

Skehan, P. (Ed.). (2014). *Processing perspectives on task performance*. John Benjamins Publishing Company. https://doi.org/10.1075/tblt.5

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, *76*(2), 129–141. https://doi.org/10.1111/j.1540-4781.1992.tb01093.x

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*, 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sydorenko, T., Smits, T. F., Evanini, K., & Ramanarayanan, V. (2019). Simulated speaking environments for language learning: Insights from three cases. *Computer Assisted Language Learning*, *32*(1–2), 17–48. https://doi.org/10.1080/09588221.2018.1466811

Tai, T. Y., Chen, H. H.-J., & Todd, G. (2022). The impact of a virtual reality app on adolescent EFL learners' vocabulary learning. *Computer Assisted Language Learning*, *35*(4), 892–917. https://doi.org/10.1080/09588221.2020.1752735

The CELBAN Centre. (2018). *History of CELBAN*. https://www.celbancentre.ca/about-us/reports-for-test-users/history-of-CELBAN.aspx

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Wang, C.-P., Lan, Y.-J., Tseng, W.-T., Lin, Y.-T. R., & Gupta, K. C.-L. (2020). On the effects of 3D virtual worlds in language learning: A meta-analysis. *Computer Assisted Language Learning*, *33*(8), 891–915. https://doi.org/10.1080/09588221.2019.1598444

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ASCD.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing*, *26*(3), 445–466. https://doi.org/10.1177/0265532209104670

Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing*, *18*(2), 187–206. 10.1177%2F026553220101800205

Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing academic speaking. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 152–199). Routledge. https://doi.org/10.4324/9781351142403-5

Youn, S. J. (2014). Measuring syntactic complexity in L2 pragmatic production: Investigating relationships among pragmatics, grammar, and proficiency. *System*, *42*, 270–287. https://doi.org/10.1016/j.system.2013.12.008

Youn, S. J. (2020). Managing proposal sequences in role-play assessment: Validity evidence of interactional competence across levels. *Language Testing*, *37*(1), 76–106. https://doi.org/10.1177/0265532219860077.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Appendix

Table A1 Email Task Scenarios for the Train Task

	Train Task
Scenario description	Today, you are scheduled to give an important presentation in your English class taught by Dr. Smith. However, your train is late, and you will not arrive on time. You may miss the first half of your class. Write an email to your professor in English to: (a) apologize; (b) explain what happened; and (c) propose a solution. You have 5 minutes to write your email.

Table A2 Task Accomplishment Scoring Rubrics

5 *A fully successful response*

The writer's email message is effective, clearly expressed, and pragmatically appropriate from the beginning to the end.

Language: The response displays

- Consistent facility in the use of language
- Effective syntactic variety
- Precise and idiomatic word choice
- Consistent use of appropriate pragmatics (e.g., hedging, use of modals, other politeness markers)
- No or very few lexical or grammatical errors

Content/writing conventions: The response includes an apology, an explanation of the situation, and a proposed solution that are effective for the given scenario. The message follows the typical writing conventions of this genre, and it includes (a) an appropriate opening and closing and (b) an appropriate professional title when addressing the professor (e.g., Dr.).

4 *A successful response*

The writer's email message is mostly effective and easily understood.

Language: The response displays

- Facility in the use of language
- Some syntactic variety
- Appropriate word choice
- General use of appropriate pragmatics (e.g., hedging, use of modals, other politeness markers)
- Few lexical or grammatical errors

Content/writing conventions: The response includes an apology, an explanation of the situation, and a proposed solution that are mostly effective for the given scenario. The message follows the typical writing conventions of this genre most of the time, but it may lack (a) an appropriate opening or closing and/or (b) an appropriate professional title when addressing the professor (e.g., Dr.).

Table A2 Continued

3	*A partially successful response* The writer's email message is partially effective. The scenario described in the message is understandable without the knowledge of the prompt. *Language*: The response displays • Some facility in the use of language • Some syntactic variety • A moderate range of vocabulary • Some noticeable errors in structure, word form, use of idiomatic language, and/or pragmatics *Content*: The response describes the basic scenario although some details may be omitted. The response may leave out an apology or a proposed solution, OR lack of language facility prevents parts of the message from being effective.
2	*A mostly unsuccessful response* The response is an attempt to address the task, but it is mostly unsuccessful. What's included in the message may be limited OR difficult to interpret. *Language*: The response displays • Some connected, sentence-level language • An accumulation of errors in sentence structure and/or language use • Limited range of vocabulary and syntax *Content*: The response is relevant to the given scenario. It contains some descriptions of the scenario, but it is only minimally elaborated, or difficult to interpret especially if you do not know the prompt.
1	*An unsuccessful response* The response is an unsuccessful attempt to write an email message about the given scenario. Language: • The response is telegraphic. • The response displays serious and frequent errors in the use of language. • Limited range of vocabulary and syntax *Content*: The message is limited to the point of being unintelligible, or it is off topic.
0	*An authentic response is not attempted*: Response is too short to judge its topic relevance (e.g., "dear dr smith this email is for"), rejects the task itself (e.g., "I do not know"), is not in English, is entirely copied from the prompt (e.g., "Dr. Brown Dr. Brown Dr. Brown"), or consists of arbitrary keystrokes.

Note: A blank response should be scored as NS (nonscoreable) rather than a score 0.

Suggested citation:

Norris, J. M., Sasayama, S., & Kim, M. (2023). *Simulating real-world context in an email writing task: Implications for task-based language assessment* (Research Report No. RR-23-05). ETS. https://doi.org/10.1002/ets2.12366

Action Editor: John Davis

Reviewers: Saerhim Oh and Veronika Timpe-Laughlin

Find other ETS-published reports by searching the ETS ReSEARCHER database.