Version of Record: https://www.sciencedirect.com/science/article/pii/S1048984320300369
Manuscript_7c44db11d6ac355064cb15a3bb40ab27

Testing the Babble Hypothesis: Speaking Time Predicts Leader Emergence in Small Groups

Neil G. MacLaren[1], Francis J. Yammarino[1], Shelley D. Dionne[1], Hiroki Sayama[1], Michael D. Mumford[2], Shane Connelly[2], Robert W. Martin[2], Tyler J. Mulhearn[3], E. Michelle Todd[2], Ankita Kulkarni[4], Yiding Cao[1], and Gregory A. Ruark[5]

[1]Binghamton University, State University of New York

[2]University of Oklahoma

[3]Neurostat Analytical Solutions, LLC

[4]Drexel University

[5]United States Army Research Institute for the Behavioral and Social Sciences

Author Note

Declarations of interest: none.

The research described herein was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army (Grant No. W911NF-17-1-0221). The views expressed in this manuscript are those of the authors and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.

Corresponding author is Neil G. MacLaren, nmaclar1@binghamton.edu, School of Management, Binghamton University, PO Box 6000, Binghamton, NY 13902-6000, UNITED STATES.

Testing the Babble Hypothesis: Speaking Time Predicts Leader Emergence in Small Groups

Attributions of leader emergence tend to be highly correlated with speaking time: those group members who speak the most also receive the highest ratings on a wide variety of leader-, communication-, and contribution-related measures (Schmid Mast, 2002). Bass (1990) rejected what he called a "babble hypothesis" of leadership that proposed that only this quantity, amount of talking, determined leader emergence. Instead, Bass argued that both quantity and quality mattered, citing studies that suggested speech must be relevant and beneficial to the group to facilitate leader emergence. Furthermore, Bass (1990) noted the essential endogeneity of speaking time as a correlate of leader emergence: group members seem to regulate the speaking time of other members, suggesting that speaking time may be a result of group processes, perceptions, or other factors instead of being a cause of leader emergence.

Three decades after Bass's (1990) assertions in the third edition of his handbook the situation remains unsettled. One concern may be that the relationship between speaking time and leader emergence has been amply, but perhaps problematically, demonstrated. For example, previous experimental studies have demonstrated a potential causal role for speaking time in leader emergence, but used study design features that may limit internal and external validity (Lonati, Quiroga, Zehnder, & Antonakis, 2018; Schmid Mast & Hall, 2004). Other studies have used observational designs in both leaderless group discussion (LGD) and problem solving environments, improving ecological validity by allowing participants to access a range of cues from other participants or the problem itself. However, these studies have not tended to control for relevant participant traits and other variables considered important in leader emergence (Ensari, Riggio, Christian, & Carslaw, 2011; Van Dijk, Meyer, Van Engen, & Loyd, 2017; Zaccaro, Green, Dubrow, & Kolze, 2018) and have not addressed Bass's endogeneity concerns (see Antonakis, Bendahan, Jacquart, & Lalive, 2010).

Ecologically valid studies supported by appropriate data collection and analysis techniques are needed to more thoroughly test the babble hypothesis. This study will attempt such a test by estimating the role of speaking time in relatively unconstrained problem-solving groups.

Although similar efforts have been made before, we will use a more comprehensive set of predictor variables, greater control of the temporal influence of those variables, and a more sophisticated set of analyses, allowing for stronger inference with respect to the babble hypothesis.

A Review of Speaking Time and Leader Emergence

Bass (1990) argued for a dichotomy of quality and quantity of speech, but it was the empirical correlation between these two constructs that initiated interest in the relationship between speaking time and leader emergence. By using electronic timers to record the duration of participant speech, Bass (1949) found that participant speaking time was highly correlated (between 0.82 and 0.92, 0.93 for all items combined) with co-participant ratings on a variety of leadership-related items. The content of the rating items referenced a range of target behaviors and rater attributions, but did not reference speaking or other contribution amount (Bass, 1949): a content-free measure of speaking act duration appeared to be closely related to an apparently content-driven act: leadership.

In the years since Bass published these findings researchers have repeatedly investigated speaking time as a correlate of perceptions of leader emergence as well as its relationship to other assessment procedures, individual differences, task expertise, and group structure. This literature has been reviewed several times (Bass, 1954, 1990; Mullen, Salas, & Driskell, 1989; Schmid Mast, 2002; Stein & Heller, 1979), and only a few studies have been published in this area since Schmid Mast's (2002) review. We will therefore focus not on the quantitative outcome of the studies (see Schmid Mast, 2002), but on how the studies were conducted and their qualitative contributions to understanding the relationship between speaking time and leader emergence. For comparability purposes we will restrict our attention to studies that explicitly assessed speaking time by recording participant speaking time directly, by counting words in a transcript or speech acts through observation, or by face-valid survey item (Norfleet, 1948).

Expertise and Speaking Time

Of the studies that Bass (1990) used to support his arguments, perhaps the clearest evidence in favor of an expertise effect comes from Gintner and Lindskold (1975). Gintner and

Lindskold (1975) used an ostensibly LGD scenario, the object of which was to guess the names of two paintings, and manipulated confederate "expertise" ("expert" vs. "inexpert"), confederate speaking time (high participation amount vs. low participation amount), and task ambiguity (high ambiguity vs. low ambiguity). The effects of these manipulations were related to each other in ways that support a stronger role for expertise than quantity of speaking time in this task environment. For example, the confederate received more leader emergence votes in the "expert" than the "inexpert" condition regardless of speaking time, but within the "inexpert" condition the confederate received more votes when she spoke more. Although Gintner and Lindskold (1975) used several design features that may not be considered best practice—including a relatively low number of replications per treatment and the use of deception (Lonati et al., 2018)—the study's results do, taken at face value, seem to support Bass's (1990) assertion that speech quality is more central to emergence than quantity.

Bass (1990) also cited Sorrentino and Boutillier (1975), but the results from that study were more mixed. In this study the information available to participants was even more constrained than in Gintner and Lindskold (1975): in addition to placing the participants in separate rooms, there was no actually correct answer to guess in the study task. After the task was over, participants were asked to rank the four members of their group, including themselves, on seven leadership-related items. Sorrentino and Boutillier (1975) found that increased speaking time, manipulated by means of a confederate, increased the confederates' ratings on each item except contribution, whereas the frequency of the confederates' "correct" answers influenced contribution but did not influence confidence, interest, or task leadership ability. Thus, when the confederate's contributions appeared to be "correct" more often, the confederate received increased scores on certain leadership-related items, supporting a quality-based view, but it is difficult to deny the apparent influence of speaking quantity in this study.

The Sorrentino and Boutillier (1975) study restricted participant access to reliable information about the task and other participants in order to implement the planned contrasts, but Bottger (1984) used a richer task environment to investigate the effects of expertise. Bottger

(1984) used the "NASA moon survival test" with groups of both managers and students: participants were asked to rank 15 equipment items in order of importance for surviving a crash landing on the moon. Bottger (1984) calculated the difference between rank orders assigned by participants and an expert opinion, the preferred ranking provided by NASA, thus gaining some estimate of participant "expertise". The difference between the individual participants' pre-discussion scores and the group's post-discussion scores provided an estimate of "actual" influence. Neither interaction nor information discovery was constrained within the discussion setting.

Bottger (1984) reported a low overall correlation between "expertise" and speaking time but a high correlation between perceived influence and speaking time and between "expertise" and "actual" influence. Although "expertise" and "influence" may be somewhat confounded in this study because of how each was calculated, Bottger (1984) concluded that, "Expertise is a stronger predictor of perceived influence when ability and air time [i.e., speaking time] are correlated than when they are unrelated... Also, expertise is a stronger predictor of actual influence when cues ['expertise' and speaking time] covary than when they are independent" (Bottger, 1984, p. 217). In other words, Bottger (1984) assessed participants as most influential when they both spoke a lot in the discussion and were inferred to have higher expertise.

Bass and Bales

The tentative conclusion from Gintner and Lindskold (1975), Sorrentino and Boutillier (1975), and Bottger (1984) is that both quality and quantity of speech appear to have independent effects, but that, perhaps depending on the situation, speech quality, presumed to be associated with expertise, may have a stronger influence. Some studies have tried to investigate that presumption directly. If, as Bass (1990) argues, quality of speech is related to its quantity and to leader emergence, then assessment procedures that rely on speech content should show incremental validity over content-free assessments. A well-known content-based method interaction process analysis (Bales, 1950), and several studies have used forms of the Bales method to investigate the relationship between speaking time and leader emergence. In the

following discussion, we will refer to direct assessment of speaking time, or its variants as noted above, as a Bass-like method (Bass, 1949) whereas the Bales interaction process analysis, or a derivative thereof, will be referred to as a Bales-like method (Bales, 1950).

Kremer and Mack (1983) randomized participants to single-gender LGD groups, but first obtained a sequence of decisions in a game called "Leader", a variant of the "Prisoner's Dilemma" game. Afterwards, the LGD participants were asked for nominations for who they considered to be either the task or socioemotional leader in their discussion group. Behavior in the "Leader" game, which was not observed by other participants, correlated with these post hoc ratings such that male leaders appeared to display different behaviors in the game than female leaders did. However, those differences were generally not evident when LGD phase leaders were assessed with either speaking time or interaction process analysis categories. In other words, both the Bass-like and the Bales-like methods appeared to differ systematically from post-hoc participant attributions in similar ways. This study relied on observed participant performance in the "Leader" and observed participant gender to delineate important contrasts. However, given that other potentially important covariates, such as intelligence and personality, were not assessed and a statistical control for endogeneity was not implemented, these findings cannot directly contradict the studies reviewed above.

Morris and Hackman (1969) directly compared the Bass-like and Bales-like methods using 3-person groups who produced four written products based on assigned tasks. Tasks were varied according to difficulty and type, and a full interaction process analysis was conducted. Morris and Hackman (1969) used creativity rankings provided by 25 judges as a performance outcome and correlated this group-level outcome with individual-level counts of behavioral categories. When raw counts of behavioral categories were used, Morris and Hackman (1969) found that most categories had significant zero-order correlations with the single leader emergence item mentioned above. However, when category counts were instead expressed as rates only 7 out of 48 pairwise correlations remained statistically significant: when accounting for speaking time, few behavioral categories appeared to differentiate leaders from non-leaders. Although Morris and

Hackman (1969) did not use a true control condition and may not have dealt with levels of analysis in the most appropriate way, the central finding would nevertheless be surprising if the Bales-like method was able to recover substantial incremental validity over and above the Bass-like method.

Morris and Hackman (1969) treated the Bass-like and Bales-like methods as distinct, but combining the two methods does not appear to yield much incremental improvement. Kirscht, Lodahl, and Haire (1959) collected observations from 22 pairs of 3-person groups who were asked to discuss a problem, then choose a representative from their subgroup to discuss a related problem. Although Kirscht et al. (1959) used an eight-category interaction process analysis, they reported on only three: one category each for giving and asking for suggestions and a third for attempts to structure what had been discussed. Kirscht et al. (1959) found that the Bass-like and Bales-like methods resulted in similar zero-order correlations with the subgroups' choice of representative (0.54 and 0.53, respectively) and that the two methods correlated (r = 0.39). Using either method alone to predict the chosen representative resulted in correctly classifying 14 out of 22 group representatives, but the sets were not entirely overlapping. However, combining the two methods post hoc resulted in an only modest improvement in classification performance, increasing from 64% successful classification to 73%.

Contribution Quantity Matters More in Some Studies

Thus, there seems to be evidence that Bales-like content-based methods provide some, but perhaps not substantial, incremental validity over Bass-like content-free methods—although, as originally suggested by Bales's analysis of proportions, accounting for content-free quantity appears to be necessary for proper interpretation of content-based counts of behaviors (Bales, 1950; Gerpott, Lehmann-Willenbrock, Voelpel, & van Vugt, 2019; Morris & Hackman, 1969). Some studies, however, have shown little to no effect of contribution quality.

Riecken (1958) had 32 four-member groups discuss solutions to a series of three business problems. In half of the groups, a hint about the solution to the third problem was given to the participant who spoke the most during the first two discussions; in the other half the hint was given to the participant who spoke the least. In 11 out of 16 groups in which the participant with

the highest speaking time received the hint, the hinted solution was chosen. However, if the hint holder was the lowest speaker, the hinted solution was not chosen in 11 out of 16 groups. Thus, the data in Riecken (1958) suggest that group solution acceptance was more correlated with participant speaking time than with possession of the hint. These results contrast with the results of Gintner and Lindskold (1975): similar operationalizations of "expertise", that is, the possession of unique and obviously helpful information, had little effect in Riecken (1958) but significant effect in Gintner and Lindskold (1975).

Jaffee and Lucas (1969), like Sorrentino and Boutillier (1975), used a game-like environment with no objectively correct solution for participants to find; like Riecken (1958) and others, Jaffee and Lucas (1969) used a high/low treatment comparison. Jaffee and Lucas (1969) presented participants with a variety of lights and shapes and asked them to discuss which light would activate next. Each participant then provided an individual guess and a vote for which participant could best lead the group in the following round. One of the group members was a confederate who, according to the treatment condition, either spoke a lot but provided no "correct" answers or spoke very little and provided "correct" answers half of the time. Each group experienced both conditions for half of the experiment session in a random order. Given this environment, speaking time correlated positively ($r = 0.63$) with leader evaluations across all conditions. Additionally, the confederate was chosen as the leader more often during the periods in which she spoke more despite the fact that she was never "correct" during those periods.

Perhaps the strongest evidence for a causal influence of speaking time, regardless of speaking quality, was the operant conditioning study of Bavelas, Hastorf, Gross, and Kite (1965). In this study, participants in the treatment condition were told that equipment in front of them would provide feedback about the quality of their contribution, encouraging more statements similar to what had supposedly been previously determined to be valuable contributions. In fact, however, a relatively quiet participant was selected and feedback was provided to discourage other participants from speaking and encourage the target participant to speak more. Participants in the control condition received no feedback. In both conditions participants were asked to complete a

questionnaire after each of three discussions that included four leadership-related items; the treatment was applied during the second discussion based on the speaking time measurements from the first. Bavelas et al. (1965) documented a significant increase in the amount of time the target participant spoke during the operant conditioning phase, accompanied by a significant increase in the average leadership scores the target participant received, which subsided somewhat in the third discussion. No such change was documented for the equivalent participant in the control group.

Bavelas et al. (1965) used a more powerful design than several other studies in this literature, but still relied on deception and a constrained information environment. However, the Bavelas et al. (1965) results were corroborated by Littlepage, Schmidt, Whisler, and Frost (1995). Littlepage et al. (1995) conducted a very similar study to Bottger (1984), using the same "expertise" and "influence" calculations and a similar task, the "desert survival" task. Littlepage et al. (1995) also used a structural equation modeling (SEM) framework, arranging variables in the model according to the temporal order in which variable values were argued to be set. Littlepage et al. (1995) found no significant influence of "expertise" on "influence", perceived or "actual". These latter findings suggest either that the "expertise" effect was not recovered in a new sample, was not robust to analysis methodology, or both. Furthermore, whereas Littlepage et al. (1995) assessed personality variables and found modest effects for confidence, extraversion, and dominance, Bottger (1984) did not assess personality, and neither study assessed general cognitive ability or gender. As such, omitted variables could be biasing the results of either or both studies. Furthermore, because Littlepage et al. (1995) was not a precise replication attempt of Bottger (1984) there may be other explanations for the differing results.

Influence of Task-Irrelevant Attributes

Only two of the studies cited above (Bavelas et al., 1965; Gintner & Lindskold, 1975) use an experiment with a formal treatment and control (see Lonati et al., 2018), and these studies seem to provide conflicting findings. The two studies that provide the richest information environment for the participants also lead to conflicting conclusions (Bottger, 1984; Littlepage et

al., 1995). Furthermore, despite the consistent, if moderate, correlation generally found between leader emergence and traits such as intelligence and extraversion, the few studies that have assessed these traits and measured speaking time have also found conflicting results (e.g., Littlepage et al., 1995; Riggio, Riggio, Salinas, & Cole, 2003; Ruback & Dabbs, 1986; Ruback, Dabbs Jr, & Hopper, 1984).

One potential weakness in this literature as reviewed thus far is the tendancy to ignore gender, a variable commonly found to be related to leader emergence (Berger, Cohen, & Zelditch Jr., 1972; Ensari et al., 2011; Van Dijk et al., 2017; Zaccaro et al., 2018). However, studies focused on speaking time and leader emergence seem to treat gender as a nuisance or confounding variable (Stein & Heller, 1979), often segregating groups of participants by gender to avoid its effects (e.g., Kremer & Mack, 1983) or using participants of only one gender (e.g., Jaffee & Lucas, 1969). Unfortunately, where neither of these mechanisms was used, gender composition was not always reported (e.g., Kirscht et al., 1959) or gender was reported but not analyzed (e.g., Bottger, 1984). This treatment of gender is unfortunate because gender can have a demonstrable effect on ratings: Riggio et al. (2003), for example, found that male participants were substantially overrepresented in the set of leaders chosen by the groups they observed. The influence of gender is strongly suggested by the expectation states literature (Joshi & Knight, 2015; Van Dijk et al., 2017), but unfortunately studies in this literature did not always record speaking time, or a similar variable, so the results are not directly comparable. Regardless, the evidence of Riggio et al. (2003) suggests that gender could play an important role in the relationship between speaking and leader emergence, a problematic situation for a rejection of the babble hypothesis. If gender is irrelevant to the task, as in Riggio et al. (2003), but correlates with leader emergence while controlling for speaking time, as Riggio et al. (2003) suggest, it makes it appear less likely that the content of the leader's speech was responsible for their emergence.

Hypothesis and Research Questions

Despite the long and general acceptance of the relationship of speaking time with leader emergence in the literature (e.g., Stein & Heller, 1979), as well as the demonstrable correlation

between the two constructs (Schmid Mast, 2002), it appears from this review that conflicting results and potentially problematic studies hamper clean interpretation of these findings. Part of these inconsistencies may be due to how the studies themselves were conducted. Experimental studies reviewed above used deception, controlled participant access to information about the task and other participants, did not use a control condition, or employed other design features that restrict generalizability (Lonati et al., 2018; Schmid Mast & Hall, 2004). The observational studies, on the other hand, tended to provide a richer information and interaction environment but did not account for a variety of omitted variables that in some cases have been found to have substantial influence on leader emergence in their own right. These limitations point to a need for verification.

The purpose of the present study, therefore, is to establish whether or not the main effects predicted by the literature are still evident when using (a) a richer task and interaction environment than experimental studies have typically allowed and (b) a more complete set of control variables in a study design and advanced analysis framework that supports stronger inference than observational studies have done. To be specific, this study tests the babble hypothesis by:

- Controlling for intelligence, personality, role assignment, and gender
- Using valenced tasks with salient and observable outcomes
- Contrasting different tasks and samples
- Addressing the endogenous nature of speaking time
- Including an exogenous manipulation in an observational design study
- Allowing for minimally constrained participant interaction and information discovery.

To our knowledge, an observational study with this comprehensive set of features has not been attempted before. Given the enhanced analysis framework stated above, we formulated one hypothesis:

Hypothesis: Speaking time will positively predict perceptions of leader emergence in initially leaderless groups.

This study was designed and primarily analyzed to test the above hypothesis, but also allowed us to pursue further research questions in an exploratory way. The first of these questions regarded the relative magnitude of the influence of different exogenous predictors of speaking time and therefore their relative indirect effect on leader emergence. This question seemed particularly relevant given that gender has been predicted to have both exogenous and endogenous effects on leader emergence in small group interactions (Berger et al., 1972); and variables such as intelligence and extraversion are found to have significant correlations with leader emergence in meta-analyses and reviews (Ensari et al., 2011; Zaccaro et al., 2018) but have received mixed support for influence in specific groups (e.g., Ruback & Dabbs, 1986; Ruback et al., 1984).

Research Question 1: What are the relative indirect and direct effects of intelligence, personality, role assignment, and gender in predicting leader emergence?

Several studies have reported on speaking time as a classifier that may distinguish leaders from non-leaders. The classification performance of speaking time has received recent interest from computer scientists (e.g., Jayagopi, Hung, Yeo, & Gatica-Perez, 2009; Sanchez-Cortes, Aran, Jayagopi, Schmid Mast, & Gatica-Perez, 2013), but researchers have reported classification performance data since the 1950s (e.g., Slater, 1955). Using speaking time to classify leaders has methodological advantages for leadership researchers because it may allow researchers to identify leaders in small groups without knowledge of leadership outcomes, supporting less problematic analysis of leader behaviors (Van Knippenberg & Sitkin, 2013). Comparing speaking time as a classifier with participant emergent leader votes as a separate behavior may also lead to insight with regard to group processes: participant rating behaviors are thought to be potentially biased in several non-random ways (Donaldson & Grant-Vallone, 2002; Fleenor, Smither, Atwater, Braddy, & Sturm, 2010; P. M. Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), so comparing a physical measurement with participant perceptions may be informative.

Research Question 2: What is the performance of speaking time as a classifier of leader

emergence?

Methods

Thirty-three ad hoc, heterogeneous student groups of 4–10 participants worked together as part of an ongoing study to solve a problem in a computer simulation environment using either a map-based military task or a similarly low-fidelity simulation for a business problem. A 10-minute planning session and a 60-minute gameplay session were recorded with two Canon VIXIA HF-series video cameras. Researchers made no attempt to control the groups' planning or gameplay activities other than to provide the structure described in detail below. Both simulations had objective assessments of performance outcomes recorded by the game which were salient to group members.

Sample

Student participants in this study were recruited at two different universities, referred to here as S1 (100 participants in 11 groups) and S2 (156 participants in 22 groups). Participants were heterogeneous in a variety of ways. First, students were recruited differentially: some participants were recruited from undergraduate psychology courses as part of a "participant pool"; other students were recruited from undergraduate and graduate courses in management and engineering in return for extra credit in those courses. Students were assigned to study sessions based on student scheduling convenience in all cases.

Second, participants were both cognitively and demographically diverse. Participants ranged in age from 18 to 38 with an overall mean of 20.7 years of age and significantly differed between the two locations (S1 = 23.2, S2 = 19.1, t = 13.349, df = 131.83, $p < 0.0001$). Additionally, 26% of participants spoke English as a second language (ESL), with a greater proportion of participants reporting ESL status at one university than the other ($\chi^2 = 42.341$, df = 1, $p < 0.0001$). Scores on intelligence and other assessment instruments, described below, varied within groups but did not vary significantly between groups or between the two samples.

Procedure

Each study session lasted approximately four hours. Students, who may or may not have known each other prior to beginning the session, were introduced to the study procedures, gave informed consent, and began completing psychometric instruments. Time was allotted to read the simulation's instruction manual, practice the game as an individual, and to discuss (five minutes) and practice (ten minutes) the game as a group. Participants were then given ten minutes to plan how they would complete their task and 60 minutes to attempt their task as a group. After the 60 minute gameplay phase students completed another set of instruments, were thanked for their participation, and given a debriefing form.

Task

Students participated in either a military- or business-themed simulation. Assignment to simulation was based on a set schedule (military-themed sessions were conducted earlier in the project timeline). The military-themed simulation, a game called BCT Commander (Shrapnel Games, Inc., 2018), is a map-based platform in which users attempt to complete a mission given by an in-game set of instructions and made salient to the participants by criteria given in their study materials. The computer software itself recorded in-game actions and provided the objective assessment of performance outcomes by which success or failure was determined. Participants found this game to be challenging and at times frustrating, often seeming to spend the majority of the planning phase learning how to use the interface, and spending less time working out a plan for task completion; that is, task-based planning did occur in these groups, but it did not seem to take up as much time during the planning session as working through the interface did.

The business-themed simulation was an on-line entrepreneurship simulation (Sterman, Miller, & Hsueh, 2018). Participants viewed simulated quarterly reports and adjusted in-game quantities according to relevant business decisions. As with the military-themed simulation, instructions and objective success criteria were provided to the participants in study materials and by the simulation itself. Participants appeared to find the interface more intuitive and spent more time in task-based planning than in the BCT simulation. Groups sometimes finished planning

early, spending the rest of the planning session speaking off topic. Thus, the two games were apparently perceived by the participants to be different in a way that was noticeable to an observer.

An operator was chosen at random from among the participants in both simulations. The operator was responsible for manipulating the game's user interface but the proctor emphasized to the groups that the operator was not responsible for making all of the decisions. No leader role was assigned, though in the business simulation certain roles, such as "Pricing", described in the simulation's user manual, were assigned at random.

Measures

Data were collected through several methods, including participant self-report psychometric assessments, demographic questionnaire, researcher assignment, and researcher assessment from video recordings. Details are provided below.

Survey Instruments

An attempt was made to use psychometric and demographic instruments that are thought to assess relatively stable individual traits, preceded the planning session in time, or both. Traits such as personality are generally thought to be stable enough that their value would not change during the study session (Hough, Oswald, & Ock, 2015), suggesting that a participants' observed value on these instruments could be thought of as representing an attribute of the participant at the start of the planning phase: trait variables should not change values based on group interaction. This feature is important for our analysis because it would suggest that individual difference variables are exogenous to the relationship between speaking time and leader emergence. There is some evidence that this assumption may not hold for all traits, with N. P. Podsakoff, Spoelma, Chawla, and Gabriel (2019) finding in their meta-analysis that an estimated 45% of variance in observed personality is within-person variance. Thus, while it may be the case that the latent trait assessed by a psychometric instrument does not change, the observed value obtained by the instrument may be vulnerable to omitted variable bias. The issues raised by N. P. Podsakoff et al. (2019) may weaken our use of personality variables as exogenous, but intelligence was assessed prior to group interactions and so should therefore not be influenced by these interactions.

All survey instruments were administered using the Qualtrics web-based platform (www.qualtrics.com). Intelligence was assessed with the Employee Aptitude Survey (30 items; Grimsley, Ruch, Warren, & Ford, 1985); personality was assessed with the NEO-FFI (60 items; Costa & McCrea, 1992). The demographic questionnaire included items regarding gender, age, and English as a second language (ESL) status; participants were not asked about ethnicity. Intelligence and game knowledge (described below) were assessed before the planning phase; demographic attributes and personality were assessed after participants had completed the gameplay phase.

Ten-item quizzes were developed for each simulation to assess game knowledge based on material in the instruction manuals provided to the participants. The game knowledge measure was developed as an in-house assessment of participant knowledge of the simulation task and interface after the initial familiarization phase had been completed. There were two versions of this quiz, one for each simulation. An example item from the military simulation knowledge quiz is: "What is the purpose of the Scout unit? (a) Carry soldiers, (b) Remove obstacles, (c) Survey the area for the enemy, (d) Repair tanks." An example item from the business simulation knowledge quiz is: "What type of employees could you hire? (a) Engineers and sales & admin, (b) Engineers and consultants, (c) Receptionists and engineers, (d) Engineers and scientists". The sum total score on all items on the quiz was the relevant score used in analyses; as such, typical assessments of reliability, such as Cronbach's α, may not be relevant.

Leader emergence was assessed by asking the students the following question: "We would like you to nominate an individual or multiple individuals that emerged as a leader or leaders during this planning (gameplay) phase. You can select as many as five leaders or as few as one. You will be choosing the leader or leaders based on workstations. Please take a moment to nominate the leader or leaders of the group." The leader emergence item was given to the participants twice, once after the planning phase and once after the gameplay phase. Overall leader emergence for a given phase was determined by a count of votes each participant received. The time consuming nature of coding speaking time from pre-recorded video motivated against

calculating speaking time from all 80 minutes of each sessions video. The 10-minute planning session was chosen to increase sample size while still capturing early interactions between group members that may be important for establishing social norms in ad hoc groups (Burroughs & Jaffee, 1969; Cashdan, 1998; Reynolds, 1984). Although only the planning sessions were coded for speaking time in this study, the correlation between leader emergence votes after the planning session and after the gameplay session was 0.85.

Speaking Time

The first author used the ELAN video annotation software (Max Planck Institute for Psycholinguistics, 2018) to record the beginning and end of each participant utterance during each of the 33 planning sessions that could be considered language or having clear semantic meaning. Thus, laughs were not recorded but some non-word utterances, such as apparent expressions of agreement, were. By and large the content of participant speech appeared relevant to the task, but all speech, even if apparently irrelevant, was recorded. Most participants appeared to speak to the group as a whole or to the operator, but some participants also spoke quietly to their neighbors at times—these speaking events were also recorded. The video data came from the two camcorders placed in the study session room without additional microphones. As such, not all utterances were clearly distinguishable or assignable to a specific participant, introducing error. The total time a participant spoke was recorded as total speaking time (TST) and used as the measure of participation amount in this study. A count of participant speaking turns was also conducted but because the total number of speaking turns correlated at 0.88 with TST it was not further analyzed in this study.

Analysis

Despite the within- and between-group heterogeneity discussed above, the ICC1 values for speaking time and leader emergence were both truncated at zero, suggesting that the focal variable relationship was relevant to the individual level of analysis. A two-stage least squares (2SLS) analysis, employed to rigorously test our notions and address endogeneity concerns, was therefore conducted analyzing the relationship between speaking time and leader emergence at the

individual level. Cluster-robust standard errors, clustered at the group level, were used for all regression models to account for potential non-independence and heterscedasticity (Cameron & Miller, 2015; McNeish, Stapleton, & Silverman, 2017). Regression models, including standard 2SLS, Lewbel, and maximum likelihood models, were estimated in Stata 15. All remaining statistical analyses, such as calculating proportions and tests of proportions, were conducted in R using standard repository packages. Analysis code is available along with the anonymized data set at (URL to be inserted here).

As suggested by Bass (1990), TST was seen as the endogenous predictor. The instrumental variables, also known as excluded instruments, were intelligence, the five personality variables from the NEO-FFI, and a dummy variable for whether or not the participant was assigned as the operator. Intelligence and personality are not directly observable by co-participants, but are consistently found to be associated with leader emergence (Ensari et al., 2011; Zaccaro et al., 2018). Given this task environment, participants should observe intelligence and personality through speaking time; other forms of communication are not analyzed in this study but are potential sources of information for participants (e.g., Gerpott, Lehmann-Willenbrock, Silvis, & Van Vugt, 2018). Thus, in this study intelligence and personality should influence leader emergence only through speaking time. Antonakis, Day, and Schyns (2012) proposed a general process model in which individual differences act on observed outcomes through leader behavior, and the 2SLS model considered here follows that general format.

The overall task environment called for an operator to implement group decisions in the simulation. Any of a variety of different methods could have been used to assign participants to the operator role. Particularly for the military simulation, with its more challenging interface, some sort of participant skill-based assignment may have been useful. However, random assignment of a participant to the operator role ensured that we had a good, what some might call "perfect" (Antonakis et al., 2010, p. 1103), instrumental variable: operator status was uncorrelated with other variables of interest by design, an important condition for instrumental variables (Antonakis et al., 2010; Sajons, in press; Wooldridge, 2010). Again based on Bass

(1990), we further assumed that operators would in fact speak more, but that increased speech pursuant to the operator role would not be considered directly relevant to leader emergence by the participants; we tested this assumption and report the results below.

The exogenous predictors, also known as included instruments, were age, gender, game knowledge, ESL status, group size, simulation (military vs. business), and institution (S1 vs. S2). These variables appeared more likely to have a direct effect on both speaking time and leader emergence than the excluded instruments by virtue of being more directly apparent to the participants. Furthermore, group size has been shown to have direct effects on patterns of speech that do not affect the relative relationship between speaking time and leader emergence (Reynolds, 1984). The differences between the tasks in terms of content and difficulty also appeared likely to directly affect both speaking time and leader emergence, as did the demographic differences between the two samples.

The two equations of the 2SLS model were therefore as follows:

Stage 1:

$$TST = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 GameKnowledge + \beta_4 ESL + \beta_5 GroupSize + \beta_6 Simulation + \beta_7 Institution + \beta_8 Intelligence + \beta_9 Conscientiousness + \beta_{10} Agreeableness + \beta_{11} Neuroticism + \beta_{12} Openness + \beta_{13} Extraversion + \beta_{14} OperatorStatus + e$$

Stage 2:

$$LeaderEmergence = \gamma_0 + \gamma_1 \widehat{TST} + \gamma_2 Age + \gamma_3 Gender + \gamma_4 GameKnowledge + \gamma_5 ESL + \gamma_6 GroupSize + \gamma_7 Simulation + \gamma_8 Institution + u$$

Results

Summary statistics for variables used in the 2SLS regression analysis are in Table 1 and correlations are in Table 2; all variables in both tables are at the individual level. Both the leader emergence votes and TST were positively skewed: 61.7% of participants spoke at or less than the mean TST. Furthermore, 26 participants did not speak at all with at least one member with TST = 0 in 19 out of 33 groups. The zero-order correlation between TST and leader emergence was 0.67, near the mean effect size calculated by Schmid Mast (2002). The zero-order correlation between

leader emergence and gender was 0.35, near the mean effect found by Ensari et al. (2011).

As discussed by Hough et al. (2015), we found correlations between personality variables. Neuroticism was negatively correlated with conscientiousness (r = -0.34), agreeableness (r = -0.22), and extraversion (r = -0.28). Extraversion was positively correlated with conscientiousness (r = 0.35) and agreeableness (r = 0.28). Contrary to other studies (see Ensari et al., 2011), the correlation between extraversion and leader emergence was relatively low (r = 0.11) compared to the correlation between openness to experience and leader emergence (r = 0.21).

As suggested by the ICC1 values, TST and leader emergence votes exhibited more variance within groups than between groups (Table 1). With the exception of age and ESL status, which reflect the differences between samples and the presence of a small number of older students, most variables have ICC1 values near zero. These variance patterns occur despite differences between the groups at the two different institutions that are reflected in the correlation table (Table 2). For example, S2 groups tended to be younger on average than S1 groups (r = -0.69) and older students were more likely to report ESL status (r = 0.45). Differences between S1 and S2 were not evident in the personality variables, nor in the intelligence or game quiz variables when controlling for ESL status (partial correlations were -0.01 and 0.07, respectively). There were nearly equal numbers of male and female participants; though proportions of male and female participants varied between groups, that variation was approximately uniformly distributed between 12.5% and 80% female participants (χ^2 = 2.3571, df = 32, p = 1.0). Overall, the patterns of correlations and variances suggest that groups were internally heterogeneous and that the groups differed from each other and there was a similar amount of variance within groups as there was between groups.

Model Estimation

The results of the initial 2SLS estimation are in Table 3. The hypothesized formulation had a significant F-statistic for the first stage (F = 5.5177, df = (7, 32), p = 0.0003), but it was below the F = 10 cutoff recommended by Wooldridge (2010). The model is correctly specified (Hansen's J = 5.187, p = 0.5201). There is no significant empirical evidence for endogeneity (cluster-robust Durbin-Wu-Hausman F = 0.1055, df = (1, 32), p = 0.7474).

Weak instruments have the potential to bias 2SLS coefficient estimates (Wooldridge, 2010). In this case, coefficient estimates could be biased by nearly 30% (Stock & Yogo, n.d.). There are several ways to address weak instruments. First, we could reduce the number of instruments we used. In this data, four of the five NEO FFI variables did not have substantial zero-order correlations with speaking time, and the five were correlated. In other words, there was more correlation among the personality variables than between the personality variables and speaking time—a situation that potentially contributed to the weakness of the instruments. A reduced model in this case could use two traits, intelligence and openness to experience, and the controlled variable, assignment to the operator role, as instrumental variables.

One problem with this variable reduction approach could be that, as a post hoc reduction based on examining the data and model results, we over-fit the model to this particular data. There is some reason to think this may be the case here: most reviews (e.g., Ensari et al., 2011) find that extraversion has an equivalent or stronger influence on leader emergence than openness to experience—yet that does not seem to be the case in our data. With these cautions in mind, we report the results of several models potentially indicated by these considerations in the Appendix.

A second approach to obtaining unbiased coefficient estimates would be to use the Lewbel procedure (Lewbel, 2012). Rather than reducing the number of instrumental variables in the model, the Lewbel approach is to increase their number by calculating new instrumental variables from heteroscedasticity in the model errors. Although Lewbel estimation may have desirable properties in general, such as improving estimation efficiency (Lewbel, 2012), the approach is useful in this case because it leaves the original data and hypothesized relationships in place, improving estimation using data already available. Additionally, there is substantial skew in some of the predictor variables, perhaps influencing the results of the typical 2SLS estimation procedure. We use the additional Lewbel instrumental variables for all further analyses below and report the augmented model in Table 4.

The count-based nature of the outcome variable, leader emergence votes, indicated that a Poisson regression may be more appropriate. However, an exponential model was a poor fit to

data in this case, particularly towards the higher end of the primary variables, speaking time and leader emergence. The qualitative conclusions of the model do not change with any of the eventualities described in this section, including Poisson regression. The Poisson model also is reported in the Appendix.

Model Results

The Lewbel model had strong instruments (F = 19.4248, df = (14, 32), $p < 0.0001$), empirical evidence of endogeneity (cluster-robust Durbin-Wu-Hausman F = 6.8199, df = (1, 32), p = 0.0136), and was correctly specified (Hansen's J = 13.731, p = 0.3930). Bias in 2SLS coefficient estimates was estimated to be below 10% (Stock & Yogo, n.d.). The regression coefficient on speaking time was significant (Table 4, Figure 1). Specifically, based on this model it takes an average of 39 seconds of speaking to earn another leader emergence vote. Being a male participant was associated with approximately an additional expected vote and was equivalent to about 45 seconds of speaking—nearly as much as the average speaking time for all participants across the data set. Figure 1 shows the relative effects of these two variables on leader emergence votes in a margins plot.

The coefficients on age, game knowledge quiz, and ESL status were not significant; of the group-level controls, only group size was significant. Despite differences in apparent task difficulty and ambiguity between the two simulations, the coefficient on simulation was not significant. The variable indicating institution was also not significant, suggesting that despite demographic and other differences between the institutions, those differences were not important to the relationship between speaking time and leader emergence. In other words, the relationship between speaking time and leader emergence appears to be robust in this data.

There is the possibility that participants attributed additional leader status to the assigned operator over and above the additional speaking time associated with the operator role predicted by Bass (1990). If participants viewed the operator thus, operator status should have a direct effect on leader emergence. Given that the overidentification statistic was not significant (Table 4), the data do not appear to support this conclusion. A similar interpretation is indicated for intelligence

and openness to experience.

It is important to note that the order in which variables appear in the 2SLS model respects the actual or presumed temporal order, and thus perhaps influence, of the variables in the study. Intelligence and game knowledge were both assessed before speaking time was recorded, and votes were taken immediately after the planning phase from which the speaking time data came. The demographic and personality questionnaires were issued after the game was over, but it is typically argued that personality variables are stable aspects of the individual (Hough et al. (2015); N. P. Podsakoff et al. (c.f., 2019)). Thus, the value of all variables assumed by the 2SLS model to be exogenous were actually or presumably fixed before group interactions occurred. Group interactions provided the data for assessing speaking time. We therefore find support for our hypothesis: there was a significant effect of speaking time on leader emergence given the controls used in the model and provided by the 2SLS structure itself.

Relative Effects

The range of values that the primary predictors of speaking time in this data are substantially different, making an interpretation of their relative effects difficult in the model presented above. Standardized regression coefficients can aid in such interpretation, although caution is warranted in their use (Baguley, 2009; Greenland, Maclure, Schlesselman, Poole, & Morgenstern, 1991). We use them here to compare the relative effects of variables within a model to support a post hoc analysis; the precise standardized coefficient estimates are not interpreted. To use standardized coefficients, we recast the Lewbel 2SLS model in an SEM format with comparable, but not identical, coefficient estimates (Table 5)—a further caution against overinterpretation of precise values in this post hoc analysis. The SEM model had strong instruments (F-equivalent = 26.1442), some empirical evidence for endogeneity (Wald test for correlation in disturbances: $\chi^2 = 3.83$, df = 1, $p = 0.0505$), and acceptable fit with cluster-robust errors (SRMR = 0.012, coefficient of determination [CD] = 0.602).

With the above cautions in mind, we used a Wald test to compare the standardized coefficients: the indirect effects of gender ($b = 0.0806$), operator assignment ($b = 0.1018$),

openness to experience (b = 0.0907), and intelligence (b = 0.0724) were not significantly different (χ^2 = 0.57, df = 3, p = 0.9043). The indirect effects of conscientiousness, agreeableness, neuroticism, and extraversion were not significantly different from zero. The combined direct and indirect effect of gender was b = 0.3093, about three times the effects of each of operator assignment, openness to experience, and intelligence. We tentatively conclude, therefore, that the effects of task-relevant individual differences, intelligence and openness to experience, were approximately equal to the effects of a presumably irrelevant individual difference, gender, and to an ostensibly non-leader role assignment, operator status. However, operator assignment, intelligence, and openness to experience did not have a significant direct effect: the combined effect of gender—that is, the direct effect of gender on leader emergence and the indirect effect on leader emergence through its effect on speaking time—was substantially larger than the other predictors analyzed here.

Classification Performance

Analysis of the performance of in-group maximum TST in correctly classifying the emergent leader, as rated by the plurality of co-participants, was also conducted post hoc. Specifically, classification was considered "correct" or "successful" when the participant with the maximum within-group TST also received the maximum number of votes. In this analysis, the votes co-participants cast for emergent leaders were considered the reference scores, or "ground truth", and the participants' TST the classifier to be tested. The question then becomes: how well does the decision rule, "the participant with the maximum TST is the leader," agree with the reference decision rule, "the participant with the maximum votes is the leader." This approach has been recently used with groups of size four (Jayagopi et al., 2009; Sanchez-Cortes et al., 2013) and in previous literature (see Table 6A).

This classification performance analysis may be of interest to leadership researchers generally because of its potential use as an assessment procedure in other studies: identification of leaders by speaking time may allow researchers to identify relevant leaders without knowledge of formal group structure or behavioral outcomes—features recommended for improved inference by

Van Knippenberg and Sitkin (2013). However, the classification performance results also demonstrate a pronounced gender bias more clearly than in the regression results above.

Previous work has found TST to be a very good classifier for social dominance in small groups (size = 4) engaged in leaderless group discussions, and as good or nearly as good as more complicated audiovisual features (Jayagopi et al., 2009; Sanchez-Cortes et al., 2013). The present study found a similar raw classification performance, 70% success, as previous studies had (see Table 6A). The groups in this study included some with similar sizes as those in previous studies, but also larger groups.

The speaking time literature reviewed above has tended to rely on participant attributions of leadership as the objective criterion by which speaking time, as an assessment procedure, should be judged. However, there are several reasons to suspect there may be systematic error in participant attributions of this kind (Donaldson & Grant-Vallone, 2002; Fleenor et al., 2010; P. M. Podsakoff et al., 2003). Among the groups whose data is analyzed in this study, there was a single consensus leader—that is, one and only one participant who received a leadership vote from each group member—in 12 of 33 groups. In two other groups two participants both received votes from every member. In the remaining groups, 58% of those analyzed, no single consensus leader emerged, suggesting that there may be some error, and perhaps non-trivial structural differences in relationships, in the leader emergence votes. If classification success includes the maximum TST participant being ranked either first or second by co-participant vote, then classification performance goes up to 91% (Table 6B). However, classification performance differs strongly by gender. Using the stricter definition of performance, female participants were classified correctly 29% of the time but 100% of the time using the looser definition, whereas performance for male participants was 81% and 88%, respectively. These differences may suggest that notable differences in leader emergence votes may be attributable to the gender of the rating target rather than to actual behavioral differences.

Discussion

The purpose of this study was to more rigorously test conclusions from decades of prior research on the relationship between speaking time and leader emergence. In support of this goal, this study combined the relative ecological validity of observation of problem-solving groups with stronger inference supported by 2SLS analysis with a more comprehensive set of covariates. We find that our hypothesis is supported: speaking time retains a substantial effect on leader emergence even when controlling for a variety of other variables also known to correlate with leader emergence.

Contrary to several other studies in this literature, we did find a significant influence of personality and cognitive ability variables on leader emergence: the effect in this data is indirect, through a behavior, speaking, as predicted by Antonakis et al. (2012). It is not immediately clear why openness to experience and not extraversion is the significant personality variable, though Colbert, Judge, Choi, and Wang (2012) had similar results and other studies have also failed to find a significant effect of extraversion (Ruback et al., 1984) or interpreted speaking time as "talkativeness" and therefore representing extraversion itself (Ensari et al., 2011); the nature of the task or the sample may also be involved in these differences.

Implications

As noted above, this study both confirms and extends prior research on the influence of speaking time in small groups. The accumulated evidence suggests that future studies of leader emergence could profitably include assessments of speaking time, whether to account for its variance in an appropriate analysis framework (Gerpott et al., 2018; Morris & Hackman, 1969) or as a variable of interest in its own right. Evidence from this study and others (e.g., Riggio et al., 2003) suggests that concurrent consideration of gender may be important for proper inference.

A second implication of this study relates to what has been called expectation states theory (EST) or status characteristics theory (Berger et al., 1972; Joshi & Knight, 2015; Ridgeway & Erickson, 2000; Van Dijk et al., 2017). EST predicts that in groups that vary with respect to external status markers, such as, potentially, gender and age, these status markers will become

predictors of within-group status through both direct, as indicated by differential voting behavior, and indirect pathways, such as through speaking time. The groups in this study were heterogeneous in a variety of ways, but all groups contained both male and female participants. The results above indicate that gender influences both the production of speaking time and the interpretation of speaking time, consistent with EST. It is probably not important that the coefficient on age was non-significant: many groups did not contain substantial variation in age, which may have reduced the observable impact of any effect due to age. However, EST is clear that in heterogeneous groups, task-related cues should not be important (Berger et al., 1972; Ridgeway & Erickson, 2000). In this study intelligence, openness to experience, and role assignment affect the production of speaking time as much as gender does—a finding that seems inconsistent with the predictions of EST.

Game knowledge was not significant in this model, which is also consistent with EST. However, game knowledge also had a low zero order correlation with planning phase votes. The game knowledge score had a higher zero order correlation with gameplay phase votes, suggesting that perhaps, as others have noted (Gerpott et al., 2019; Kalish & Luria, 2016) different features are more important for leader emergence at different times. If speaking time is autoregressive within a given group and situation (Burroughs & Jaffee, 1969; Cashdan, 1998; Reynolds, 1984) it seems plausible that a similar model as tested here but fit to the gameplay phase (using the later in time gameplay phase speaking time and vote data) might find a significant coefficient on game knowledge—that finding, although conjecture at this time, would not be consistent with EST. Future work could use a data collection method that is easier to code for speaking time and more isomorphic tasks to better test these ideas.

Third, this study seems to support the growing interest in the use of countable features in organizational behavior (Matusik et al., 2018): the use of physical measurements of individual behavior and interactions between individuals, such as duration of speech, counts of eye fixations (Gerpott et al., 2018), and vocal pitch (Cheng, Tracy, Ho, & Henrich, 2016), may have consistent and important relationships with relevant group processes and outcomes. There is some evidence

that these effects may be more widespread and more integral to group process than has been previously considered. For example, both Riggio et al. (2003) and the present study found a high correlation (≈ 0.90) between speaking time and speaking turns. Cashdan (1998), Reynolds (1984), and Ruback et al. (1984) all found consistent patterns in the distribution of speaking time, pauses, and turns across several structural contrasts, including group size, presence of strangers, and task types. This consistency seems relevant in light of the Woolley, Chabris, Pentland, Hashmi, and Malone (2010) findings on collective intelligence: Woolley et al. (2010) found that increased variance in within-group speaking turns decreased average group performance outcomes across a range of tasks. Unfortunately, it is difficult to tell from Woolley et al. (2010) variance in speaking time also had a correlation with group performance outcomes, but given the tight correlation other studies have found, it seems reasonable to suspect that increased variance in speaking time was also present. Because increased speaking time appears to lead to leader emergence, higher variance in speaking time in an initially leaderless group may mean more consensus among other group members as to who the leader is—a condition that has been associated with increased group member satisfaction (Berkowitz, 1953). In other words, a more even distribution of speaking time may indicate, or even facilitate, an increase in average group performance outcomes, but may also decrease satisfaction of group members, another important consideration. Although these are somewhat isolated studies without direct comparison to each other, this could be a promising area of future research.

Finally, although there is no necessary conflict between speech quantity and quality—indeed, these two constructs should be somewhat correlated—the evidence in this study appears to support the very babble hypothesis Bass (1990) sought to refute. Task-relevant aspects of cognition (intelligence), personality (openness to experience), and role (operator assignment) did have an influence on speaking time, but gender had an equivalent influence on speaking time and had an additional, substantial impact on the ratings themselves. Differences between institutions, though significant in some dimensions, did not significantly influence this relationship. Task differences, which to an observer seemed apparent in the behavior of the

participants, were also not significant. Increased speaking time was attributable to the operator but did not have a direct effect on leader emergence. Game knowledge was not a significant predictor of either speaking or leader emergence, suggesting a reduced role for task-specific knowledge in leader emergence in this study. Taken together, these findings are most consistent with a babble hypothesis of leader emergence: participant features that may reflect speech quality appear reduced in importance when compared with features that are correlated with speech quantity but may be uncorrelated with quality.

Limitations

There are at least three important limitations of our study that suggest against over-interpretation. First, our study includes no direct assessment of the quality of participant speech. Several studies suggest that several standard behavioral assessment methods may be inadequate to judge speech quality, including some of the earliest work in this literature (Bass, 1949; Juola, 1957; Morris & Hackman, 1969; Riggio et al., 2003). However, without a direct assessment of quality it is not possible to reject the hypothesis that it is quality, or both quality and quantity, that determine leader emergence. To fully reject—or convincingly fail to reject—a babble hypothesis of leadership, future studies should systematically assess speech quality directly through more sophisticated methods than have been tried in the past.

While this study does not address speech content per se, the study design does mitigate this limitation to a certain extent. For example, by including variables in the regression model that should correlate with speech quality we estimate the average effect of speaking time, as an endogenous variable, on leader emergence while holding these other variables constant. Thus, we can conclude that for a given set of participants who are equivalent in intelligence and other quality-related variables, speaking time has a significant expected effect on leader emergence. Understood this way, while we can not reject an influence of quality, we can suggest that, holding potential correlates of speech quality constant, speaking time does have a significant effect.

The game knowledge variable itself provides additional information. As noted earlier, some of the initial difficulties participants seemed to face were tied to the basic concepts of the

game and how to use the interface appropriately; this appeared to be particularly noticeable for the military simulation. Based on the content of the items, the game knowledge instrument should have captured this type of knowledge, however imperfectly. Particularly for the military simulation, if the contribution quality of participant speech was important for leader emergence there should have been a correlation between the game knowledge variable and leader emergence above and beyond that accounted for by speaking time—that does not seem to have been the case.

A second limitation is the lack of analysis of directed behaviors between individuals. EST in particular considers pairwise status assessments of status to be critical to group sociometry, and the lack of consensus in many groups as to the emergent leader suggests that there may be non-trivial pairwise interactions to capture. Observation of the groups in our study suggested that most speech was directed to the group in some way, but many dyadic interactions may have been missed by our coding methodology. A better test of EST in a similar study would require solving the problem of how to obtain and analyze pairwise interactions between group members alongside an analysis of speaking time and its derivations in an ecologically valid setting. Specifically, a study that controlled task and status cues alongside diarized speaking in a rich information environment could provide this important test in the future.

Finally, there are limits to the ecological validity of this study. For example, some student participants had relevant business training through their degree programs, but few if any participants could be said to have relevant expertise for either simulation; this condition was most noticeable for the military simulation. Furthermore, whereas the simulations were valenced tasks, the simulation outcomes were not consequential to the participants.

Conclusions

The data presented in this study suggest that speaking time predicts leader emergence despite variation in task-related demands and group composition, among other variables. Corroborating previous studies, this study provides evidence of a potential causal relationship between speaking time, an endogenous variable, and leader emergence, an outcome, through the use of additional control variables and an exogenously manipulated variable, operator assignment,

in an observational setting. Furthermore, the data suggest that although there is some influence of task-relevant variables on leader emergence, when operationalized as co-participant votes, gender, a presumably task-irrelevant variable, has a more substantial influence. These findings are consistent with a babble hypothesis of leader emergence.

References

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120.

Antonakis, J., Day, D. V., & Schyns, B. (2012). Leadership and individual differences: At the cusp of a renaissance. *The Leadership Quarterly*, *23*(4), 643–650.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617.

Bales, R. F. (1950). *Interaction process analysis; a method for the study of small groups.* Addison-Wesley.

Bass, B. M. (1949). An analysis of the leaderless group discussion. *Journal of Applied Psychology*, *33*(6), 527.

Bass, B. M. (1954). The leaderless group discussion. *Psychological Bulletin*, *51*(5), 465.

Bass, B. M. (1990). *Bass & Stogdill's handbook of leadership: Theory, research, and managerial applications* (3rd ed.). New York: Free Press.

Bavelas, A., Hastorf, A. H., Gross, A. E., & Kite, W. R. (1965). Experiments on the alteration of group structure. *Journal of Experimental Social Psychology*, *1*, 55–70.

Berger, J., Cohen, B. P., & Zelditch Jr., M. (1972). Status characteristics and social interaction. *American Sociological Review*, *37*(3), 241–255.

Berkowitz, L. (1953). Sharing leadership in small, decision-making groups. *The Journal of Abnormal and Social Psychology*, *48*(2), 231–238.

Blevins, D. P., Tsang, E. W., & Spain, S. M. (2015). Count-based research in management: Suggestions for improvement. *Organizational Research Methods*, *18*(1), 47–69.

Bottger, P. C. (1984). Expertise and air time as bases of actual and perceived influence in problem-solving groups. *Journal of Applied Psychology*, *69*(2), 214.

Burroughs, W. A., & Jaffee, C. L. (1969). Verbal participation and leadership voting behavior in a leaderless group discussion. *The Psychological Record*, *19*(4), 605–610.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference.

807 *Journal of Human Resources*, *50*(2), 317–372.

808 Cashdan, E. (1998). Smiles, speech, and body posture: How women and men display sociometric
809 status and power. *Journal of Nonverbal Behavior*, *22*(4), 209–228.

810 Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals
811 of dominance predict emergent social rank in humans. *Journal of Experimental*
812 *Psychology: General*, *145*(5), 536.

813 Colbert, A. E., Judge, T. A., Choi, D., & Wang, G. (2012). Assessing the trait theory of
814 leadership using self and observer ratings of personality: The mediating role of
815 contributions to group success. *The Leadership Quarterly*, *23*(4), 670–685.

816 Costa, P. T., & McCrea, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO*
817 *five-factor inventory (NEO-FFI)*. Psychological Assessment Resources.

818 Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational
819 behavior research. *Journal of Business and Psychology*, *17*(2), 245–260.

820 Ensari, N., Riggio, R. E., Christian, J., & Carslaw, G. (2011). Who emerges as a leader?
821 meta-analyses of individual differences as predictors of leadership emergence. *Personality*
822 *and Individual Differences*, *51*(4), 532–536.

823 Fişek, M. H., Berger, J., & Norman, R. Z. (2005). Status cues and the formation of expectations.
824 *Social Science Research*, *34*(1), 80–102.

825 Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. E. (2010). Self-other
826 rating agreement in leadership: A review. *The Leadership Quarterly*, *21*(6), 1005–1034.

827 Gerpott, F. H., Lehmann-Willenbrock, N., Silvis, J. D., & Van Vugt, M. (2018). In the eye of the
828 beholder? an eye-tracking experiment on emergent leadership in team interactions. *The*
829 *Leadership Quarterly*, *29*(4), 523–532.

830 Gerpott, F. H., Lehmann-Willenbrock, N., Voelpel, S. C., & van Vugt, M. (2019). It's not just
831 what is said, but when it's said: A temporal account of verbal behaviors and emergent
832 leadership in self-managed teams. *Academy of Management Journal*, *62*(3), 717–738.

833 Gintner, G., & Lindskold, S. (1975). Rate of participation and expertise as factors influencing

leader choice. *Journal of Personality and Social Psychology*, *32*(6), 1085.

Greenland, S., Maclure, M., Schlesselman, J. J., Poole, C., & Morgenstern, H. (1991). Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology*, 387–392.

Grimsley, G., Ruch, F., Warren, N., & Ford, J. (1985). *Manual for the employee attitude survey, test of verbal reasoning*. Glendale, CA: Psychological Services.

Gustafson, D. P., & Harrell, T. W. (1970). A comparison of role differentiation in several situations. *Organizational Behavior and Human Performance*, *5*(3), 299–312.

Hough, L. M., Oswald, F. L., & Ock, J. (2015). Beyond the big five: New directions for personality research and practice in organizations. *The Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 183–209.

Jaffee, C. L., & Lucas, R. L. (1969). Effects of rates of talking and correctness of decisions on leader choice in small groups. *The Journal of Social Psychology*, *79*(2), 247–254.

Jayagopi, D. B., Hung, H., Yeo, C., & Gatica-Perez, D. (2009). Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(3), 501–513.

Joshi, A., & Knight, A. P. (2015). Who defers to whom and why? dual pathways linking demographic differences and dyadic deference to team effectiveness. *Academy of Management Journal*, *58*(1), 59–84.

Juola, A. E. (1957). Leaderless group discussion ratings: What do they measure? *Educational and Psychological Measurement*, *17*(4), 499–509.

Kalish, Y., & Luria, G. (2016). Leadership emergence over time in short-lived groups: Integrating expectations states theory with temporal person-perception and self-serving bias. *Journal of Applied Psychology*, *101*(10), 1474.

Kirscht, J. P., Lodahl, T. M., & Haire, M. (1959). Some factors in the selection of leaders by members of small groups. *The Journal of Abnormal and Social Psychology*, *58*(3), 406.

Kremer, J. M., & Mack, D. (1983). Pre-emptive game behaviour and the emergence of leadership.

British Journal of Social Psychology, *22*(1), 19–26.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, *30*(1), 67–80.

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & Frost, A. G. (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology*, *69*(5), 877.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*, 19–40.

Matusik, J. G., Heidl, R., Hollenbeck, J. R., Yu, A., Lee, H. W., & Howe, M. (2018). Wearable bluetooth sensors for capturing relational variables and temporal variability in relationships: A construct validation study. *Journal of Applied Psychology*, *104*(3), 357–387.

Max Planck Institute for Psycholinguistics. (2018, August). *ELAN (version 5.3)* [Computer Software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from `https://tla.mpi.nl/tools/tla-tools/elan/`

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140.

Morris, C. G., & Hackman, J. R. (1969). Behavioral correlates of perceived leadership. *Journal of Personality and Social Psychology*, *13*(4), 350.

Mullen, B., Salas, E., & Driskell, J. E. (1989). Salience, motivation, and artifact as contributions to the relation between participation rate and leadership. *Journal of Experimental Social Psychology*, *25*(6), 545–559.

Norfleet, B. (1948). Interpersonal relations and group productivity. *Journal of Social Issues*.

Podsakoff, N. P., Spoelma, T. M., Chawla, N., & Gabriel, A. S. (2019). What predicts within-person variance in applied psychology constructs? an empirical examination. *Journal of Applied Psychology*, *104*(6), 727–754.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method

888 biases in behavioral research: A critical review of the literature and recommended
889 remedies. *Journal of Applied Psychology*, *88*(5), 879.
890 Reynolds, P. D. (1984). Leaders never quit: Talking, silence, and influence in interpersonal
891 groups. *Small Group Behavior*, *15*(3), 404–413.
892 Ridgeway, C. L., & Erickson, K. G. (2000). Creating and spreading status beliefs. *American*
893 *Journal of Sociology*, *106*(3), 579–615.
894 Riecken, H. W. (1958). The effect of talkativeness on ability to influence group solutions of
895 problems. *Sociometry*, *21*(4), 309–321.
896 Riggio, R. E., Riggio, H. R., Salinas, C., & Cole, E. J. (2003). The role of social and emotional
897 communication skills in leader emergence and effectiveness. *Group Dynamics: Theory,*
898 *Research, and Practice*, *7*(2), 83.
899 Ruback, R. B., & Dabbs, J. M. (1986). Talkativeness and verbal aptitude: Perception and reality.
900 *Bulletin of the Psychonomic Society*, *24*(6), 423–426.
901 Ruback, R. B., Dabbs Jr, J. M., & Hopper, C. H. (1984). The process of brainstorming: An
902 analysis with individual and group vocal parameters. *Journal of Personality and Social*
903 *Psychology*, *47*(3), 558.
904 Sajons, G. (in press). Estimating the causal effect of measured endogenous variables: A tutorial
905 on the experimental instrumental variable approach. *The Leadership Quarterly*.
906 Sanchez-Cortes, D., Aran, O., Jayagopi, D. B., Schmid Mast, M., & Gatica-Perez, D. (2013).
907 Emergent leaders through looking and speaking: from audio-visual data to multimodal
908 recognition. *Journal on Multimodal User Interfaces*, *7*(1-2), 39–53.
909 Schmid Mast, M., & Hall, J. A. (2004). Who is the boss and who is not? accuracy of judging
910 status. *Journal of Nonverbal Behavior*, *28*(3), 145–165.
911 Schmid Mast, M. (2002). Dominance as expressed and inferred through speaking time. *Human*
912 *Communication Research*, *28*(3), 420–450.
913 Shrapnel Games, Inc. (2018, June). *BCT Commander.* Retrieved from
914 `http://www.shrapnelgames.com/ProSIM/BCT/BCT_page.html`

Slater, P. E. (1955). Role differentiation in small groups. *American Sociological Review*, *20*(3), 300–310.

Sorrentino, R. M., & Boutillier, R. G. (1975). The effect of quantity and quality of verbal interaction on ratings of leadership ability. *Journal of Experimental Social Psychology*, *11*(5), 403–411.

Stein, R. T., & Heller, T. (1979). An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology*, *37*(11), 1993.

Sterman, J., Miller, D., & Hsueh, J. (2018, June). *Cleanstart: Simulating a clean energy startup.* Retrieved from `https://mitsloan.mit.edu/LearningEdge/simulations/\cleanstart/Pages/default.aspx`

Stock, J. H., & Yogo, M. (n.d.). Testing for weak instruments in linear iv regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of thomas rothenberg* (pp. 80–108). Cambridge: Cambridge University Press.

Van Dijk, H., Meyer, B., Van Engen, M., & Loyd, D. L. (2017). Microdynamics in diverse teams: A review and integration of the diversity and stereotyping literatures. *Academy of Management Annals*, *11*(1), 517–557.

Van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: Back to the drawing board? *Academy of Management Annals*, *7*(1), 1–60.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, *330*(6004), 686–688.

Zaccaro, S. J., Green, J. P., Dubrow, S., & Kolze, M. (2018). Leader individual differences, situational parameters, and leadership outcomes: A comprehensive review and integration.

The Leadership Quarterly, *29*, 2–43.

Appendix

Although speaking time is theoretically endogenous to the process of leader emergence (Bass, 1990), the original model formulation used in this study had relatively weak instruments and no empirical evidence of endogeneity. As described above, there are several potential post hoc corrections that could serve to improve estimates. The main concern is the lack of strong instruments: instrumental variables that are not correlated enough with the first stage outcome variable may substantially bias coefficient estimates (Stock & Yogo, n.d.; Wooldridge, 2010). Although individual coefficient estimates vary, the basic qualitative conclusions from each model reported below are the same as the conclusions from the Lewbel model in the main text.

There are at least two ways to try to correct for weak instruments in this study. One choice, used in the main text, is to follow Lewbel (2012) and calculate new instrumental variables from heteroscedasticity in model errors. A different approach would be to select only the strongest instrumental variables to use in a reduced model. A model reduction approach may be indicated in this study because of the personality variables: although they are of theoretical interest, in this study none of these variables is a particularly strong covariate of speaking time and there are correlations among the personality variables that are higher than the zero-order correlation of any personality variable with speaking time. Correlations among personality variables are a known concern (Hough et al., 2015). Although studies using the five factor model (Costa & McCrea, 1992) typically use all factors together, recent emphasis on personality facets and personality traits not addressed in the five factor model (Hough et al., 2015), as well as previous work in the speaking time literature (e.g., Riggio et al., 2003), suggest that personality variables may be analyzed without the rest of the now-standard five factors.

Given this conceptual framework, we chose a reduced model that retained intelligence, operator assignment, and openness to experience as instrumental variables. Model estimation results are reported in Table A1. In this formulation, the coefficient on speaking time is about the same as the Lewbel model and gender is about 9% lower. The standard errors the speaking time coefficients are about the same but the coefficient for gender is about 4% larger in the reduced

model.

Another way to address the weakness of the personality variables as instrumental variables is to remove them from the set of excluded variables, but retain them in the model as included variables. This approach would allow the model to control for personality, retaining all originally hypothesized variables, without using these variables as instrumental variables. The results of this model are presented in Table A2.

One potential advantage of a reduced model is to simplify data collection and analysis. Personality assessment procedures can be time consuming or distracting to participants and are not always available. Removing the personality variables entirely does not change the qualitative conclusions or test results of the reduced model (Table A3), suggesting that a model with two instrumental variables, operator status and intelligence, was largely similar to the model including the personality variables. In the future, it may be sufficient to have a general cognitive ability assessment and an experimentally introduced variable available to use as excluded instruments.

Models with count-based outcome variables are often fit with regressions using a Poisson-like distribution assumption (Blevins, Tsang, & Spain, 2015). Both the leader emergence votes and TST had Poisson-like distributions in this data. Although the exponential curve fit by Poisson regression methods was a poor fit to data at the high end of both variables, we estimated a Poisson model, using the Lewbel instruments for comparability, as a robustness check. As with the reduced model discussed above, the qualitative conclusions do not change with this formulation but the precise coefficient estimates do vary. The results are reported in Table A4.

Table 1

Summary Statistics

		Mean	SD	Min	Max	ICC1
1	Leader Emergence	2.61	2.63	0.00	10	0*
2	TST	48.16	54.51	0.00	276.25	0*
3	Gender	0.5	0.5	0.00	1	0.01
4	Age	20.7	2.93	18	38	0.52
5	Game Knowledge	7.34	1.82	2	10	0.08
6	ESL	0.26	0.44	0.00	1	0.22
7	Group Size	8.2	1.68	4	10	NA
8	Simulation	0.32	0.47	0.00	1	NA
9	Institution	0.61	0.49	0.00	1	NA
10	Intelligence	24.11	6.59	6	42	0.08
11	Operator Status	0.13	0.34	0.00	1	0*
12	Conscientiousness	3.66	0.53	2.17	4.92	0.00
13	Agreeableness	3.5	0.52	2.17	4.75	0*
14	Neuroticism	2.87	0.66	1.17	4.67	0*
15	Openness	3.25	0.54	1.83	4.75	0.06
16	Extraversion	3.54	0.53	2.08	4.92	0.03

Note: Means, standard deviations, ranges, and ICC1 values of all variables used in this study. Gender: female participant = 0, male participant = 1; English is a second language (ESL): no = 0, yes = 1; Operator Status: no = 0, yes = 1; Simulation: BCT = 0, CleanStart = 1; Institution: S1 = 0, S2 = 1. Total speaking time (TST) is expressed in seconds. The entry, "0*", indicates that an ICC1 value has been truncated at zero.

Table 2

Correlations

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Leader Emergence																
2	TST	0.67															
3	Gender	0.35	0.24														
4	Age	-0.04	-0.09	0.05													
5	Game Knowledge	0.20	0.19	0.09	-0.17												
6	ESL	-0.18	-0.21	-0.07	0.45	-0.26											
7	Group Size	0.10	-0.23	-0.02	0.39	-0.07	0.31										
8	Simulation	0.05	-0.01	-0.04	-0.31	0.31	-0.28	-0.21									
9	Institution	-0.01	0.08	-0.03	-0.69	0.18	-0.42	-0.58	0.55								
10	Intelligence	0.18	0.25	0.10	-0.18	0.31	-0.31	-0.04	0.11	0.13	**0.77**						
11	Operator Status	0.12	0.25	0.01	0.03	0.01	0.04	-0.10	0.01	0.04	0.02						
12	Conscientiousness	0.06	0.01	-0.16	0.02	0.12	-0.12	0.06	0.05	0.00	0.06	-0.05	**0.78**				
13	Agreeableness	-0.06	-0.02	-0.26	-0.08	0.06	-0.19	0.03	-0.02	0.02	0.10	-0.09	0.17	**0.74**			
14	Neuroticism	-0.08	-0.07	-0.16	-0.05	-0.07	0.15	-0.10	0.00	0.05	-0.05	-0.07	-0.34	-0.22	**0.81**		
15	Openness	0.21	0.32	0.08	0.11	0.11	-0.04	-0.05	-0.04	-0.10	0.22	-0.01	0.00	0.06	0.00	**0.71**	
16	Extraversion	0.11	0.09	0.00	-0.13	-0.01	-0.15	-0.04	0.00	0.07	0.07	0.08	0.35	0.28	-0.28	-0.08	**0.76**

Note: Correlations between all variables used in this study. Gender: female participant = 0, male participant = 1; English is a second language (ESL): no = 0, yes = 1; Operator Status: no = 0, yes = 1; Simulation: BCT = 0, CleanStart = 1; Institution: S1 = 0, S2 = 1. Total speaking time (TST) is expressed in seconds; Cronbach's α is in bold.

Table 3

Initial Model

First Stage Total Speaking Time	*b*	SE	*p*
Intelligence	0.881	0.358	0.015
Operator Status	37.245	8.73	< 0.001
Conscientiousness	2.088	6.372	0.743
Agreeableness	-3.414	8.528	0.689
Neuroticism	-0.06	5.658	0.992
Openness	26.855	6.589	< 0.001
Extraversion	7.125	7.406	0.337
Gender	19.324	8.224	0.02
Age	-0.866	1.128	0.444
Game Knowledge	2.815	1.673	0.094
ESL	-14.442	8.299	0.083
Group Size	-5.863	1.924	0.003
Simulation	-9.564	6.853	0.164
Institution	-8.673	7.66	0.259
Constant	-38.296	56.478	0.498
F = 5.5177, df = (7, 32), p = 0.0003			
Second Stage Leader Emergence	*b*	SE	*p*
$\widehat{TST}$	0.031	0.005	< 0.001
Gender	1.035	0.223	< 0.001
Age	-0.024	0.047	0.613
Game Knowledge	0.055	0.056	0.321
ESL	-0.385	0.311	0.217
Group Size	0.52	0.073	< 0.001
Simulation	0.34	0.18	0.059
Institution	0.319	0.305	0.295
Constant	-3.795	1.106	0.001
R^2 = 0.5776			
DWH F = 0.1055, df = (1, 32), p = 0.7474			
Hansen's J = 5.187, df = 6, p = 0.5201			

Note: First and second stage results and model performance criteria for the initial model. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test. Standard errors are cluster-robust at the group level.

Table 4

Lewbel Estimation Model

First Stage Total Speaking Time	*b*	SE	*p*
Intelligence	1.054	0.421	0.013
Operator Status	35.878	9.735	< 0.001
Conscientiousness	2.504	6.021	0.678
Agreeableness	-6.402	6.529	0.328
Neuroticism	1.804	5.036	0.72
Openness	20.553	5.138	< 0.001
Extraversion	5.892	4.772	0.218
Gender	18.25	7.989	0.021
Age	-0.405	0.961	0.674
Game Knowledge	1.551	1.729	0.377
ESL	-15.394	7.585	0.044
Group Size	-5.684	1.875	0.003
Simulation	-9.605	7.057	0.175
Institution	-7.489	9.242	0.419
Constant	-15.339	60.186	0.799
F = 19.4248, df = (14, 32), p < 0.0001			
Second Stage Leader Emergence	*b*	SE	*p*
$\widehat{TST}$	0.026	0.004	< 0.001
Gender	1.152	0.185	< 0.001
Age	-0.024	0.047	0.607
Game Knowledge	0.08	0.044	0.072
ESL	-0.483	0.308	0.117
Group Size	0.479	0.066	< 0.001
Simulation	0.285	0.18	0.114
Institution	0.258	0.291	0.375
Constant	-3.353	1.169	0.004
R^2 = 0.5611			
DWH F = 6.8199, df = (1, 32), p = 0.0136			
Hansen's J = 13.731, df = 13, p = 0.3930			

Note: First and second stage results and model performance criteria for the model estimated with the Lewbel (2012) procedure. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test. Standard errors are cluster-robust at the group level.

Table 5

Maximum Likelihood Estimation

First Stage Total Speaking Time	b	SE	p
Intelligence	1.221	0.428	0.004
Operator Status	33.679	10.578	0.001
Conscientiousness	0.243	5.288	0.963
Agreeableness	-5.214	6.091	0.392
Neuroticism	-0.591	4.746	0.901
Openness	18.749	4.935	< 0.001
Extraversion	3.286	4.526	0.468
Gender	17.889	7.381	0.015
Age	-0.411	0.936	0.66
Game Knowledge	1.349	1.715	0.432
ESL	-14.315	7.311	0.05
Group Size	-6.102	1.85	0.001
Simulation	-10.084	6.957	0.147
Institution	-7.842	8.825	0.374
Constant	12.169	56.718	0.83
F equivalent = 26.1441			
Second Stage Leader Emergence	*b*	SE	*p*
$\widehat{TST}$ 0.024	0.006	< 0.001	
Gender	1.198	0.186	< 0.001
Age	-0.025	0.048	0.61
Game Knowledge	0.09	0.047	0.065
ESL	-0.522	0.342	0.127
Group Size	0.463	0.073	< 0.001
Simulation	0.263	0.204	0.197
Institution	0.234	0.29	0.419
Constant	-3.177	1.217	0.009
CD = 0.602			
Wald χ^2 = 3.83, df = 1, p = 0.0505			
SRMR = 0.012			

Note: First and second stage results and model performance criteria for the model estimated by maximum likelihood. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test, SRMR: standardized root mean square residual, CD: coefficient of determination. Standard errors are cluster-robust at the group level. The endogeneity test for the SEM model was the Wald χ^2 test for the correlation of disturbances.

Table 6

Classification Performance

A	Comparison with Other Studies		
		% Correct	Group Size
	Slater (1955)	55	3–7
	Kirscht et al. (1959)	63	3
	Gustafson and Harrell (1970)	77 (1966), 42 (1967)	5
	Riggio et al. (2003)*	39	5–6
	Jayagopi et al. (2009)	77	4
	Sanchez-Cortes et al. (2013)	55	4
	This Study	70	4–10
B	Classification Performance by Gender		
		Ranked 1st	1st or 2nd
	Male and Female Participants	70	91
	Female Participants Only	29	100
	Male Participants Only	81	88

Note: Note: Performance of TST as a classifier of leader emergence as compared with similar studies (A) and in this study when separated by gender and whether being ranked second is included in the definition of leader emergence (B). *Riggio et al. (2003) combined their extraversion variable with speaking time in assessing classification performance.

Table A1

Reduced 2SLS Model

First Stage Total Speaking Time	b	SE	p
Intelligence	0.901	0.363	0.014
Operator Status	38.479	9.259	< 0.001
Openness	26.069	6.604	< 0.001
Gender	19.938	6.527	0.003
Age	-0.949	1.088	0.384
Game Knowledge	2.741	1.629	0.094
ESL	-15.105	7.81	0.054
Group Size	-5.846	1.798	0.001
Simulation	-9.568	6.776	0.159
Institution	-8.833	7.559	0.244
Constant	-13.495	39.313	0.732
F = 10.7768, df = (3, 32), p < 0.0001			
Second Stage *Leader Emergence*	b	SE	p
$\widehat{TST}$	0.029	0.005	< 0.001
Gender	1.075	0.224	< 0.001
Age	-0.024	0.047	0.61
Game Knowledge	0.064	0.056	0.256
ESL	-0.418	0.316	0.186
Group Size	0.506	0.072	< 0.001
Simulation	0.321	0.177	0.07
Institution	0.299	0.301	0.321
Constant	-3.644	1.112	0.001
R^2 = 0.5742			
DWH F = 0.4356, df = (1, 32), p = 0.5140			
Hansen's J = 1.286, df = 2, p = 0.5258			

Note: Model results for linear two-stage least squares models with post hoc reduction in instrumental variables. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test. Standard errors are cluster-robust at the group level.

Table A2

Reduced 2SLS Model: Personality Variables as Included Instruments

First Stage Total Speaking Time	*b*	SE	*p*
Intelligence	0.881	0.358	0.015
Operator Status	37.245	8.731	0.000
Openness	26.855	6.589	0.000
Gender	19.324	8.224	0.020
Age	-0.866	1.128	0.444
Game Knowledge	2.815	1.673	0.094
ESL	-14.442	8.299	0.083
Group Size	-5.863	1.924	0.003
Simulation	-9.565	6.853	0.164
Institution	-8.673	7.660	0.259
Conscientiousness	2.088	6.372	0.743
Agreeableness	-3.414	8.528	0.689
Neuroticism	-0.060	5.658	0.992
Extraversion	7.125	7.406	0.337
Constant	-38.296	56.478	0.498
F = 11.2632, df = (3, 32), p < 0.001			
Second Stage Leader Emergence	*b*	SE	*p*
$\widehat{TST}$	0.030	0.005	0.000
Gender	1.118	0.243	0.000
Age	-0.018	0.046	0.688
Game Knowledge	0.061	0.057	0.279
ESL	-0.440	0.327	0.178
Group Size	0.518	0.069	0.000
Simulation	0.325	0.173	0.060
Institution	0.299	0.301	0.321
Conscientiousness	0.272	0.216	0.208
Agreeableness	-0.160	0.220	0.467
Neuroticism	0.272	0.207	0.189
Extraversion	0.283	0.255	0.267
Constant	-6.107	1.893	0.001
R^2 = 0.5841			
DWH F = 0.2909, df = (1, 32), p = 0.5933			
Hansen's J = 1.526, df = 2, p = 0.4662			

Note: Model results for linear two-stage least squares models with post hoc reduction in instrumental variables. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test. Standard errors are cluster-robust at the group level.

Table A3

Reduced 2SLS Model: Openness Removed from the Model

First Stage			
Total Speaking Time	b	SE	p
First Stage			
Total Speaking Time	b	SE	p
Intelligence	1.393	0.409	0.001
Operator Status	37.058	9.98	< 0.001
Openness			
Gender	20.928	7.253	0.004
Age	-0.329	1.121	0.77
Game Knowledge	3.266	1.779	0.067
ESL	-15.45	7.661	0.048
Group Size	-7.368	1.721	< 0.001
Simulation	-9.234	7.118	0.196
Institution	-13.627	7.777	0.081
Constant	57.776	33.034	0.082
F = 12.5511, df = (2, 32), p = 0.0001			
Second Stage			
Leader Emergence	b	SE	p
$\widehat{TST}$	0.026	0.007	< 0.001
Gender	1.156	0.224	< 0.001
Age	-0.024	0.047	0.606
Game Knowledge	0.081	0.056	0.148
ESL	-0.485	0.335	0.129
Group Size	0.478	0.08	< 0.001
Simulation	0.283	0.186	0.129
Institution	0.256	0.307	0.403
Constant	-3.341	1.223	0.006
R^2 = 0.5604			
DWH F = 0.9343, df = (1, 32), p = 0.3410			
Hansen's J = 0.927, df = 1, p = 0.3358			

Note: Model results for linear two-stage least squares models with post hoc reduction in instrumental variables. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test. Standard errors are cluster-robust at the group level.

Table A4

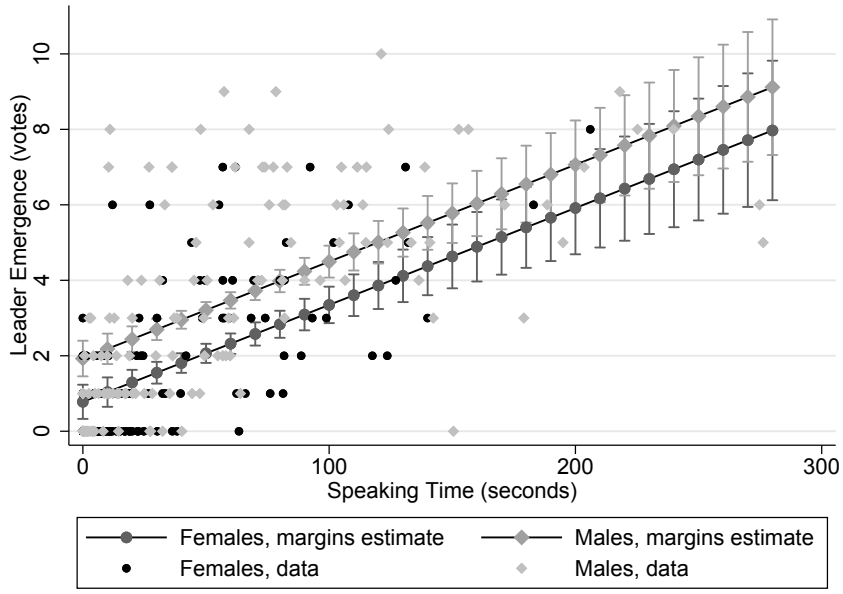
Poisson 2SLS with Instrumental Variables from the Lewbel (2012) Estimation Procedure

First Stage Total Speaking Time	b	SE	p
Gender	18.25	7.535	0.015
Age	-0.405	0.907	0.655
Game Knowledge	1.531	1.631	0.348
ESL	-15.394	7.154	0.031
Group Size	-5.683	1.768	0.001
Simulation	-9.605	6.66	0.149
Institution	-7.489	8.717	0.39
Intelligence	1.054	0.397	0.008
Operator Status	35.878	9.182	< 0.001
Conscientiousness	2.503	5.678	0.659
Agreeableness	-6.402	6.158	0.299
Neuroticism	1.804	4.75	0.704
Openness	20.553	4.846	< 0.001
Extraversion	5.892	4.5	0.19
Constant	-15.339	56.764	0.787
Second Stage **Leader Emergence**	**b**	**SE**	**p**
Gender	0.646	0.154	< 0.001
Age	0.01	0.032	0.766
Game Knowledge	0.069	0.039	0.074
ESL	-0.257	0.202	0.204
Group Size	0.188	0.056	0.001
Simulation	0.051	0.154	0.741
Institution	-0.028	0.266	0.917
$\widehat{TST}$	0.011	0.002	< 0.001
Constant	-2.431	0.965	0.012
$\hat{\rho}$	0.006821	0.00242	0.005
Wald $\chi^2 = 7.94$, df = 1, $p = 0.0048$			

Note: Model results for Poisson two-stage least squares model with Lewbel (2012) instrumental variables. The above is the residuals coefficient from the control function estimation of the Poisson instrumental variables regression. ESL: English is a second language, TST: total speaking time, DWH: Durbin-Wu-Hausman endogeneity test. Standard errors are cluster-robust at the group level.

Figure 1

Margins Plot of Two-Stage Least Squares Model with Lewbel (2012) Estimation



Note: Margins plot of main model results using the Lewbel (2012) estimation procedure. The margins estimates are model predictions averaging over all predictor variables not explicitly plotted.