RESEARCH

Line, please? An analysis of the rehearsed speech characteristics of native Korean speakers on the English Oral Proficiency Interview—Computer (OPIc)

Gwyneth Gates[1,2], Troy L. Cox[1*] , Teresa Reber Bell[1] and William Eggington[1]

* Correspondence: troyc@byu.edu
[1]Brigham Young University, Provo,
UT 84602, USA
Full list of author information is
available at the end of the article

Abstract

Two assumptions of speaking proficiency tests are that the speech produced is spontaneous and the the scores on those tests predict what examinees can do in real-world communicative situations. Therefore, when examinees memorize scripts for their oral responses, the validity of the score interpretation is threatened. While the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines identify rehearsed content as a major hindrance to interviewees being rated above Novice High, many examinees still prepare for speaking tests by memorizing and rehearsing scripts hoping these "performances" are awarded higher scores. To investigate this phenomenon, researchers screened 300 previously rated Oral Proficiency Interview-computer (OPIc) tests and found 39 examinees who had at least one response that had been tagged as rehearsed. Each examinee's responses were then transcribed, and the spontaneous and rehearsed tasks were compared. Temporal fluency articulation rates differed significantly between the spontaneous and rehearsed segments; however, the strongest evidence of memorization lay in the transcriptions and the patterns that emerged within and across interviews. Test developers, therefore, need to be vigilant in creating scoring guidelines for rehearsed content.

Keywords: Rehearsed speech, Spontaneous speech, Temporal fluency, Fluency features, Speaking assessment, OPIc

Introduction

Globalization has led to an increased demand for English skills. Some companies, even those founded and based in countries where English is not an official or common language, are enforcing language policies to make themselves more competitive and better communicators in a global market through requiring English tests. One assumption of speaking proficiency tests is that the speech produced is spontaneous so that the interpretation of the scores on those tests will be predictive of what examinees can do in these real-world communicative situations. Therefore, when examinees memorize scripts for their oral responses, validity is threatened. For tests to be valid, responses need to be consistently rated,

and score interpretations need to be both meaningful and impartial (Kelly et al. 2017); however, when raters encounter memorized responses intermingled with the spontaneous, scoring can be impacted. Thus, the prevalence of rehearsed responses in speaking tests should be evaluated to determine the effect it might have on test validity.

In recent years, South Korean corporations such as Samsung, LG, Hyundai, Kia, and Daewoo have begun to require their employees to take the Oral Proficiency Exam-computerized (OPIc), a computerized adaptation of the ACTFL (American Council on the Teaching of Foreign Languages) Oral Proficiency Interview (OPI) to determine employees' levels of proficiency in English (Jeon 2010). The OPIc holds the same objective as the OPI of holistically assessing a speaker's proficiency through a series of oral prompts designed to elicit certain types of speech. In justifying the OPIc, one Korean executive explained that "the English ability required by the Samsung Group is a person who can speak their own thoughts as they wish" (Myung 2012). Similarly, ACTFL defines proficiency as the "ability to use a language to communicate meaningful information in a spontaneous interaction, and in a manner acceptable and appropriate to native speakers of the language" (ACTFL 2012c, p. 4).

However, it appears that examinees still attempt to "cram" for the interview by memorizing scripts and, as actors reciting lines, search for what they prepared for (Cox 2017). Raters in his study indicated the presence of rehearsed material: within the sample of 300 Korean OPIc exams, 20% of rater comments included notes of rehearsed material. Although Cox's study is not ongoing, memorization on the OPIc has proven pervasive enough that ACTFL trains raters on how to recognize and detect rehearsed speech (for example, each online rating includes a checkbox wherein the rater can indicate rehearsed material). Memorized utterances are mentioned explicitly in the ACTFL guidelines as an indicator of Novice speech, which demonstrates the significance of this problem from a proficiency rating standpoint. Understanding the nature of this issue lies in identifying not only what speaking proficiency is but also how educational traditions and practices differ across cultures.

Literature review

Rote learning and language proficiency

Rote learning is defined as the memorization and recall of material when the time comes for test or performance. Memorization can be a helpful strategy for mastering L2 vocabulary and formulaic sequences (Boers et al. 2006; van Hell and Mahn 1997), improving speaking (Takeuchi 2003), and performing well on tests (Kember 2000); however, the usefulness of relying heavily upon rote learning to attain second language proficiency is questioned (Wray and Fitzpatrick 2008). Bloom (1956) similarly argues in his Taxonomy of Educational Objectives that critical thinking and creating with learned material are better demonstrators of higher-level processing than recall alone: "Knowledge is of little value if it cannot be utilized in new situations or in a form very different from that in which it was originally encountered" (p. 29). In that same vein, ACTFL proficiency standards are designed so that to move beyond the Novice level, the interviewee must prove the ability to spontaneously create with the language.

While Novice-level speakers primarily use "isolated words and phrases that have been encountered, memorized, and recalled," a key characteristic of intermediate speakers is their ability to "create with the language when talking about familiar topics. . . [and] recombine learned material in order to express personal meaning" (ACTFL 2012a, p. 7).

In essence, intermediate speakers can use the words and phrases they know to spontaneously create sentence length utterances they have never heard or studied before. Because of this, raters cannot use any rehearsed material observed during the interview as evidence of intermediate-level speech. By ACTFL's standards, reliance upon rehearsed material, even if that material is applied to various tasks, separates language *performance* from *proficiency* (ACTFL 2012c). *Performance* describes language that is practiced in designated contexts while *proficiency* refers to spontaneous language used in conditions that are unanticipated (Rubio and Hacking 2019). The Center for Applied Linguistics (CAL) validation framework notes that a test's validity can be threatened when the interpretation of the test score is impacted (Kelly et al. 2017), and this could be by a raters' failure to detect or appropriately score scripted speech samples.

Influences within Asian education systems

Asian countries have a learning tradition and testing culture that, some have posited, traces its roots to Confucianism (Cooke and Kim, 2017; Seth 2002)—a term coined in reference to the Chinese scholar Confucius's lifelong promulgation of virtue, morality, and education (Yao, 2000). Confucianism, "both a tradition of literature and a way of life" (Yao, 2000, p. 11), rose to Chinese state orthodoxy by the second century BCE. Though it is no longer the official state dogma of any country today, its influence has permeated Asian beliefs, societal expectations, government, and educational practices (Shin, 2012; Yao, 2000).

The influence of testing started in China in the eighth century, when entry into civil service transitioned from a system based on personal connections to a meritocracy that prized mental acumen, which was measured primarily through the ability to memorize. The Korean government, which adopted these Confucian practices after the colonization in the north by the Chinese in approximately the start of the Common Era, established its imperial civil service examination (the *gwangeo*) in the eighth century and administered the test intermittently until the conclusion of the Joseon dynasty in the 1890s (Cooke and Kim, 2017; Shin, 2012).

Standardized tests became the trusted method to assess qualifying contenders for government positions, and test preparation became increasingly significant as not only an employment tool but an obscurant of traditional class lines (Deuchler 1992; Seth 2002). Deuchler explained the significance of achievement according to Confucianism in Deuchler 1992:

> Confucianism was then and later intimately connected with the search for rational standards that would weaken, if not sever, the indigenous link between the prerogatives of birth and political participation and would condition advancement to high office on achievement (p. 15).

In his use of ancient texts to justify his teachings, Confucius modeled a tendency toward selecting well-established solutions, a practice in which "free thought and individual expression were discouraged, giving way to the safer and surer domain of classical quotation" (Becker 1986, p. 77). Likely resulting from the combination of Confucian emphasis on reference to authority and testing culture, rote learning and production became a practical and encouraged method of learning throughout Asia (Becker 1986; Cooke and Kim, 2017; Keong-kyu, 2000).

Increased demand for English proficiency assessment in South Korea

South Korea is a notable example of globalization's impact on the use of English in the international workplace. In the wake of the Asian financial crisis of 1997, Korean companies became more interested in assessing applicants' flexibility and adaptability within a global workplace—particularly their ability to communicate in that international environment (Park 2011). In a series of interviews, Jambor found that 17% of South Korean job interviews were conducted 50–80% in English, while 6% of interviews were carried out entirely in English (Jambor 2011).

However, as Bae Jae-keun, CEO of the e-learning company Credu, explained, "Although many companies conduct English interviews, it is not easy to evaluate a lot of candidates at one time, and interviewers may not give unbiased, objective evaluations" (Whan-yung 2009). For several years, South Korean employers used the Test of English for International Communication (TOEIC) as a standard of English proficiency. Compared to the traditional English tests used in years past, the TOEIC (which focuses on business English) was a far more practical option for both companies and their employees, and the number of test takers soared in the following years (Park 2011).

Until 2006, the TOEIC featured no speaking or writing sections, and all questions were presented in multiple-choice format. The International Communication Foundation, which administered the TOEIC in South Korea, justified the recycling of past questions for the sake of maintaining anchor questions. The practice was discontinued in 2003, but not before one teacher "was able to tell his students that if the term 'South Korea' is found in a question, the correct answer is 'strategic'" (Jae-gon 2005). Messick (1982) calls such a problem "test-taking artifice;" in other words, it is among the "stratagems and answer-selection tricks resulting in increased test scores that are inaccurately high assessments of ability" (p. 67). By 2005, many corporations had done away with the TOEIC requirement entirely (Jae-gon 2005), while others demoted its make-or-break status to simply one more criterion of many within the job application process (Park 2011). The TOEIC may have been a more standardized, objective test of workplace English than past assessments, but there has been a gradual decline in popularity among companies.

The OPI and OPIc

The OPI is a global assessment in which speakers are rated holistically, receiving a score according to the ACTFL Proficiency Guidelines. Interviewees participate in a conversation with a trained rater who asks a series of questions and analyzes participant responses to determine their speaking level. The interview is subsequently rated by a second trained evaluator, after which a rating is determined. This proficiency level could be any one of the 10 rated proficiencies designated by ACTFL: low, mid, or high at the novice, intermediate, or advanced levels, or superior. (ACTFL 2012a).

Increased demand for the OPI has led to the development of the OPIc. Delivered via computer avatar, the OPIc allows for proctored testing anywhere with internet access. After completing a brief survey of their self-assessed language level and personal interests, interviewees are assigned one of five OPIc test forms. The avatar gives prompts relating to the interviewee's interests, and the interviewee's responses are recorded and double rated (ACTFL 2012b).

The fundamental difference between the OPI and the OPIc is that the first is an interview format while the second does not allow for participants to communicate with raters in real time. Although conducting the interview via avatar increases test practicality and convenience, Thompson et al. (2016) found that the majority of participants preferred the OPI, despite 31.8% of them scoring higher on the OPIc. Several participants cited the OPIc's lack of personal interaction—and, consequently, feedback and ease of conversation—as a primary reason they preferred the OPI.

Since many companies have "switched their preferred mode of assessment to methods that can directly observe the candidate's oral language skills" (Park 2011, p. 449), oral proficiency tests such as the OPI and the OPIc have begun to eclipse the TOEIC in South Korea. The trend toward the OPIc is evidenced by YouTube OPIc prep videos (Hackers Education Group 2014; Lee 2013), courses designed to help attendees reach their ideal OPIc score and job postings specifically for OPIc teachers (HiEnglish Korea 2015).

Previous studies involving the OPIc and Korean learners of English

In addition to the resources mentioned in the previous section, several articles investigate English test preparation and strategy specifically for Koreans taking the English OPIc. Park and Lee (2019) conclude that higher-scoring OPIc interviewees speak more quickly and pause for less time. Ko recommends "individual one-way speaking practice about daily life and past experiences" Ko (2010), concludes that language control and a large amount of speaking content "seems to guarantee [participants'] level-up," and recommends 12 storytelling strategies Ko (2017). While some findings caution against sacrificing true language development and narrative identity for memorization and test preparation (Kim 2016), related research appears to more often recommend than advise against such performance strategies, however unfit they may be in a holistic proficiency assessment.

In a study of Korean OPIc examinees, Shin et al. 2010 found that only small percentages of participants agreed or strongly agreed that they understood the ACTFL criteria of global tasks and functions, context/content, accuracy, and text type. Twice as many or more of the participants disagreed or strongly disagreed that they understood these crucial proficiency criteria despite nearly half of the participants (43.9%) agreeing that OPIc speech samples are reviewed and evaluated by certified raters. An additional noteworthy finding was the participants' agreement that OPIc examinees should "memorize vocabulary and common phrases before the test" (63.4%), "practice oral presentation and debate" (78.1%), and "practice visual and oral description and story narration" (84.2%). Although 75.6% of participants also agreed that conversation practice with a native speaker is an important practice method, the authors noted the heavy reliance upon the three previous strategies combined with a general lack of understanding the aims of the proficiency interview:

> Despite this lack of knowledge, [participants] believed that memorizing language forms is useful even in preparation for speaking tests. The primary goal of OPIc, however, is to assess oral proficiency for authentic language use. Hence there is a major gap between test preparation and the intent of the test itself. (p. 284).

Seo and Chang (2013) conducted a case study to determine the effectiveness of a recurring recording assignment in a month-long OPIc test prep course. Participants, all

employed by major Korean companies, took an OPIc pre-test before beginning the study. During class, they practiced model OPIc responses, useful phrases for the given topic, and direct translation of sentences from Korean to English; they then recorded themselves answering a sample OPIc question after each class period. Post-test OPIc scores increased for only three of the eight participants and for none of those who had not completed the assignment. The authors suggested hinderances of OPIc participants as a topic of future research and recommended voice recording assignments for others preparing for the OPIc. While Xie (2013) and Messick (1982) concede that drilling and other rote memorization practices *can* yield higher test scores, the assessment is inflated and its validity is compromised. Such an issue only underscores the pertinence of the present study and the need for additional research in rehearsed speech and oral proficiency assessment.

The research of Cox (2017), however, suggests that strategies like those previously suggested may not achieve desired results in a speaking proficiency assessment. In a study funded by Credu (a Samsung affiliate) that aimed to identify the strengths and weaknesses of Intermediate-level English learners based upon linguistic characteristics of particular tasks, Cox analyzed recordings of 300 Korean OPIc examinees, seeking to determine which linguistic features of the intermediate subgroups' (low, mid, and high) sampled responses prevented the examinees from achieving the next highest sublevel. One surprising observation was that OPIc raters noted "rehearsed material or canned/ memorized responses" (Cox 2017, p. 17) in approximately 20% of the intermediate low (IL) recordings and 12% of the intermediate mid (IM) recordings. These detected occurrences of rote speech were noted qualitatively in raters' comments, but in several cases the presence of rehearsed material resulted in a holistic rating of *Does Not Meet Expectations* for that segment. This is particularly problematic with the OPIc because while OPI raters can redirect, clarify, and probe in real time, OPIc raters "must listen for telltale signs of rehearsed responses and then exclude that sample as evidence that the examinee can create with the language" (Cox 2017, p. 104).

According to the ACTFL proficiency guidelines, a novice-level English speaker "can communicate minimally with formulaic and rote utterance, lists, and phrases" (2012a, p. 4), so memorized material offers evidence of only Novice-level speech. An examinee who has the ability to create with language at the intermediate level but reverts to only memorized scripts is underrepresenting their true language ability; therefore, Cox (2017) identifies rehearsed content as yet another factor which hinders OPIc interviewees from being rated higher.

Previous studies have examined the validity and the reliability of the OPIc (Surface et al. 2008), comparability of human-versus computer-delivered oral interviews (Surface et al. (2009), strengths, and weaknesses of the OPIc (Isbell and Winke 2019), engagement authenticity in speaking tasks (Lam 2015), and speech authenticity as perceived by raters (Burton 2020). Suggestions for further research have included investigation of individual speaker differences and their impacts upon an OPIc score and OPIc rater effectiveness (Surface et al. 2008), response analysis for better understanding of what proficiency features are and are not best elicited in the OPIc (Isbell and Winke 2019), examination of additional speech assessment samples for features of inauthenticity (Luk 2010), and how certain characteristics of speech are perceived by raters to be rehearsed (Burton 2020). OPIc raters, who may be unsure of how to proceed objectively

when presented with what sounds like rehearsed material, stand to benefit from these calls for research, particularly the latter three, in a closer examination of rehearsed speech characteristics. While scoring segments, raters in the Cox (2017) study in most cases did not specify why they thought the material had been memorized. The research questions of this study thus revolve around *how* these raters decided the responses had been rehearsed beforehand; in other words, identifying the characteristics which distinguish the segments marked as memorized from those which were not.

Speed, measured by articulation rate, and breakdown, measured by the number of silent pauses and the mean length of utterances, were selected as the variables of interest for the temporal fluency measures because previous research indicated that these accounted for much of the variance in speaking test scores (Ginther et al. 2010). To measure rate of speech, articulation rate was selected because it is "a pure measure of speed" (De Jong 2016, p. 212). Articulation is defined as the total number of syllables divided by *phonation* time. The number of silent pauses and the mean length of utterances (MLU) were selected as measures of speaker breakdown. These measures are correlated, being the inverse of one another, and were chosen as indicators of breakdown within speech because as a speaker's sentences become shorter and more halted, the MLU also decreases. A lower MLU and higher number of silent pauses can indicate speech that is less proficient due to its lack of connectivity.

By identifying recurring patterns of linguistic characteristics within these rehearsed segments, this study aimed to determine key differences between memorized and spontaneous speech on the OPIc, potentially authenticating observations based solely on rater impressions and not technical evidence. The research questions of this study are as follows:

1. What quantitative differences can be found in temporal fluency measures between the rehearsed and spontaneous samples?
2. What qualitative observations can be made from the transcripts of the audio recordings?

Methods

Instrument

This research used extant ACTFL OPIc test data from which a subset of different task types was sampled. The test as a whole has up to 17 tasks depending on the test form. For the IL interviewees, analyses were conducted on six speech task types: *talk about thing or place* (for which there were two tasks), *talk about activity or routine, ask questions, an intermediate role play,* and *past description.* The analyses for the IM interviewee included seven task types: the five task types given in the IL form and additionally *past narration* and *advanced role play.* This analysis focused on a six- to seven-task subset which had been selected in the previous study by Cox (2017, see Table 1).

Participants

In the study conducted in 2017 by Cox, the researcher conducted a stratified random sampling from a bank of English OPIc assessments. In order to control for L1 variance, the only exams sampled were those of native Korean speakers. These participants all worked for Samsung and had taken the English OPIc as required. To qualify for selection,

Table 1 Descriptions of task type analyzed by intermediate sublevel

Task type description	Intermediate low (IL)	Intermediate mid (IM)
Talk about thing or place	2	1
Talk about an activity or routine	1	1
Ask questions	1	1
Intermediate role play	1	1
Past description	1	1
Past narration	0	1
Advanced role play	0	1

each interview had to have been scored in perfect agreement by two or more raters. This selection process yielded 300 interviews, which were analyzed by nine trained raters who identified memorization as a hindrance to sublevel advancement (Cox 2017).

Instrument

To examine the quantitative temporal fluency differences, the syllable nuclei Praat script of De Jong and Wempe (2009) was used to find the MLU, number of silent pauses (SP), and articulation rate (AR) for each segment (rehearsed and spontaneous) in all flagged interviews (see example, Table 2). As mentioned earlier, it was determined that the speed measure of each sample would be determined by articulation rate, while speaker breakdown would be gauged through MLU and number of silent pauses. To determine what qualitative differences were present between the rehearsed and spontaneous tasks, all 247 segments were transcribed and analyzed. The qualitative analysis of the segments was conducted using a grounded theory approach, the rationale for which will appear later in this section.

Procedure

The rater comments for each participant were then analyzed and sorted. "Canned/memorized responses" and "rehearsed material" (p. 100) were rater terms quoted in the Cox (2017) study. A closer look at comments revealed that raters often used the abbreviation of "RM". Participants who at any point were flagged for rehearsed material were found by filtering rater comments with the words *rehearsed, memorized, canned, practice,* or *RM*. For the purposes of this study, any segments not flagged for memorization were operationalized as *spontaneous*. Thirty-nine participants IL (*n* = 24) and IM (*n* = 15) had been flagged by the original raters in any one of the six or seven OPIc tasks for memorization,

Table 2 Praat data sample of IL interviewee

Task type	Task level	Status	MLU	SP	AR
Talk about thing or place	Intermediate	Spontaneous	3.03	0.47	3.47
Talk about activity or routine	Intermediate	Spontaneous	3.32	0.44	3.68
Intermediate roleplay	Intermediate	Spontaneous	2.96	0.53	3.66
Talk about thing or place	Intermediate	Spontaneous	3.05	0.59	3.60
Ask questions	Intermediate	Spontaneous	3.60	0.43	3.84
Past description	Advanced	Rehearsed	2.97	0.54	3.97

yielding 69 rehearsed and 178 spontaneous segments (Fig. 1). Each interview contained an average of 1.77 rehearsed tasks and 4.56 spontaneous tasks.

Research question 1: differences in temporal fluency features

After compiling data for all three fluency measures, the overall average MLU, SP, and AR were noted for each of the 39 interviewees. One of the interviews had existing background noise that prevented the detection of MLU and number of silent pauses but not articulation rate, resulting in an n of 38 for MLU and number of silent pauses and an n of 39 for articulation rate. The overall averages for the entire sample in both spontaneous and rehearsed segments were then calculated.

Research question 2: qualitative differences between rehearsed and spontaneous speech

The segments were transcribed and analyzed for differences between the rehearsed and spontaneous segments of individual interviews as well as for patterns occurring in the rehearsed segments throughout all participants' interviews.

A grounded theory approach, rather than a typical discourse analysis, was used to conduct the qualitative analysis of the segments because "in grounded theory, which is

Fig. 1 Research procedure flow chart
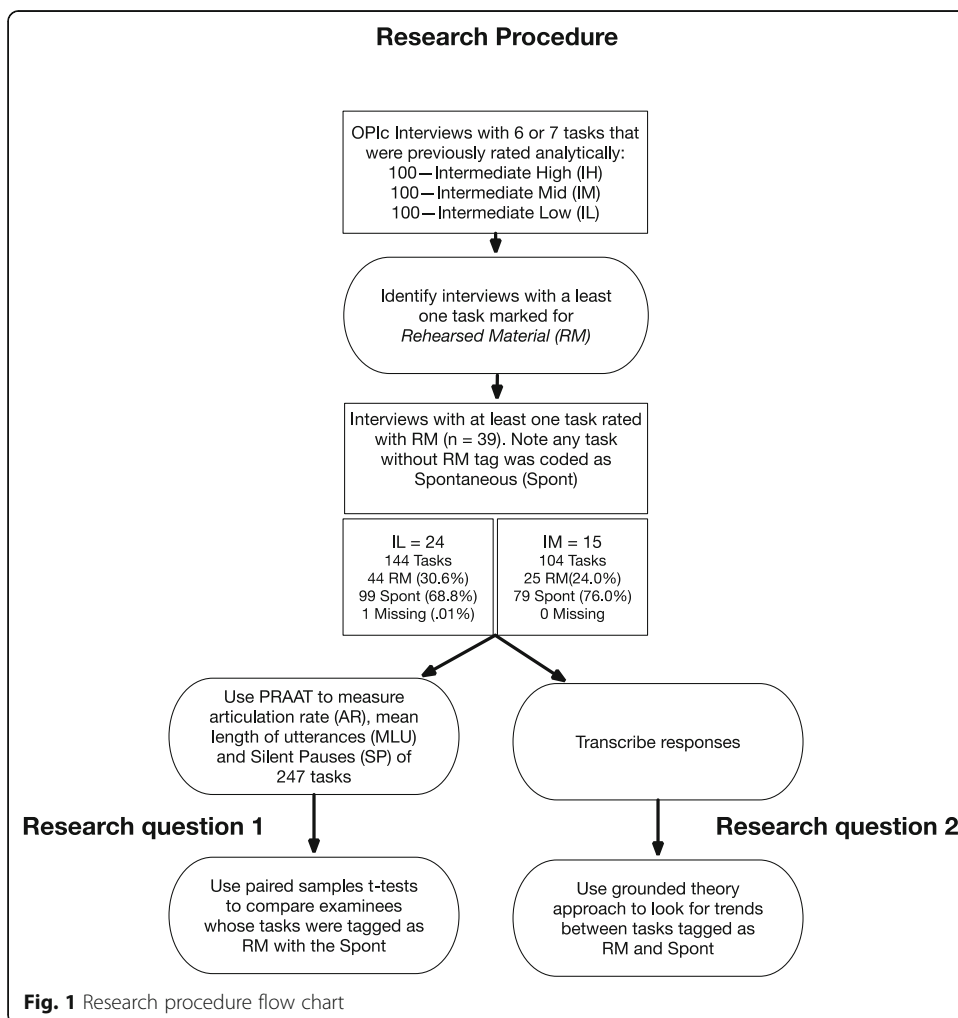
sometimes described as breaking down and then putting together the data, analysis seems often to consist of progressively more abstract categorization vs. analysis in the sense used in discourse analysis" (Wood and Kroger 2000, p. 29). In other words, grounded theory is not the use of data to support or reject a preformed hypothesis but rather the development of a theory *from* that collected qualitative data (Glaser and Strauss 1967).

Throughout the transcription process, one of the researchers made brief notes of any repeating patterns. Once all segments were transcribed, the researcher again reviewed the segments and rater comments, noting particular words or phrases (e.g., unusually high-level vocabulary or text type or patterns that recurred in other interviews). To ensure the dependability of the first researcher's conclusions, a separate spreadsheet listing all of the prompts, transcribed responses, and rater comments for all 39 flagged interviews was given to a second reader, who was instructed to read each item and note any patterns distinguishing the spontaneous samples from the rehearsed samples. This reader was a graduate research assistant who was familiar with the ACTFL proficiency guidelines. After completing an independent analysis, the two readers met to discuss their findings.

Results

Research question 1: differences in temporal fluency features

To determine whether there was a significant difference between the MLU, SP, and AR of spontaneous versus rehearsed segments, a paired samples *t* test was utilized to compare the sample means. There was no significant difference between the spontaneous and rehearsed segments in terms of MLU (Table 3, Fig. 2) or SP (Table 3, Fig. 3). There was, however, a significant difference between spontaneous and rehearsed segments in terms of articulation rate (Table 3, Fig. 4). This is not surprising, as sentences which have been practiced prior to an interview would be expected to have a faster delivery.
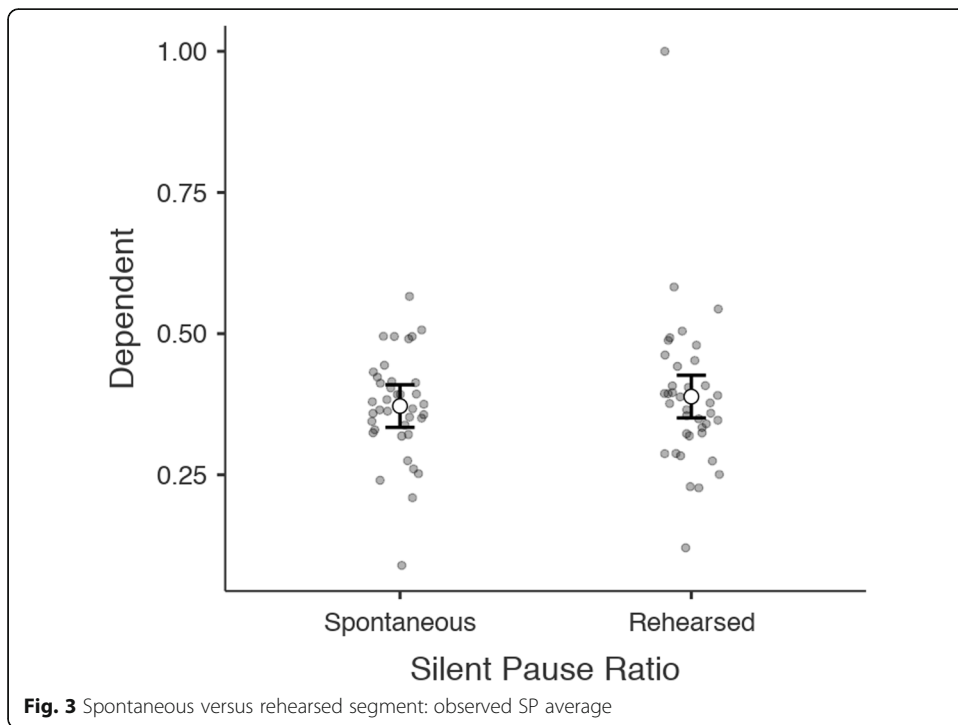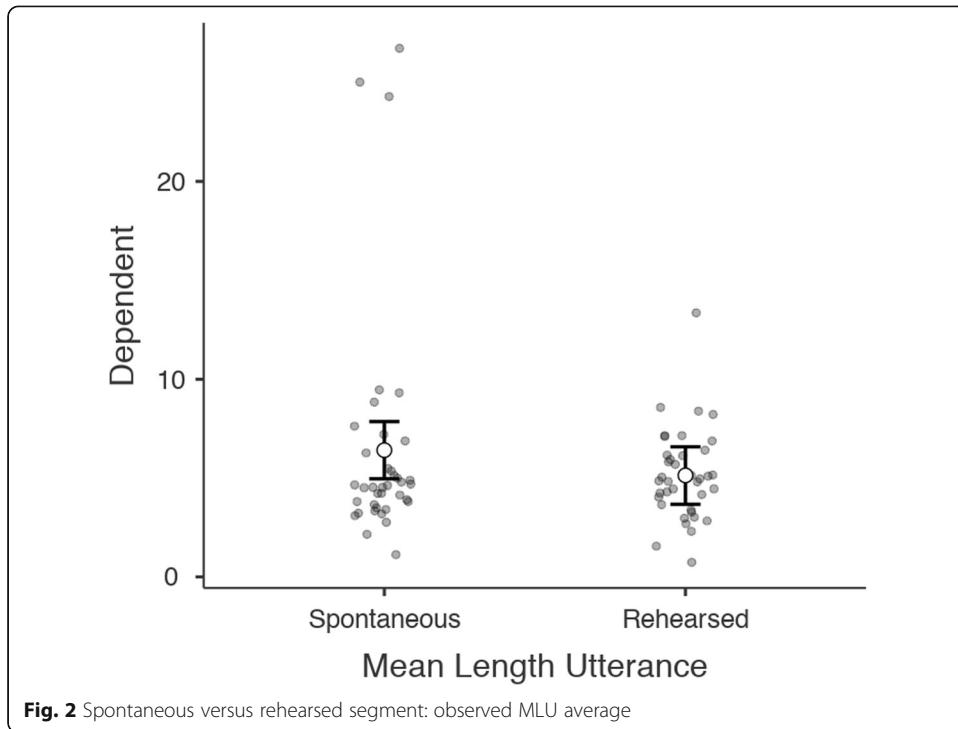
According to the quantitative statistical analyses, articulation rate was the only fluency measure that was significantly different between the rehearsed and spontaneous segments. This was somewhat expected because material that has been practiced would probably come more naturally to the speaker. The breakdown measures might not be significant because, even with memorization, there may still be pausing and hesitation involved in recitation. However, because only one temporal fluency feature yielded significant results, the quantitative analysis has not proven to be a strong demonstrator of speech type differences.
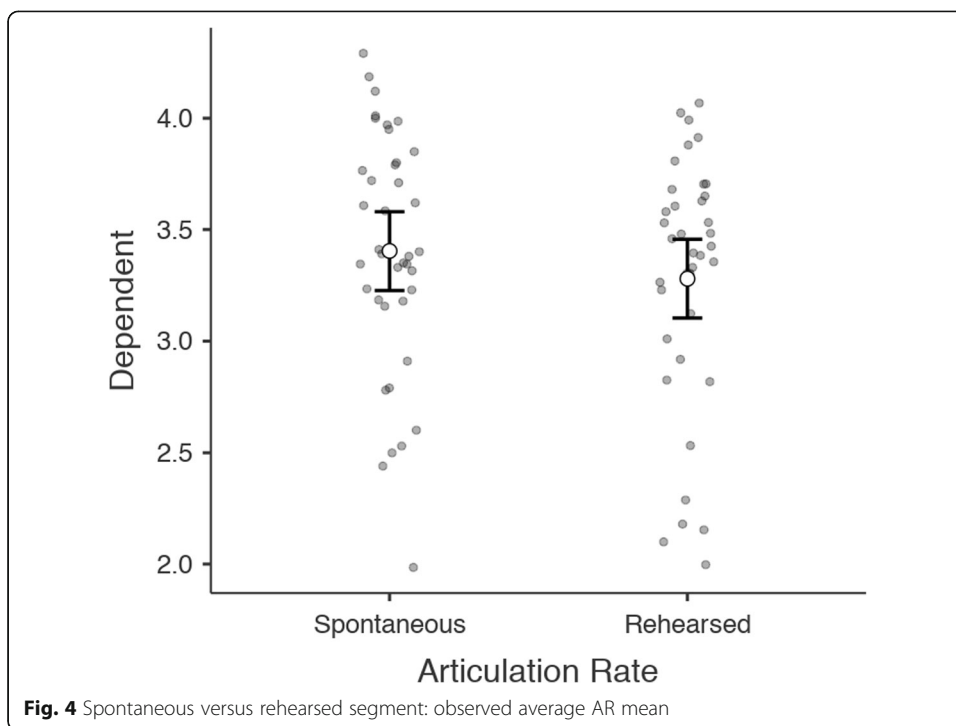
Research question 2: qualitative differences between rehearsed and spontaneous speech

The two qualitative researchers found two trends: within- and between-speaker. The within-speaker trend refers to a single speaker who had tasks that were tagged rehearsed and spontaneous. These were compared and aggregated to the whole group. Researchers noticed two within-speaker trends: off-topic responses and recurring material. The

Table 3 Statistics for spontaneous and rehearsed segments

	Spontaneous			Rehearsed							
	M	SD	95% CI	M	SD	95% CI	*n*	*t*	df	*p*	Cohen's *d*
MLU	6.41	5.92	± 1.88	5.13	2.29	± 0.73	38	1.29	37	0.20	0.21
SP	.37	.09	± 0.03	.39	.14	± 0.04	38	− 0.75	37	0.45	0.12
AR	3.28	.55	± 0.17	3.40	.54	± 0.17	39	− 2.74	38	0.01	0.44

Fig. 2 Spontaneous versus rehearsed segment: observed MLU average

Fig. 3 Spontaneous versus rehearsed segment: observed SP average

Fig. 4 Spontaneous versus rehearsed segment: observed average AR mean

between-speaker trends refer to how examinee responses to the same task types, due to the recurring phrases and repeating organizational patterns, may have been based on response "templates."

Within-speaker: off-topic responses

To analyze off-topic responses, researchers identified features as to why certain segments were marked as memorized. Some of these clues could be identified from within the single prompt response. One clue is a response that has little to do with the question asked; in other words, off-topic—a tag assigned by many raters. The off-topic responses to tasks 5 and 6 in Table 4 could be due to the speaker's misunderstanding of the prompt: confusion of *walk* for *work* (this was noted in the rater comments). However, while task 14 asks about a healthy friend, the interviewee speaks only about the gym he attends—a vaguely related subject, but clearly off-task.

It is worth noting that the accepted rhetorical strategy of one culture does not always easily transfer to another. Scholars have observed that although the communication style of Anglo-American educational traditions tends to be direct or linear (Eggington 2015; Wierzbicka 2006), Korean sociocultural norms call for rhetoric that is more relative to context (Eggington 2015; Goddard and Wierzbicka 2007). While native English

Table 4 Rehearsed tasks of interviewee 2

Task	Prompt topic	Response
5	Walking location	*I will talk about- I will talk about some household chores…*
6	Walking routine	*I will talk- I will- I will talk about my work…*
14	A healthy friend	*I will talk about gym I go to…*

writers favor direct, predictable rhetorical patterns in communication, Korean cultural scripts typically reject this assertive approach out of concern for imposing upon the reader's own interpretation of and assumptions from the text. Consequently, much of the Korean discourse which an American reader or listener may perceive as disorganized or circumlocutory is culturally valid and does indeed have a rhetorically based focus located deeper within the text (Eggington 2015). However, given that OPIc participants have a limited response time, and considering the degree of tangentiality in the examples, we feel that Korean rhetorical style does not have a significant influence upon the data at hand.

The final rehearsed task of interviewee 2 is markedly more off-topic and disorganized than the previous examples (Table 5). The speaker refers only twice to an element of the given prompt: childhood. Throughout the response, interviewee 2 attempts to steer the topic toward the neighborhood, the local park, and descriptions of these locations.

Other responses are similarly off-topic. Interviewee 3 connects the response to the given prompt in what seems like an afterthought ("because of what had in [unintelligible] Beach"), and otherwise describes only a music festival and her reasons for attending it (Table 6). It is suspected that the speaker prepared an experience to insert at some point in the interview.

Responses of interviewee 4 to tasks 5 and 14 are also off-task (Table 7). In task 5, numerous details are given about a trip (e.g., location, activities, and safety precautions) but interviewee 4 makes no mention of the people with whom the vacation is spent, as the question specifically asks.

Similarly, when asked to describe household responsibilities in task 14, the speaker gives a detailed description of a house but none on the members or their duties. Still, the task 5 and 14 responses are within the general topics of vacations and the home, respectively; the task 11 response, however, is completely off-topic. The speaker talks about a broken car that has resulted in missing a meeting (nothing to do with an e-reader as the prompt mentions). It seems that the speaker recognizes the prompt for roleplay, but was unable to deliver one that fit the scenario provided.

Within-speaker: recurring material

In addition to segments flagged by raters for rehearsed material manifest in off-topic responses, a second pattern was recurring (often verbatim) phrases and descriptions. These were noted in several rater comments: "This is a canned response we hear over and over," "Have heard this material before," "After hearing RM [rehearsed material] in response to other prompts, you realize this is also totally memorized." Remarks about repeating content formed 6% of rater comments overall.

The search for recurring patterns was widened to include segments marked as spontaneous to determine if the pattern continued. Interviewee 5 repeats most of the same response on two difference prompts by recombining memorized chunks of material.

Table 5 Interviewee 2 excerpted transcript

Prompt topic	Response
Childhood Home and Furniture	*I will talk about on- I will talk about on- I will talk about on in my childhood. Uh, I- I- I living in [unintelligible] Dong, in Seoul. Mm, it is- it is recreational area. Uh, I think my neighborhood in around very beautiful, because there are many trees. It is- no- sorry, sorry, I didn't- didn't [unintelligible] talk. I will talk about- I'll talk about event- I'll talk about…neighborhood events. I'll talk about- I'll talk about on my childhood event. I go to park in my neighborhood for running and walking.*

Table 6 Interviewee 3 excerpted transcript

Prompt topic	Response
Typical Beach Activities	*Um, one of my favorite events was- uh, one of my favorite events in [unintelligible] was the music festival in last year. I like music so much, so I like- I like go to the mu- um, concert. Um, uh, because of what had in [unintelligible] Beach. Anyway, I heard that- um, uh, that there will be a many famous celebrities, like Jason Mraz. So I was so excited. I really want- uh, want to go there. I-*

Responses are off- and on-topic for tasks 6 and 14, respectively, but are strikingly similar in several ways Table 8). Not only are many phrases repeated word-for-word, but the speaker's organization within both tasks, with the exception of some deviations in task 14, is virtually identical.

Between-speakers: recurring phrases

In other cases, patterns emerge across multiple interviews. An example of this is seen in task 14, an intermediate description prompt. Three interviewees were asked to describe three unique locations. Despite each prompt asking for three distinct places, the descriptions shared several characteristics (Table 9). It is worth noting that of these three very similar segments, only one was flagged as rehearsed.

Between-speakers: reuse of speaking templates

In addition to phrases, similar patterns of organization emerged across interviewees. Table 10 features the transcripts of three different interviewees, each of whom was asked to describe a memorable vacation. All three participants use not only similar sentence structure ("which is located," "it was for summer vacation," "on the first day," etc.) but the same sentence order: the name of destination, its location, summer vacation, and 3 days of activities.

One of the most striking patterns of memorization noted during transcription was the repeating story of encountering a long-lost friend. Table 11 shows three stories that are nearly identical in organization and content. All three of these responses were to task 10, which asks the speaker to describe an unexpected event that occurred while traveling.

Discussion

As mentioned earlier, statistical analyses showed that only the mean articulation rate differed significantly between the rehearsed and spontaneous segments. The qualitative analysis, in contrast, revealed far more about the nature of these rehearsed segments.

Table 7 Rehearsed tasks of interviewee 4

Task	Prompt topic	Response
5	Vacation companions	*I like to drive through, uh, lakeside area called Changpyeong for my vacation. . . .Um, when I- when- I carry very little cash when I travel. It makes me feel much safer.*
11	Role-play purchase of technological item	*Hello, I need to talk to you about something. The problem is that- the problem is that I can't attendant the meeting today because my car is broken. Um, I'm really sorry, but can you please reschedule the meeting?....*
14	Household responsibilities	*Um, sorry, I lived third floor on apartment building in [unintelligible]. There are four room, two bathroom, and a living room. It's different size for my family.. . . In my free time, I watch TV or relax- relax in the living room. In my, uh, my-my- my room is study room....*

Table 8 Recurring patterns within tasks 6 and 14 of interviewee 5

Prompt topic	Formulaic response	Unique response	Status
Work training	*Speaking my favorite sport, my favorite sport is with my family every Sunday. I usually in the park near my house because the park is very large and beautiful. Especially there is a very good I recently.*	*basketball. . . . I like to play basketball play basketball basket- basketball court in the park. played. . .*	R
Jogging location	*Speaking of my favorite sport, my favorite sport is with my family every Sunday. I usually in the park near my house because the park is very large and beautiful. Especially there is a very good I recently*	*jogging. . . . I like to go jogging go jogging walkway in the park. So, I go there by bus. When I go there, I usually bring something to drink. When went jogging. . . .*	S

Of the rehearsed segments that had been marked as rehearsed for off-topic material, several contained a polished enough response that they were assumed to have been prepared before the interview. Spence-Brown (2001) encountered a similar challenge while researching authenticity in course-based assessment when a particular student "indicated that he did not understand the preceding utterance, but drew on a rehearsed response that suggested that he did" (p. 470). The author cautions that such communication, however linguistically competent it may seem, is not authentic and thus lacks validity. That being said, it is difficult to know whether the response is off-topic because the speaker has intentionally made it so or because the prompt was misunderstood—a question that calls for further investigation

Several raters also justified their conclusions by noting that the speaker used a higher text type than expected. Text type is defined by ACTFL (2012c) as "that which the learner is able to understand and produce in order to perform the functions of the level" (p. 8) and includes words, questions, connectivity of sentences, and paragraph-length discourse. Text type that does not match what a speaker has produced in the remainder of an interview is a potential indicator of rehearsed material.

However, reaching agreement upon which segments demonstrated notably higher or lower text type than other segments within the interview is a challengingly subjective process. The most specific rater comments as to why a segment was marked as memorized involved the observation of recurring material; that is, the most striking evidence of memorization lay in recurring storylines given for the same task number or prompt across interviews (for example, encountering an old friend). Among the Korean-

Table 9 Recurring themes in task 14 responses of interviewees 7, 8, and 9

Prompt	Response
Describe where you go jogging.	*In the park, there are* **many trees. In the middle** *of park, uh, there's a- uh, uh, there's a- uh, uh, large* [unintelligible]. *Uh, from the* [unintelligible] *middle, there is a big runni- running* **track**. . . .
Describe the baseball field where you play.	*Um, it is surrounded with* **many flowers and trees**. *Um,* **in the middle**, *uh, there is a* **lake**; *around the lake, uh, walking- a walking track and a jogging* **track**. . . . *Uh,* **on the right**, *I can see, uh, soccer field*. . . .**On the left**, *many flowers and trees*. . . .
Describe your language school	*It is located in- um, uh, middle part of language campus*. . . .**In the middle**, *I can see the* **lake**. . . .**On the right**, *I can see the cafeterias and cafes,* **trees, and flowers**. *Uh,* **on the left**, *they are dormitory buildings*. . .

Table 10 Recurring themes in the intermediate description tasks 9 and 5 across interviews

Task	Response
9	*I took a trip to Jeju Island,* `which is located in` *south of Korea. . . .I went there by air,* `it was my summer vaca- summer vacation. When I stayed there a week, I- I had- I had, um,` *many thing.* `On a first day, I- I climb up the-` *I climbed up the Halla Mountain, it took about five hours. . . .On the sec- on the second of trip, I-* `I had to swimming. Uh, I wa- we- we were swimming, sunbathing, and sn- snorkeling and beach volleyball. And we ha- we had barbeque party at night.` *On the last day, we tried-* `we tried their local food,` *such as raw fish-*
9	*Last year I took a trip to Manila* `which is located in` *the Philippine. Uh,* `it is for my summer vacation. . . .I stayed there for one week. I enjoyed many things. On the first day,` *we visited famous place such as national museum.* `On the second day, we enjoy many summer sports, uh, and sunbathing. We had a barbeque party at night.` *We drank all night. On the last day, we went shopping and bought some souvenir.* `We try the local food. . . .`
5	*Mm, um, last year I got a trip to Jeju Island,* `which is located in` *south of Korea. Uh,* `it was for summer vacation. Mm, so, I go there with my friends. Uh.. We flew there by Korean Air, it took one hour. We stayed, uh, for a week and enjoyed many things. Uh, first, we climb the, uh,` *Halla Mountain. Uh...*

produced OPIc prep videos investigated for this study (Hackers Education Group 2014; Lee 2013), several of the most-viewed videos presented templates very similar to those heard in these interviews.

The videos included an overwhelming number of scenarios, specific vocabulary, expressions, and pronunciation tips (Lee 2013). Although one online commenter recommended that test takers should be careful to not mirror the examples too closely, and that it was wise to insert one's own experiences, examples, and ideas (Hackers Education Group 2014), these video resources would likely be more effective if they provided OPIc participants with an idea of what speech forms to expect but then encouraged the practice of those forms in realistic and personalized contexts instead of providing a script. Luk (2010) similarly noted student participants whose efforts to appear more proficient to their evaluators "might have resulted in a form of ritualized and colluded talk that does not represent an authentic replica of ordinary conversations" (pp. 49–50). In other words, participants should be encouraged to improve their actual English proficiency instead of memorizing prepared scripts which, when heard alongside several similar recitations in a testing group, are recognized by raters as rehearsed.

Table 11 Responses from three different interviewees to task 10 (all marked as rehearsed)

Uh, last year I took a trip, uh, Manila which is located in the Philippine. Uh, uh, one day I went to Manila while- while I was working on the shopping mall `. I met my old friend` *So Ra Kim.* `Suddenly I saw one person who looks like my friend.` `At first I just ignored her.` `A few seconds later, she called my name.` `I was very surprised` *to see her there. At first I was* [unintelligible]. *Soon,* `we realized each other and were so glad to meet again.` *We went to the bar for a casual drink.* `We talk about our life and school.` *I had a good chat-* `a good time.` `I had a good time.` `We still keep in touch each other.`
Um, I, uh, last- last month, I went- I went the Europe with my best friend CheolSu. Uh, when I go there, fortunately `,` `I met a old friend,` *YoungHee. First-* `at first, I ignore her, because-` *uh, because- ah, no no, uh,* `at first, I ignore her.` *Uh,* `few sec- few seconds later, I and her- I and she- I and she are recognize each other.` `I was very surprise.` *So- so, and sh- because* `she called my name,` *we- we- we have met- we have not met for twenty years- no, ten years. So, very, very,* `very glad to meet again.` *Uh, after- after- after traveling these place,* `we go to the bar to talk about life and school.` *And we- we kee-* `we still keep in touch each other.` *Thank you.*
Wow, it is very interesting question, let me see. Uh, two years ago, I went there Tokyo with my family. At the time I was very surprised experience. Uh `,` `I was very surprised` *because* `I met my old friend` *SuRa Kim. Um, before, we never met each other for five years . . .* `. I- firstly, I ignored her, but` `she called my name.` `I really surprised` *to see her there. We went shoppi- we shopping together, and after that, I- oh, after that* `we went to the cafeteria` *for lunch. And then,* `we talked about our future and school.` `We had a good time. . . .`

Limitations and considerations for future research

A limitation of this research lies in the selection of the three fluency measures (AR as a measure of speed and MLU and number of SP as a measure of breakdown) and how much these vary between speech in the L1 versus the L2. For example, memorized speech had a significantly higher articulation rate than spontaneous speech, but this evidence would be stronger had the articulation rates of spontaneous and rehearsed *Korean* speech also been calculated. De Jong et al. (2015) collected both L1 and L2 data which demonstrated that pausing behavior overlaps in L1 and L2 speech, as well as that the inverse articulation rates of both are more strongly correlated after taking into consideration L1 fluency behavior.

Another limitation is this research used existing data that was not designed from the outset to find examples of memorization. Since the raters were tasked with judging speech samples, it is possible that they focused instead on the degree that the response was off-topic or off-task instead of detecting rehearsed speech. While this is a strength in that it reflects real-world rating, it is also a weakness in that it is difficult to know the extent to which other memorized samples were missed. A similar issue is the fact that rehearsed speech characteristics were found in segments which had not been flagged for memorization by original raters (segments that, for the purposes of this study, were referred to as spontaneous instead of rehearsed). This resonates with the findings of Xie (2013) and Messick (1982) that a test taker's ability can be overestimated from rote memorization practices, rendering the assessment invalid. Thus, a carefully designed experimental study to follow as a follow-up would be in order.

Conclusion

When actors are unable to remember their part, they will often call out "Line, please" to help them as they rehearse their performance. However, there is a difference between the performative aspect of rehearsed scripts and actual language proficiency assessment. The tendency to memorize rote responses threatens the score validity based on the CAL validation framework (Kelly et al. 2017) because memorized responses may not apply to other speaking situations and raters may not consistently detect or score rehearsed speech. While rehearsed speech is acknowledged as a proven strategy for acquiring vocabulary or formulaic sequences and preparing for speeches or exams, it does not reflect real-world communication. A person cannot be considered proficient (using the ACTFL guidelines) beyond the level in which they can use language spontaneously: "An individual cannot 'cram' for a proficiency test . . . . According to the research on second language acquisition, several hundred hours . . . meaningfully engaged in honing and using language skills . . . are required to progress to the next highest sublevel or major level" (ACTFL 2004). And while this is ACTFL's position, this does not stop more adept actors from attempting to receive higher ratings without the prerequisite proficiency.

However, we may wisely question some decisions based upon test scores vis-à-vis the CAL Validation framework (Kelly et al. 2017). McNamara et al. (2019) note the difference between *fairness* and *justice* in language testing. Rating all samples with same standards of detecting rehearsed speech and then applying the same criteria in scoring the samples is an issue of fairness. Justice, however, determines if the use of a test is appropriate for the circumstances. "Our current [language] models. . . emphasize aspects of language that seem salient to us but may not be the ones that serve people well in

the complex acts of communication that they engage in" (Elder et al. 2017, p. 20). Many participants in this study aimed for an intermediate high rating, but their use of rehearsed material and consequent lack of observable proficiency awarded them a score of intermediate mid or even intermediate low. Could this be because the level of English proficiency actually needed for their job is not as high as the one set by their employers?

If this is the case, it raises additional questions for future research: what level of English proficiency is required to perform their job functions? Do they need their employees to communicate well in English to remain competitive in a globalizing economy or is English proficiency a proxy measure for some other factor such as socioeconomic status? Increasing the standard of English proficiency adds yet another element of rigor to hiring and promotion that perhaps is not enforced for the sake of international communication alone and can threaten the *justice* of the test use.

Returning to Messick's scholastic coaching outcomes (Messick 1982), perhaps many OPIc test takers are producing rehearsed responses because they believe this reflects "genuine improvements in the abilities measured by the test" or "enhanced test-taking sophistication" when in reality this strategy is a "test-taking artifice" (p. 67). Given trained raters' recognition of rehearsed material and its impact on proficiency ratings, many test takers are adopting a seemingly helpful stratagem that will yield disappointing results on an oral proficiency interview. Employers, policy-makers, educators, and researchers ought to consider a reasonable standard of English proficiency for their institution's purposes, what a proficiency interview aims to assess, and how to increase test takers' awareness of those measures.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s40468-020-00110-5.

Additional file 1. LTIA Revision chart.

Author details
[1]Brigham Young University, Provo, UT 84602, USA. [2]Divine Word College, Epworth, IA 52045, USA.

References

American Council on the Teaching of Foreign Languages. (2004). *ACTFL Policy on Oral Proficiency Interview Retesting.* https://www.regent.edu/acad/undergrad/pdf/ACTFL_Policy

American Council on the Teaching of Foreign Languages (2012a). *Oral Proficiency Interview Familiarization Manual.* Alexandria, VA: ACTFL.

American Council on the Teaching of Foreign Languages (2012b). *Oral Proficiency Interview Computerized Familiarization Manual.* Alexandria, VA: ACTFL.

American Council on the Teaching of Foreign Languages (2012c). *Performance Descriptors for Language Learners, 2012 Edition.* Alexandria, VA: ACTFL.

Becker, C. B. (1986). Reasons for lack of argumentation and debate in the Far East. *International Journal of Intercultural Relations*, *10*(1), 75–92.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* London: Longman, Green, & Co. Ltd.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*(3), 245–261.

Burton, J.D. (2020). "How scripted is this going to be?" Raters' views of authenticity in speaking-performance tests. *Language Assessment Quarterly*, 1-18. https://doi.org/https://doi.org/10.1080/15434303.2020.1754829

Cox, T. (2017). Understanding Intermediate-level Speakers Strengths and Weaknesses: An Examination of OPIc Tests from Korean Learners of English. *Foreign Language Annals*, *50*(1), 84–113. https://doi.org/10.1111/flan.12258.

De Jong, N. H. (2016). Fluency in second language assessment. In D. Tsagari, & J. Banerjee (Eds.), *Handbook of Second Language Assessment*, (pp. 203–218). Berlin: Mouton deGruyter.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

Deuchler, M. (1992). *The Confucian transformation of Korea: A study of society and ideology.* Cambridge: Harvard University Press.

Eggington, W. G. (2015). When everything's right, but it's still wrong. In N. W. Evans, N. J. Anderson, & W. G. Eggington (Eds.), *ESL readers and writers in higher education: Understanding challenges, providing support*, (pp. 198–208). New York: Routledge.

Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language and Communication*, *57*, 14–21.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399 https://doi.org/10.1177/0265532210364407.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* New Brunswick: Transaction Publishers.

Goddard, C., & Wierzbicka, A. (2007). Semantic primes and cultural scripts in language learning and intercultural communication. In F. Sharifan, & G. Palmer (Eds.), *Applied cultural linguistics: Implications for second language learning and intercultural communication.* Amsterdam: John Benjamins.

Hackers Education Group. (2014). *If you prepare an OPIc, you must be sure to listen to the essential OPIcs!* [Video file]. https://www.youtube.com/watch?v=CoJKLbTv3c4

HiEnglish Korea. (2015). [Job posting for corporate OPIc instructor]. www.myenglishjob.co.kr/view/employer_info_view.php?select_num=6873&select_no=9 169

Jae-gon, K. (2005). The end of the road for the TOEIC? *The Chosun Ilbo.* http://english.chosun.com/.

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

Jambor, P. Z. (2011). English Language Necessity: What It means for Korea and Non-English Speaking Countries. ERIC Institute of Education Sciences Document Number 528279. Retrieved from https://eric.ed.gov/?id=ED528279.

Jeon, J. (2010). Issues of English tests and assessments: A view from Korea. In Y. Moon, & B. Spolsky (Eds.), *Language Assessment in Asia: Local, regional or global?* (pp. 55–80). Seoul: Asia TEFL.

Kelly, J., Renn, J., & Norton, J. (2017). Addressing consequences and validity during test design and development. *Useful assessment and evaluation in language education*, 185-200.

Kember, D. (2000). Misconceptions about the learning approaches, motivation, and study practices of Asian students. *Journal of Higher Education*, *40*, 99–21.

Kim, N. (2016). Korean test takers TOEIC-speaking and OPIc test preparation. [Abstract]. *Korean Journal of Applied Linguistics 32*(3), 51-76. Abstract retrieved from https://doi.org/10.17154/kjal.2016.9.32.3.51

Ko, M. S. (2010). Improving English speaking proficiency with OPIc: Focusing on the contents of the story. [Abstract]. *Research Institute of Curriculum & Instruction, 14*(1), 21-45. Abstract retrieved from http://doi.org/https://doi.org/10.24231/rici.2010.14.1.21

Ko, M.S. (2017). Studying textbooks for speaking OPIc-focusing on content composition and writing practice for speaking English about yourself. [Abstract]. English Education Research, 50, 51-71. Abstract retrieved from http://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE09233061

Kim, S., Fu, P., & Duan, J. (2017). Confucianism and human resource management in East Asia. In F. L. Cooke & S. Kim (Eds.) Routledge Handbook of Human Resource Management in Asia (pp. 46- 65). London: Routledge.

Lam, D. M. K. (2015). Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong. In G. Yu, & Y. Jin (Eds.), *Assessing Chinese learners of English: Language constructs, consequences and conundrums*, (pp. 38–60). Basingstoke: Palgrave Macmillan. Reproduced with permission of Palgrave Macmillan.

Lee, H. S. [MulticampusEdu]. (2013, July 3). *Dr. OPlc Lee Hyun Suk's OPlc concert* [Video file]. Retrieved from https://www.youtube.com/watch?v=CoJKLbTv3c4

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, *7*(1), 25–53. https://doi.org/10.1080/15434300903473997.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment*. Oxford: Oxford University Press.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, *17*(2), 67–91. https://doi.org/10.1080/00461528209529246.

Myung, J. K. (2012). [Sole] Samsung, TOEIC (abolition) … Unique with OPlc (OPlc). *Asian Economy*. http://view.asiae.co.kr/news/view.htm?idxno=2012103011033848604&nvr=Y

Park, J. S. (2011). The promise of English: Linguistic capital and the neoliberal worker in the South Korean job market. *International Journal of Bilingual Education and Bilingualism*, *14*(4), 443–455.

Park, D., & Lee, S. (2019). A study on suprasegmentals in the rated Korean English learners' OPlc speech data. Lingua Humanitatis, 21(2), 333-351. https://www.earticle.net/Article/A36716.

Rubio, F., & Hacking, J. (2019). Proficiency vs. performance: what do the tests show? In S. Gass, & P. Winke (Eds.), *Foreign language proficiency in higher education*, (pp. 137–152). Berlin/New York: Springer.

Seo, M., & Chang, S. (2013). A pilot case of an OPlc test oriented class: The effect of providing formative assessments. *Journal of the Korea English Education Society*, *12*(2), 217–236.

Seth, M. J. (2002). *Education fever: Society, politics, and the pursuit of schooling in South Korea*. University of Hawaii Press.

Shin, D., Kim, N., & Kang, S. (2010). Korean examinees perceptions of testing spoken English: A move toward proficiency. *Studies on English Language and Literature*, *36*(2), 263–288. https://doi.org/10.21559/aellk.2010.36.2.015.

Spence-Brown, R. (2001). The eye of the beholder: authenticity in an embedded assessment task. *Language Testing*, *18*(4), 463–481.

Surface, E. A., Harman, R. P., Watson, A. M., & Thompson, L. F. (2009). *Are human- and computer- administered interviews comparable?* In *Paper to be presented at the 24th annual meeting of the Society for Industrial and Organizational Psychology*. LA: New Orleans.

Surface, E. A., Poncheri, R. M., & Bhavsar, K. S. (2008). *Two studies investigating the reliability and validity of the English ACTFL OPlc with Korean test takers*. Raleigh, NC: SWA Consulting Retrieved from www.languagetesting.com/pub/media/wysiwyg/research/ACTFL-OPlc-English-Validation-2008.pdf.

Shin, D. C. (2011). Confucianism and democratization in East Asia. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139084086.

Takeuchi, O. (2003). What can we learn from good foreign language learners? A qualitative study in the Japanese foreign language context. *System*, *31*, 385–392.

Thompson, G., Cox, T., & Knapp, N. (2016). Comparing the OPI and the OPlc: The effect of test method on oral proficiency scores and student preferences. *Foreign Language Annals*, *49*(1), 75–92 https://doi.org/10.1111/flan.12178.

Van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, *47*(3), 507–546.

Whan-yung, K. (2009). [In depth interview] English more than just a resume-builder. *Korean JoongAng Daily*. http://mengnews.joins.com/view.aspx?ald=2910932

Wierzbicka, A. (2006). *English: Meaning and culture*. Oxford: Oxford University Press.

Wood, K. M., & Kroger, R. O. (2000). *Doing discourse analysis: Methods for studying action in talk and text*. Sage Publications, Inc.

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier, & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching*, (pp. 123–146). Amsterdam: John Benjamins Publishing.

Yao, X. (2000). An introduction to Confucianism. Cambridge: Cambridge University Press.

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly*, *10*(2), 196–218. https://doi.org/10.1080/15434303.2012.721423.

Publisher's Note