DOI 10.3758/s13428-010-0019-y

Quantifying talk: developing reliable measures of verbal productivity

Margaret Wardle · Katherine Cederbaum · Harriet de Wit

Published online: 18 November 2010

Abstract Measuring talkativeness is of interest to several areas of research. However, there are few brief, validated measures available. We examined test-retest reliability, inter-relationships and convergent/divergent validity for five brief measures of verbal productivity. Nineteen men and 32 women participated in four sessions, completing five speech tasks that varied in demand, purpose of speech and sociability. Several potential metrics (word count, duration and rate) were examined. All tasks except a novel Unprompted Speech task demonstrated good word count test-retest reliability (interclass correlation coefficients from .71 to .85). Factor analysis revealed low-demand, non-functional tasks formed one factor ("Voluntary Talkativeness"), while higher demand tasks formed a second factor ("Speech Ability"). This finding and examination of relationships with IQ, personality and gender indicate "Voluntary Talkativeness" is not wholly accounted for by verbal ability, and is only weakly related to self-reported personality. Recommendations for the measurement of "Voluntary Talkativeness" are made.

Keywords Talkativeness · Speech · Task validation · Factor analysis

Introduction

Talking is a fundamental human behavior that reflects a range of underlying traits and behavioral processes. "Talkativeness" has been utilized as a diagnostic symptom of mental disorders, a behavioral indicator of personality traits and an index of drug effects. However, talkativeness has rarely been studied in its own right. Indeed, it is not entirely clear which aspects of speech comprise "talkativeness." Talkativeness might be comprised of desire to communicate, verbal intellectual ability, psychomotor speed or other factors, and might be best measured by the number of words produced, rate of speech or amount of time filled with speech. The measures used to assess talking across different areas of psychology vary widely, making it difficult to compare across studies, and little is known about the validity or reliability of these measures. Further, as different measures tend to emphasize different aspects of talking, little is known about how aspects such as rate and amount of speech relate. Thus, there is a need for sensitive and operationally defined empirical measures of talking, and for an examination of relationships between measures and metrics of talkativeness that might shed light on what constitutes the key characteristics of "talkativeness."

Measures of talkativeness are potentially important in several fields of psychology, including the study of psychopathology, personality and drug effects. In psychopathology, aspects of speech production are used in the diagnosis of many psychiatric and neurological conditions, including Parkinson's disease, depression, mania and schizophrenia (Andreasen, 1984; Lebowitz, Shear, Steed, & Strakowski, 2001; Logemann, Fisher, Boshes, & Blonsky, 1978; Ragin, Pogue-Geile, & Oltmanns, 1989; Sims, 1988). Despite the diagnostic importance of verbal output, clinicians have typically relied on subjective ratings of patient speech (Andreasen, 1984; Taylor, Reed, & Berenbaum, 1994). Although such clinical interviews employ standardized scales with behavioral anchors, subjective ratings may nevertheless be biased by patient inflection and speech rate, and do not correspond well with a computerized measure of verbal productivity in patients

M. Wardle · K. Cederbaum · H. de Wit (✉)
Department of Psychiatry and Behavioral Neuroscience,
University of Chicago,
5841 S. Maryland Ave., MC 3077,
Chicago, IL 60637, USA
e-mail: hdew@uchicago.edu

(Cohen, Alpert, Nienow, Dinzeo, & Docherty, 2008). Short, standardized measures that do not require an extensively clinically trained observer might thus be helpful to researchers of a variety of disorders.

A reliable quantitative measure of speech production would also be valuable for use in healthy normal populations. Verbal productivity has been studied in relation to personality (Digman, 1990; Thorne, 1987), gender differences (James & Drakich, 1993; Leaper & Ayres, 2007) and interpersonal attitude formation (Stewart & Ryan, 1982; Street, Brady, & Putman, 1983). In studies with healthy normal subjects, verbal productivity has typically been measured using self-report and peer ratings of "talkativeness." However, several factors can influence subjective ratings of talkativeness, including gender, age, speech rate and inflection of the speaker. Further, these observer ratings are not always related to objective measures of the amount of time spent talking. For example, ratings of talkativeness correlate with a speaker's trait extraversion, but neither ratings of talkativeness nor extraversion are related to time spent talking (Thorne, 1987).

In contrast, studies of substance use have typically used brief objective measures of talking. These measures have included speaking during a monologue task, descriptions of film clips or standardized verbal fluency measures, and have been shown to be sensitive to manipulation with administration of amphetamines, alcohol and marijuana (Higgins & Stitzer, 1989; Marrone, Pardo, Krauss, & Hart, 2010; Stitzer, Griffiths, & Liebson, 1978). However, these measures sometimes yield conflicting results (Eckardt et al., 2006; Higgins & Stitzer, 1989), supporting the need for standardization and comparison of measures.

Thus, talking and speech production have been studied in both clinical and healthy populations, but the measures used vary widely. For this study we chose to concentrate on brief, objective measures of talkativeness that were drawn from a variety of fields of research. We utilized several key potential metrics of talkativeness, including word count, duration of time spent talking and rate of speech. We selected tasks that varied in terms of demand characteristics, the purpose of the speech and whether the speech was social in nature to represent a range of underlying processes from psychomotor abilities to social motives.

The tasks selected were: (1) the Controlled Oral Word Association task (COWAT), a high-demand task in which only correct verbal output was counted (measuring psychomotor ability to produce speech accurately under external time pressure); (2) the Map task, a moderate-demand task in which speech was goal oriented (required for a functional purpose), but amount and rate were left up to the participant; (3) the Monologue Speech task, a low-demand task that prompted non-goal-oriented speech in a non-social environment; (4) the Interpersonal Speech task, a low demand task that prompted non-goal-oriented speech in a social environment, and (5) the Unprompted Speech task, in which speech was completely voluntary and unstructured with no explicit instruction to talk.

We examined the stability and reliability of these objective measures of verbal behavior, and the correspondence among these measures, to see whether different methods of eliciting speech would cohere around a single underlying construct of talkativeness, or whether separate psychomotor, social or other dimensions would become apparent. We also examined the relationships of these measures to written communication, using a Hypergraphia task, to see whether individual differences in potentially modality-independent aspects such as "desire to communicate" would be apparent across communication methods. Finally, we examined the relationship of these tasks to other individual characteristics such as sex, personality and global intellectual functioning to establish convergent and divergent validity relative to individual differences that might be expected to influence verbal output.

Methods

Subjects

Male (N = 19) and female (N = 32) native-English speakers aged 18–35 with at least a high school level of education were recruited from the University of Chicago and the surrounding community via Internet advertisements. Volunteers with serious medical conditions, a current diagnosis of a Major Axis I DSM-IV disorder or who took daily medication besides hormonal birth control were excluded (exclusions were made on the basis of a physical conducted by a doctor or nurse, an electrocardiogram and a psychiatric interview using an abbreviated version of the Structured Clinical Interview for DSM-IV; First, Spitzer, Gibbon, & Williams, 1996). All volunteers gave informed consent and were debriefed following the study. The University of Chicago Hospital's Institutional Review Committee for the use of human subjects approved the experimental protocol. Demographics of the sample may be seen in Table 1.

Procedure

The study utilized a four-session, within-subject, repeated measures design. On each session, subjects performed six behavioral tasks designed to measure verbal productivity. Subjects also completed personality questionnaires and intelligence assessments at an initial orientation session.

Orientation

First, subjects provided informed consent and underwent screening for psychiatric, physical health and drug use

Table 1 Subject characteristics

	N	
Sex (males/females)	19 males/32 females	
Ethnicity/race		
Caucasian	37	
Asian	5	
African-American	4	
Hispanic/Latino	3	
Multi-racial	2	
	Mean	SEM
Age	23.0	0.5
Current drug use		
Alcohol drinks/week	4.4	0.7
Caffeine cups/week	8.0	1.1
Cigarettes /week	0.8	0.4
Marijuana times/week	1.4	0.9
MPQ personality		
Positive emotionality	54.3	1.5
Negative emotionality	49.2	1.4
Constraint	40.1	1.2
Shipley scale		
IQ	116.3	4.8

history. They completed the Abbreviated Version of the Multidimensional Personality Questionnaire (MPQ; Patrick, Curtin, & Tellegen, 2002) and the Shipley Institute of Living Scale (Shipley, 1986). The MPQ measures three higher order personality factors (Positive Emotionality, Negative Emotionality and Constraint) and consists of 11 primary trait scales. We focused on scores from four of these: Well-Being, Social Closeness, Social Potency and Achievement. These trait scales coalesce around the factor of Positive Emotionality (PEM) and correspond to overall Extraversion, a trait expected to relate to talkativeness. The Shipley is an IQ test designed to assess general intellectual functioning. We used a single global measure of IQ as our outcome measure.

Experimental sessions

Four 1-h sessions were conducted in the laboratory at least 24 h apart. Participants abstained from alcohol and recreational drugs for 24 h prior to each study session. Pre-session urine tests were obtained to detect recent drug use. The sessions took place in comfortable rooms with a computer for administering questionnaires, and sessions were audio recorded with an Olympus DS-2 Digital Voice Recorder. Subjects completed the six tasks, listed below, in the same order on each session. No reading materials or cell phones were permitted during sessions.

Behavioral tasks

(1) Controlled Oral Word Association Test of Verbal Fluency (COWAT)

The COWAT is a widely used measure of phonemic verbal fluency (Loonstra, Tarlow, & Sellers, 2001; Ruff, Light, Parker, & Levin, 1996), involving word associations. The COWAT is most typically utilized in neuropsychology settings to measure age- or disease process-related declines in verbal ability, and has had a lengthy history of use since its development (Bechtoldt, Benton, & Fogel, 1962). Subjects are given 1 min for each letter to produce as many words as possible beginning with F, A and S. In our study, the letters were given to participants in random order to minimize practice effects across sessions. The primary outcome measure was the total number of correct words generated, minus number of incorrect words. The COWAT is distinct from the other measures used in that it has been extensively validated (Dikmen, Heaton, Grant, & Temkin, 1999), and thus can be considered somewhat as a "standard" for short verbal tasks against which our more novel other tasks can be measured. However, the high demand nature of the COWAT (i.e., instructions to produce as many words as possible in a short time frame) may also obscure other processes of "talkativeness," such as the spontaneous desire to communicate. Thus, we consider the COWAT as a standardized measure of speaking ability that should relate to IQ, but may or may not closely relate to our other, less structured, talkativeness tasks.

(2) Map task

This task was an adaptation of the Map Task Corpus (Anderson et al., 1991), and provided a measure of goal-oriented speaking to communicate a specific purpose. The Map task was designed to elicit dialog between two participants with the goal of one participant reproducing a schematic map based solely on instructions given by another participant. In our version, rather than two participants collaborating, a single participant was given a map with landmarks and a marked route, and was instructed to describe the route to the experimenter, who purportedly had the same map without the route marked. The experimenter's back was turned to the participant to minimize nonverbal communication. The experimenter used standardized verbal responses to ensure consistency and neutrality across subjects. A different route map was used for each session, but all subjects received the same maps in the same order. This task has previously been used to study a variety of linguistic questions, including what communication strategies lead to more or less successful reproductions of the original map (Anderson et al., 1991) but has not been used to examine individual differences in talkativeness per se.

The Map task provided three measures of verbal productivity: word count, i.e., the total number of words spoken over the duration of the task; speech duration, i.e., the total amount of time spent talking; and speech rate, i.e., the number of words spoken per second while the participant speaks (see below).

(3) Monologue Speech task

This relatively unstructured task, adapted from Higgins and Stitzer (1989), measured spontaneous speaking. It consisted of a 10-min period in which subjects were allowed to talk while they were alone in a room with a voice recorder (the original task was 40 min in duration). Subjects were told that they could talk as much or as little as they liked, about any topic. They were given some suggestions, but not limited to, a list of topics to talk about (e.g., family, friends, travel, current events), and could change topics as often as they wanted. This type of task has been used previously to measure the effects of drugs such as alcohol and amphetamines on speech produced in a non-social environment (Higgins & Stitzer, 1989), and provides a measure of relatively low-demand, non-social and non-goal oriented speech. Outcome measures on this task were word count, speech duration, and speech rate.

(4) Interpersonal Speech task

This task was a modified version of the interpersonal perception task employed by Janowsky (2003) and Janowsky, Kraft, Clopton, and Huey (1984), which was originally designed to study the effects of mood on interpersonal relationships. The Interpersonal task consists of a semi-structured social interaction in which the participant is asked to talk to an experimenter about topics of personal significance. In our version, participants were asked to spend 5 min talking with a female experimenter about a significant person in their lives (this is adapted from the original task, which consisted of a 15-min interview with a mental health professional). The participant nominated four different significant individuals during the orientation session, and a different significant individual was discussed during each experimental session. Experimenters were trained in basic active listening skills (Klerman & Weissman, 1993), and did not interrupt or initiate conversation unless the participant had not spoken for 10 s. If the participant did not speak for 10 s, the experimenter prompted them with one of several set questions, such as: “Tell me about how you met [this person],” “How has your relationship changed over time?” or “Is there anything else you’d like to tell me about [this person]?” This type of task has previously been used to examine whether mood influences subject assessments of the interviewer’s interpersonal skills, but has not previously been used to measure talkativeness. This task was chosen for adaptation because it provided a sample of relatively low-demand, non-goal oriented social speech to compliment the previously described monologue task, which measured non-social speech. Outcome measures on this task were word count, speech duration and speech rate.

(5) Unprompted Speech task

This was a completely novel, unvalidated task designed to measure unprompted, social speech in the presence of another individual. This task was designed based on informal observations of greater unprompted speech during amphetamine intoxication in another experiment in our laboratory. These observations suggested that completely unsolicited speech might capture an important dimension of talkativeness that was unrepresented by any measures we were able to find in the literature. In this task, the experimenter informed the participant that she needed a few moments to set up a questionnaire on a computer and turned her back to the participant for 60s. Spontaneous speech during this 60-s period was surreptitiously recorded. This task complimented the above-described tasks by providing a completely unsolicited and very low demand opportunity to speak, primarily emphasizing desire to communicate. Outcome measures on this task were word count, speech duration and speech rate.

(6) Hypergraphia task

Lastly, we designed a brief novel task to evaluate writing productivity as an exploratory analysis to examine whether factors contributing to individual differences in talkativeness (such as desire to communicate) would be evident across communication modalities. At the end of each session, participants were given 2 min to write “Any thoughts or feelings you have about the session you just completed.” Thus, this task was most similar to the Monolog task, in that it produced non-goal oriented writing in a non-social environment. The dependent measure for this task was word count.

Operational definition of verbal productivity dimensions

Word count, speech duration and speech rate were scored as follows:

(1) Word count

To count the total number of words within a task, research assistants listened to audio-recordings of the task and tallied the words with a mechanical counter. Non-word utterances such as “umm,” “uhh,” “err” and “ehh” were included in the count for all tasks except the COWAT (which is normatively scored using correct words only). The decision to include non-word utterances was made on the basis that these "filled pauses," indicate intent or urge to speak, and thus should be treated differently than silence. Indeed, in previous studies, filled pauses have been particularly sensitive to drug effects (Marrone et al., 2010), indicating that they do differ qualitatively from unfilled pauses. A 20-s segment of each task was re-counted by a "standard" rater (the second author). If the "standard" rater's count of the sample differed by more than two words from the count given by the research assistant, the entire recording was recounted. This procedure resulted in less than 1% of individual tasks needing to be re-counted, and none of these final counts were discrepant from the original counts by more than ten words.[1]

(2) Speech duration

To calculate speech duration, research assistants measured the duration of each *discrete utterance*, defined as a speech event that represents the expression of a coherent idea (Evans & Green, 2006). A speaking episode was considered halted when an individual pause reached 3 s (similar to criteria previously used to distinguish "private" utterances in children from utterances directed at someone else; Winsler, Fernyhough, McClaren, & Way, 2005). Although this measure is likely to be positively correlated with word count, it is also dependent on the speed of pronunciation and pacing (as brief pauses between words would be counted as part of the time taken to produce a discrete utterance), and it is unknown how closely these measures will relate. The Interpersonal, Monolog and Unprompted tasks had designated task lengths, lasting 300 s, 600 s and 60 s, respectively, so for these tasks duration may be considered as how much of the available time was voluntarily filled by the participant. The Map task did not have a fixed length, but rather lasted as long as required to complete the route. Thus, duration for this task may be influenced by a number of other factors, including how good the participant was at giving directions. Therefore, speech duration may or may not closely relate across tasks.

As with the word count measure, a 20-s segment of each task was re-scored by the standard rater, and if the standard rater's score differed from that of the research assistant by more than 1 s, the entire rating was re-scored. However, as with the word count, comparatively few tasks needed to be re-scored, and final scores did not differ appreciably from the original scores. This measure was available for all measures except the COWAT and the supplementary Hypergraphia task. The COWAT produces single words, rather than distinct ideas expressed as phrases, and the participant was prompted to continue naming words for the entire duration of the task; thus this measure would not be particularly meaningful for the COWAT.

(3) Speech rate

Speech rate was defined as the word count divided by duration of talk time (measured as described above), resulting in a score of words per second. Speech rate was rounded to the nearest hundredth. Speech rate measures the rapidity of speaking during speaking episodes, excluding pauses of more than 3 s. This measure was available for all tasks except the COWAT and the supplementary Hypergraphia task.

Statistics

Reliability and stability We used the intraclass correlation coefficient (ICC) in a one-factor random effects model to calculate test-retest reliability across study sessions (John & Benet-Martínez, 2000). Like other measures of reliability, the ICC ranges from 0 to 1 with larger numbers indicating a larger percentage of overall variance accounted for by between-subject variance, signifying stability within subjects across time (Shrout & Fleiss, 1979). Although rules of thumb are controversial with reliability statistics, the originator of the ICC recommended that measures with scores of 0 - .1 be considered to have virtually no reliability, .11 - .4 slight, .41-.60 fair, .61 - .80 moderate and .81 - .90 substantial reliability (Shrout, 1998). To examine stability, we conducted a repeated measures analysis of variance across the four sessions for each measure to determine whether there were systematic increases or decreases in each measure across the four sessions, which might indicate either practice or form effects.

Relationships between tasks Correlations between tasks were measured, and an exploratory factor analysis was conducted to explore the extent to which the observed correlations formed stable factors.

Relationships between talking and demographic characteristics Correlations were conducted between the task measures and MPQ personality scores, and Shipley intelligence scores. Independent t-tests were used to examine gender differences.

Results

Descriptive statistics and distributional characteristics

Examining the tasks, Unprompted speech showed a severe right skew on all measures, as only 26% of subjects spoke during the unprompted period at any session. Examining

[1] At the time of this writing, commercially available voice recognition software required extensive training of the speaker/software for accuracy. We feared training our participants to speak in a particular way to be recognizable to the computer would interfere with the natural patterns of speech. Future developments in technology may enable greater automation of this process.

the potential metrics (word count, speech rate and speech duration) for the other measures, both word count and speech rate were normally distributed for all other measures. However, speech duration showed a left skew on the Monolog and Interpersonal tasks, indicating that when there was a time limit, individuals tended to fill almost the entire time period with continuous utterances, thus presumably varying mainly in speech rate and the number of short pauses taken (as word count was much more normally distributed). Speech duration was also highly correlated with word count (between $r = .56$ and $r = .94$, $p < .01$ for all, when averaged across all sessions). Because of the likelihood of high overlap with word count and the poorer distributional characteristics of speech duration, the decision was made to focus on word count and speech rate as the primary metrics of interest. There was one high outlier for word count on the Map task. After it was determined that this outlier was inflating correlations with some other measures, this outlier was removed from all analyses, leaving $n = 50$ participants. Descriptive statistics for word count and speech rate for all tasks at all sessions may be seen in Table 2.

Behavioral task reliability and stability

The mean number of days between sessions was 2.63 (SD = 1.39, range = 1 – 13). Four of the six behavioral tasks demonstrated moderate to substantial reliability in word counts across these repeated administrations (see Table 2). Only the unvalidated Unprompted and Hypergraphia tasks did not reach acceptable reliability. In the case of the Unprompted task this was probably because many subjects did not speak at all, suggesting that this was not a sensitive measure. It is notable that word count for most of our other novel tasks reached reliability levels equivalent to or greater than the more standardized COWAT task. Speech rate also reached moderate levels of reliability for the Interpersonal task, while the Map and Monolog tasks had fair reliability.

Regarding stability, we did observe effects of time (which might indicate practice or form effects, see Table 2) on some of our dependent measures (effects are Greenhouse-Geisser corrected for violations of sphericity where necessary). On the COWAT there was a linear effect such that participants reported more correct words over time, F(2.51, 123.06) = 26.80, $p < .001$, $\eta_p^2 = .35$. There was also a significant effect of time on Map word count; however, this was not the linear decrease that would be expected of a simple practice effect, but rather a quadratic trend that might indicate variations in the difficulty of the alternate map forms used, F(2.12, 103.81) = 7.60, $p = .001$, $\eta_p^2 = .13$. There was, however, a linear effect of time on Map speech rate, with participants speaking more rapidly across the sessions, perhaps indicating increasing fluidity with direction-giving, F(2.46, 120.74) = 5.47, $p = .003$, $\eta_p^2 = .10$. In contrast, the non-goal-oriented Monolog task did not show significant effects of time on either word count or speech rate, so participants did not appear to become fatigued with the task over subsequent days. The Interpersonal task showed a quadratic effect of time on word count, with

Table 2 Means and standard deviations for all tasks at all sessions, and intraclass correlation coefficient (ICC) test-retest reliability

	Session 1		Session 2		Session 3		Session 4		ICC
	M	SD	M	SD	M	SD	M	SD	
COWAT									
Word count	54	14	58	14	61	14	64	15	0.77*
Route reproduction									
Word count	221	110	190	98	200	104	218	101	0.85**
Speech Rate	2.08	0.42	2.17	0.46	2.19	0.42	2.30	0.42	0.57
Monolog speech									
Word count	1256	342	1153	401	1139	420	1179	423	0.71*
Speech rate	2.32	0.48	2.27	0.47	2.23	0.53	2.20	0.61	0.59
Interpersonal speech									
Word count	685	158	669	154	680	140	711	144	0.77*
Speech rate	2.44	0.52	2.37	0.49	2.43	0.43	2.53	0.44	0.76*
Spontaneous speech									
Word count	4	10	2	11	2	8	4	12	0.30
Speech rate	0.55	1.27	0.22	0.90	0.19	0.74	0.52	1.45	0.25
Hypergraphia									
Word count	52	26	46	22	37	21	32	30	0.42

Reliablity: 0 - .1 virtually no, .11 - .4 slight, .41- .60 fair, *.61 - .80 moderate and **.81 - .90 substantial (Shrout, 1998).

word count decreasing on days 2 and 3, and then increasing again on day 4, F(2.86, 140.31) = 3.12, η_p^2 = .06. This is difficult to explain, as it does not fit a simple pattern of fatigue, or of increasing comfort with the experimenter. This was also a very small effect and may have been due to chance variation. A similar quadratic effect was seen on speech rate on this task, F(2.63, 129.16) = 4.01, *p* = .008, η_p^2 = .08. There were no systematic effects of time on the Unprompted task. Hypergraphia showed a linear decrease, perhaps indicating fatigue with the task over subsequent days, F(2.37, 115.88) = 12.23, *p* < .001, η_p^2 = 20.

Relationships between measures

Due to the observed effects of time on several measures, we chose to examine relationships between measures using the first session scores only. The time effects noted above might alter the relationship between the measures during subsequent sessions, so we wished to first establish the relationship between measures at a single administration before confirming whether these relationships persisted over repeated administrations.

Correlations Word count and speech rate were positively and significantly correlated within each task (as can be seen in Table 3). These correlations ranged from low-moderate on the Map task (r = .39, *p* < .01) to very high on the Interpersonal task (*r* = .94, *p* < .001). The lower correlation on the non-time-limited Map task makes intuitive sense, given (as noted above) that on tasks with time limits the participants tended to utilize nearly the entire task time for utterances, meaning that word count was primarily determined by speech rate on these tasks.

As can be seen in Table 3, the more novel tasks developed for this study were not significantly correlated with the COWAT (except for speech rate on the Route Reproduction task), suggesting the aspects of talkativeness measured by these tasks were not wholly accounted for by psychomotor ability to produce correct words. Speech rates were moderately to highly correlated across the Map, Monolog and Interpersonal tasks, as were word counts on the Monolog and Interpersonal tasks. Unprompted speech did not correlate significantly with any of the other tasks, which might be due to limitations in range noted above. Finally, there were few cross-modal correlations with Hypergraphia, except for the word count on the Monolog task (this was the most similar task to Hypergraphia in terms of demand characteristics and level of sociability).

Factor analysis Although we had a comparatively small sample of subjects for this technique (Tabachnick & Fidell, 2001), we did conduct an exploratory Principal Factor Analysis using varimax rotation on first session scores to see if the cluster of correlations identified between the Map, Monolog, and Interpersonal tasks above would load adequately on a single factor (indicating shared underlying processes). We included all of the behavioral measures with the exception of Hypergraphia, which was secondary to our main interests, and the Spontaneous task, which demonstrated a lack of sensitivity and poor correlation with the other dependent variables. After an initial extraction of all factors with Eigenvalues > 1, examination of the scree plot,

Table 3 Correlations between all behavioral measures

	Route reproduction		Monolog speech		Interpersonal speech		Spontaneous speech		Hyper-graphia
	WC	SR	WC	SR	WC	SR	WC	SR	WC
COWAT									
Word count	.20	.33*	.24	.35*	.29*	.31*	-.08	.03	.00
Route reproduction									
Word count	–	.39**	-.06	.05	.16	.12	-.04	.07	.27
Speech rate		–	.23	.41**	.22	.31*	.13	.06	.10
Monolog speech									
Word count			–	.76**	.56**	.51**	.09	.13	.30*
Speech rate				–	.69**	.79**	.07	.08	.18
Interpersonal speech									
Word count					–	.94**	.19	.18	.15
Speech rate						–	,18	.15	.12
Spontaneous speech									
Word count							–	.74**	-.12
Speech rate								–	.09

* *p* < .05, ** *p* < .01

Velicer's MAP test and parallel analyses (O'Connor, 2000) indicated that a two-factor solution was most appropriate. Loadings of the individual tasks on these two factors are presented in Table 4. The first factor accounted for 50% of the observed variance in scores, with Monolog word count and rate, and Interpersonal word count and rate loading strongly. The second accounted for 19% of the observed variance with Map word count and rate and COWAT word count loading adequately (loadings > .32 were considered adequate; Tabachnick & Fidell, 2001). Based on the characteristics of the tasks comprising the first factor, this factor appeared to represent lower-demand, non-goal directed speech, which we called "Voluntary Talkativeness." In contrast, the second factor contained goal-oriented tasks likely to be more related to cognitive abilities, which we called "Speech Ability." To quantify the extent to which each factor reliably tapped a single-construct Cronbach's alpha was calculated for each. The Voluntary Talkativeness factor had α = 0.91, indicating that these items appear to adequately assess a single construct. The Speech Ability factor had α = .56, which may indicate that these tasks may differ more in the aspects of speech/cognition that they index.

As noted above, the observed time effects might lead to alterations in this factor structure over repeated administrations. The ideal way to address this question would be to conduct a confirmatory factor analysis on subsequent time points to verify the observed factor structure. However, our small sample size made a type II error highly likely (Tabachnick & Fidell, 2001); thus, this was outside the scope of the current study. We did examine the test-retest reliability for the extracted factors, by producing factor scores for each time point (multiplying z-scores for the tasks by the factor loadings and then summing those scores) and calculating the ICC coefficient for these scores. Changes in factor composition (indicating that the factor is measuring different processes at different time points) would make good test-retest reliability less likely. The ICC for the "Voluntary Talkativeness" factor was .81, indicating substantial reliability, and for the "Speech Ability" factor was .79, indicating moderate reliability.

Personality, intelligence and gender relationships

As a first step to establishing convergent and divergent validity, we examined the relationships between our tasks and personality characteristics (extraversion), intelligence and gender, each of which might be expected to relate to talkativeness. Due to the time effects observed above, we calculated correlations between the behavioral tasks and six personality factors as measured by the MPQ and Shipley IQ scores and t-tests of gender differences using the subjects' scores from the first testing occasion only. Monolog speech word count was positively correlated with MPQ Social Closeness. No other behavioral task dimensions were related to personality traits. Shipley IQ scores were correlated only with verbal fluency (r = .39, p < .01). Women had a higher speech rate than men on both the Monolog, t(48) = 2.14, p = .04 and the Interpersonal task, t (48) = 3.27, p = .002, with .28 words per second more on the Monolog (SE = .13) and .46 more words per second on the Interpersonal task (SE = .11). Women also had a higher word count on the Interpersonal task, t(48) = 3.48, p = .001, speaking 138 more words than men on average (SE = 42). No other behavioral task showed an effect of gender.

Discussion

In this study, we investigated several short measures of verbal behavior, for reliability, stability, relationships between measures, and initial convergent and divergent validity.

This study first addressed the question of whether the talkativeness (word count and speech rate) of an individual

Table 4 Factor loadings for exploratory factor analysis

	Factor 1 "voluntary talkativeness"	Factor 2 "speech ability"
COWAT		
Word count	–	.37
Route reproduction		
Word count	–	.61
Speech rate	–	.65
Monolog speech		
Word count	.70	–
Speech rate	.89	–
Interpersonal speech		
Word count	.84	–
Speech rate	.88	–

Loadings below .32 are replaced with –

on these tasks was reliable and stable across days, on several different measures of talking. All the tasks except Unprompted Speech and Hypergraphia reached adequate reliability on word count across four different study sessions. Our more novel Map, Monolog and Interpersonal tasks were equivalent to the more established COWAT in terms of test-retest reliability. Speech rate was somewhat less reliable, with only the Interpersonal task showing good reliability for speech rate.

Some of the measures of talking did show systematic effects of time across the 4 days. These effects were particularly pronounced for the COWAT and Map Tasks, which may be more subject to practice effects. Additionally, in the case of the Map task, differences between the forms introduced a bias. In contrast, the Monolog and Interpersonal tasks showed small or non-significant effects of time.

Correlations and factor analysis were conducted using data from only the first session, and the factor analysis excluded the unreliable Unprompted and the secondary Hypergraphia task. We found two main factors: one well-correlated factor that we labeled "Voluntary Talkativeness," comprising all measures for the Monolog and Interpersonal tasks, and a second smaller and more diverse factor labeled "Speech Ability," comprising the higher demand COWAT and Map task. Thus, the first factor was not highly related to the psychomotor ability to produce speech quickly and accurately. The loading of the COWAT on the second factor was low, suggesting that additional underlying processes contribute to this form of talking. Although we did not determine whether this factor structure was stable with repeated administrations of the tasks, the high test-retest reliability of the "Voluntary Talkativeness" factor, and the comparative lack of practice effects on the scales comprising this factor suggests that at least this factor would be suitable for repeated use.

One way of assessing convergent and divergent validity was to examine relationships with other measures that might be expected to correlate with talkativeness. The correlation between Monolog and Hypergraphia tasks suggests convergent validity across modalities among certain of the tasks (although the Interpersonal task was highly related to the Monolog task, but not to the Hypergraphia task). The correlation between IQ and verbal fluency performance is consistent with other studies (Boone, 1999), but none of the other measures of talkativeness were related to intelligence, suggesting that these tasks do not simply measure linguistic ability. We expected, but did not find, scales from the MPQ that measure extraversion-related factors to be related to talkativeness. The one exception was a correlation between Monolog word count and social closeness, although this was puzzling as the Monolog task was not considered social. However, nomothetic personality measures often have low correlations with single point measures of actual behavior, including talking behavior (Cervone & Shoda, 1999; Thorne, 1987), so this does not necessarily indicate that our measures lack validity. Future studies might compare talking elicited in the laboratory with talking gathered through ecological means to determine the extent to which these laboratory samples correlate with average real-world behavior (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). Finally, we found that women had higher speech rates and word counts in the Monolog and Interpersonal tasks, but that they did not speak more on the tasks that were directive and goal-oriented (as in the COWAT and Map task). There is controversy about whether women speak more than men (Leaper & Ayres, 2007). Our results suggest that part of this controversy may be related to the manner in which the speech is elicited. Future studies might also consider systematically varying the gender of the experimenter. We held gender constant, which might also contribute to different effects in same vs. opposite sex dyads.

Limitations of this study included the comparatively small sample size, lack of alternate forms for many of the tasks, the restricted range of participants (mostly college students in a large city in the Midwest), the fixed order of administration for the tasks/forms both within and across days, and lack of knowledge about the content of the speech samples. The limited range of participants may have inflated reliability (for example, there was a general lack of regional accents in the current sample), as might practice effects (although the extent to which such systematic effects increase reliability is a matter of debate; McKelvie, 1992). Further, the fixed order of tasks may have influenced the observed factor structure, with tasks administered closer together showing more relationship because of extraneous phenomena such as fatigue. Thus, priorities for future investigation include examination of this factor structure in a larger and more diverse sample using counterbalanced order of administration, and explicit investigation of the extent to which practice effects do or do not alter the observed factor structure using confirmatory factor analysis over at least two administrations.

Regarding the content of speech, it is always possible that our decision to include filled pauses in all tasks aside from the COWAT may have produced some of the separation between the "Speech Ability" and the "Voluntary Talkativeness" factor. In future studies, combining the techniques used here with transcription of the verbal samples would allow both for easy assessment of the impact of the inclusion of non-words on the findings, and analysis of the content of speech for positive vs. negative words or other potential speech structures of interest (Pennebaker, Booth, & Francis, 2007).

The evidence for convergent and divergent validity of these tasks should also be further investigated. We found only slight evidence of convergent validity with personality (although, as noted, ecological momentary assessment of talkativeness may be a more valid way to examine the question of whether these samples represent behavior in the real word). Additionally, if these tasks are to be extended to clinical populations, they will need to be examined against clinical interviews such as the schizophrenia scales described in the introduction (Andreasen, 1984).

Last, validation of alternate equivalent forms of the COWAT and Map task will be an important next step, particularly for researchers interested in assessing the "Speech Ability" factor. Addition of other measures expected to cluster with the Map task and COWAT may also help elucidate the nature of the processes underlying this factor, as there were indications that this factor had less internal consistency than the "Voluntary Speech" factor.

Conclusions and future directions

In summary, by studying several different forms of talkativeness, we have identified two underlying processes, one relating to relatively unstructured speech (Voluntary Talkativeness) and the other related to speech produced to achieve specific goals (Speech Ability). For future studies wishing to assess Voluntary Talkativeness, we would recommend use of word counts on the Monolog or Interpersonal tasks, as these demonstrated the best reliability, had small or non-significant practice effects, and coherently indexed a single underlying factor.

Overall, this study provides the basis for future, hypothesis-driven studies investigating the underlying factor structure of verbal output, investigating the social, cognitive and psychomotor components of talkativeness, and measuring the effects of drugs or other environmental factors on speech. By studying why, when and how people talk, future research may improve our understanding of the motivations and manifestations of talk, and the mechanisms of individual differences in talkativeness. Future research will also help to identify different types of speech and sources of variability in the relationships between speaking tasks and components of speaking.

Acknowledgements This research was supported by DA02812. We would like to thank Divya Anchan, Robert Oliver, Erica Ting and Andrea Wan for their contributions.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E. A., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and Speech, 34*, 351–366.

Andreasen, N. C. (1984). *Scale for the assessment of positive symptoms*. Iowa City: University of Iowa.

Bechtoldt, H. P., Benton, A. L., & Fogel, M. L. (1962). An application of factor analysis in neuropsychology. *The Psychological Record, 12*, 147–156.

Boone, K. B. (1999). *Neuropsychological assessment of executive functions: Impact of age, education, gender, intellectual level, and vascular status on executive test scores. The human frontal lobes: Functions and disorders* (pp. 247–260). New York: Guilford Press.

Cervone, D., & Shoda, Y. (1999). Beyond traits in the study of personality coherence. *Current Directions in Psychological Science, 8*, 27–32.

Cohen, A. S., Alpert, M., Nienow, T. M., Dinzeo, T. J., & Docherty, N. M. (2008). Computerized measurement of negative symptoms in schizophrenia. *Journal of Psychiatric Research, 42*, 827–836.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan neuropsychological test battery. *Journal of the International Neuropsychological Society, 5*, 346–356.

Eckardt, M., File, S., Gessa, G., Grant, K., Guerri, C., Hoffman, P., et al. (2006). Effects of moderate alcohol consumption on the central nervous system. *Alcoholism, Clinical and Experimental Research, 22*, 998–1040.

Evans, V., & Green, M. (2006). *Cognitive linguistics: An introduction*. Mahwah: Erlbaum.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. (1996). *Strutured clinical interview for DSM-IV Axis I disorders*. New York: Biometrics Research Department.

Higgins, S. T., & Stitzer, M. L. (1989). Monologue speech: Effects of d-amphetamine, secobarbital and diazepam. *Pharmacology, Biochemistry and Behavior, 34*, 609–618.

James, D., & Drakich, J. (1993). *Understanding gender differences in amount of talk: A critical review of research. Gender and conversational interaction* (pp. 281–312). New York: Oxford University Press.

Janowsky, D. S. (2003). Depression and dysphoria effects on the interpersonal perception of negative and positive moods and caring relationships: Effects of antidepressants, amphetamine, and methylphenidate. *Current Psychiatry Reports, 5*, 451–459.

Janowsky, D. S., Kraft, A., Clopton, P., & Huey, L. (1984). Relationships of mood and interpersonal perceptions. *Comprehensive Psychiatry, 25*, 546–551.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction *Handbook of research methods in social and personality psychology.* (pp. 339-369): New York: Cambridge University Press.

Klerman, G. L., & Weissman, M. M. (1993). *New applications of interpersonal psychotherapy*. Washington: American Psychiatric Association.

Leaper, C., & Ayres, M. M. (2007). A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology Review, 11*, 328–363.

Lebowitz, B. K., Shear, P. K., Steed, M. A., & Strakowski, S. M. (2001). Verbal fluency in mania: Relationship to number of manic episodes. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 14*, 177–182.

Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, E. R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *The Journal of Speech and Hearing Disorders, 43*, 47–57.

Loonstra, A. S., Tarlow, A. R., & Sellers, A. H. (2001). COWAT metanorms across age, education, and gender. *Applied Neuropsychology, 8*, 161–166.

Marrone, G. F., Pardo, J. S., Krauss, R. M., & Hart, C. L. (2010). Amphetamine analogs methamphetamine and 3, 4-methylenedioxymethamphetamine (MDMA) differentially affect speech. *Psychopharmacology, 208*, 169–177.

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *The Journal of General Psychology, 119*, 59–72.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorded (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers, 33*, 517–523.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*, 396–402.

Patrick, C. J., Curtin, J. J., & Tellegen, A. (2002). Development and validation of a brief form of the multidimensional personality questionnaire. *Psychological Assessment, 14*, 150–163.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC 2007*. Austin: LIWC.

Ragin, A. B., Pogue-Geile, M., & Oltmanns, T. F. (1989). Poverty of speech in schizophrenia and depression during in-patient and post-hospital periods. *The British Journal of Psychiatry, 154*, 52–57.

Ruff, R. M., Light, R. H., Parker, S. B., & Levin, H. S. (1996). Benton Controlled Oral Word Association Test: Reliability and updated norms. *Archives of Clinical Neuropsychology, 11*, 329–338.

Shipley, W. (1986). *Shipley institute of living scale*. Los Angeles: Western Psychological Services.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*, 301.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Sims, A. (1988). *Symptoms in the mind: An introduction to descriptive psychopathology*. London: Bailliere Tindall Publishers.

Stewart, M. A., & Ryan, E. B. (1982). Attitudes toward younger and older adult speakers: Effects of varying speech rates. *Journal of Language and Social Psychology, 1*, 91.

Stitzer, M. L., Griffiths, R. R., & Liebson, I. (1978). Effects of d-amphetamine on speaking in isolated humans. *Pharmacology, Biochemistry and Behavior, 9*, 57–63.

Street, R. L., Brady, R. M., & Putman, W. B. (1983). The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers. *Journal of Language and Social Psychology, 2*, 37–56.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights: Allyn and Bacon.

Taylor, M. A., Reed, R., & Berenbaum, S. A. (1994). Patterns of speech disorders in schizophrenia and mania. *The Journal of Nervous and Mental Disease, 182*, 319–326.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology, 53*, 718–726.

Winsler, A., Fernyhough, C., McClaren, E. M., & Way, E. (2005). Private speech coding manual, *Unpublished manuscript*. Fairfax, VA: George Mason University. Available at: http://classweb.gmu.edu/awinsler/Resources/PsCodingManual.pdf