JRRD

Volume 51, Number 4, 2014

Pages 579–590

Reliability of brain-computer interface language sample transcription procedures

Katya Hill, PhD, CCC-SLP;[*] Thomas Kovacs, MA; Sangeun Shin

Research Services, Highland Drive Department of Veterans Affairs Healthcare System, Pittsburgh, PA; and School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA

Abstract—We tested the reliability of transcribing language samples of daily brain-computer interface (BCI) communication recorded as language activity monitoring (LAM) logfiles. This study determined interrater reliability and interjudge agreement for transcription of communication of veterans with amyotrophic lateral sclerosis using a P300-based BCI as an augmentative and alternative communication (AAC) system. KeyLAM software recorded logfiles in a universal logfile format during use of BCI-controlled email and word processing applications. These logfiles were encrypted and sent to our laboratory for decryption, transcription, and analysis. The study reports reliability results on transcription of 345 daily logfile samples. The procedure was found to be accurate across transcribers/raters. Frequency of agreement ratios of 97.6% for total number of words and 93.5% for total utterances were found as measures of interrater reliability. Interjudge agreement was 100% for both measures. The results indicated that transcribing language samples using LAM data is highly reliable and the fidelity of the process can be maintained. LAM data supported the transcription of a large number of samples that could not have been completed using audio and video recordings of AAC speakers. This demonstrated efficiency of LAM tools to measure language performance benefits to BCI research and clinical communities.

Key words: amyotrophic lateral sclerosis, augmentative and alternative communication, BCI, brain-computer interface, encryption, language activity monitor, logfile, performance report tool, reliability, transcription.

INTRODUCTION

Brain-computer interfaces (BCIs) are emerging from laboratory testing to become a consideration as augmentative and alternative communication (AAC) assistive technology. The field of AAC applies evidence-based practice to address the human need for interactive and intentional communication [1–2].[†] BCIs are becoming a viable communication option for veterans with the most severe communication and movement disorders, such as advanced amyotrophic lateral sclerosis (ALS) or locked-in syndrome.

[†]Hill K. Augmentative and alternative communication. State of the Science: Assistive Technology Devices; 2006 May 26; Washington, DC.

Abbreviations: AAC = augmentative and alternative communication, ALS = amyotrophic lateral sclerosis, BCI = brain-computer interface, CSP = Cooperative Studies Program, EEG = electroencephalography, LAM = language activity monitoring, PeRT = Performance Report Tool, SALT = Systematic Analysis of Language Transcripts, SGD = speech-generating device, TNW = total number of words, VA = Department of Veterans Affairs.

[*]**Address all correspondence to Katya Hill, PhD, CCC-SLP; Department of Communication Science and Disorders, University of Pittsburgh, 6017 Forbes Tower, Pittsburgh, PA 15260; 412-383-6659; fax: 412-383-6555.**

Email: khill@pitt.edu

http://dx.doi.org/10.1682/JRRD.2013.05.0102

A BCI uses central nervous system outputs by recording brain signals, extracting measures from them, and converting these measures into commands that operate assistive technologies to augment the user's natural functions [3]. A BCI using the P300 event-related potential can allow users, such as people with ALS, to choose among items in a matrix as a means of communication [4–5].

Until recently, capacity to use BCI technologies for communication has been evaluated under controlled laboratory conditions. BCIs for AAC intervention must measure performance during spontaneous, novel communication. Collection and analysis of language samples that represent independent, daily BCI use is an essential aspect of measuring the effectiveness of BCIs as AAC devices. A critical step in language sample analysis is reporting the reliability of the procedures for transcribing spontaneous language samples without complete knowledge of conversational context.

Measures of Brain-Computer Interface Performance

BCI research has focused on continued improvement of the technology and calibration process in order to increase the accuracy and speed of text generation whether the BCI functions as an invasive or noninvasive system. Invasive BCI systems where the electrode array is surgically implanted have been used for elicited production of isolated phonemes with a speech synthesizer [6]. Invasive BCI systems have never been used for production of connected speech or generative language. More external evidence has been published on the performance of noninvasive BCIs in highly controlled laboratory settings.

Several attempts have been made in noninvasive BCI studies to improve the accuracy and speed of the control movement by improving the selection of signal features, by optimizing information transfer rates, and by improving the interaction between the user and the BCI system [7].

Wolpaw et al. developed a response verification procedure in which stimuli were presented across two trials to confirm that subjects gave consistent responses to the same stimulus across trials [8]. Miner et al. used the response verification task to investigate the use of an electroencephalography (EEG)-based cursor control to answer yes/no questions [9]. In the second presentation of same stimuli questions in the response verification task, it showed 64 to 87 percent consistency to the initial answers.

Ahi et al. demonstrated that the BCI with a modified user interface in which the location of letters in the conventional A to Z matrix were rearranged through an error analysis improved accuracy in spelling four-letter words on a character-by-character basis when a set of target words are stored in a dictionary [10]. All 14 healthy subjects in this study increased their accuracy in selecting cued letters (copy-spelling) to form four-letter words. However, it is not clear whether improved accuracy using an interface modified for a word-level task can generalize to improved performance in interactive conversation.

Ryan et al. investigated whether word prediction would increase speed of BCI use by reducing the number of selections needed to generate text with a BCI device [11]. They found that the time needed to copy a 58-character sentence using a BCI device with spelling and word prediction was less than the time needed to copy the same sentence using a BCI device with spelling alone. Although speed increased, accuracy using the BCI device was lower when the word-prediction program was used.

Other studies have used experimental methods to compare user performance using BCI systems with different configurations of visual stimuli. For example, several studies have compared user performance across multiple P300 spellers with different keyboard configurations on copy-spelling tasks [12–14]. Some studies have also compared user performance across multiple P300 spellers when users spontaneously generated short messages in a free-spelling phase [13–14], but the production of spontaneous messages has been limited to no more than one sentence per subject.

Most prior studies use a copy-spelling task to measure subjects' speed and accuracy in order to report on the efficacy of the BCI for communication [15–16]. Little attention has been paid to analyzing the BCI user's performance during spontaneous communication, even though the primary goal of studies investigating use of BCI devices for communication is to enhance quality of life by improving the BCI user's communication performance [17–18].

Current evidence indicates limitations in accounting for the user's communication competence with the BCI system by only measuring speed and/or accuracy [6–16]. In addition, the copy-spelling task [9–16] has inherent limitations for evaluating the effectiveness of expressive language performance in either structured or unstructured communication environments. Quantitative performance data in terms of language-dependent variables (e.g., vocabulary frequency, mean length utterance) and speed-dependent variables (e.g., average and peak communication rate) are critical to optimize the effectiveness of AAC systems in

both the research laboratory and clinical settings. For decades, language sample analysis has been considered an important part of the assessment for people with communication difficulties. Language sample analysis not only allows the speech-language pathologist to describe the client's oral language skills in a naturalistic context, but also supports the development of treatment goals and activities [19–20]. Although time and effort are required for language sampling and transcription, the efficiency and accuracy of language sample analysis can be improved using computer software [19].

Measures of Language Performance

Language sample transcription has been the foundation of reporting linguistic performance and competence of various adult populations with nondisordered and disordered communication [21–22]. Studies that report results based on language samples typically report the reliability of the transcription procedures that have been operationalized for the protocol [22–23]. Highly consistent transcription of language samples indicates that the language or performance measures are both reliable and valid. Routinely, language sample analysis starts by generating a word-by-word transcript of each recorded utterance [24]. These transcripts are used to determine the reliability or consistency of the transcription process, and therefore, the performance measures are accurate or valid.

Although high reliability is no guarantee of high validity, reliability sets the limits for the validity of a study. Low reliability is evidence of low validity [25]. In other words, high validity cannot result from low reliability, because of the inconsistency or unstableness in the measurement. Language sampling typically reports global measures of language performance such as total utterances, total number of words (TNW), mean length of utterance in morphemes, and communication rate in words per minute. Performance Report Tool (PeRT) software developed by the AAC Institute (Pittsburgh, Pennsylvania) was used for transcription and analysis. PeRT computes these global measures of language performance based on the utterances transcribed by an operator. Although we used a reliable analysis program, transcription of language samples must be highly reliable to draw conclusions about the quality of the data and validity of the performance and outcome measures. The dependent variables of this study have been traditionally and routinely reported to measure linguistic competence [26–28]. In addition, communication rate has been recognized as essential for measuring overall AAC performance [29–31].

Language activity monitoring (LAM) is a process whereby logfile data are automatically recorded when an augmented communicator uses an AAC device [29,32]. The LAM protocol has at least two components. The first is a time stamp that indicates the time of an action in absolute real time in a 12 or 24 h format. The second component of a data logging record is the activity that occurred at the indicated time. Thus for a language event, the LAM format is hh:mm:ss "Any continuous text that is transmitted by the AAC device." For a non-language event, such as use of a hot key command to send an email, the protocol is hh:mm:ss "non-language information in continuous text." Over a given time period, the logfile is uploaded to a computer and visually inspected before being used to generate a transcript for analysis. Performance measures based on transcripts using LAM data can be automatically calculated using analysis software such as PeRT, or the Systematic Analysis of Language Transcripts (SALT) software developed by Jon Miller and SALT Software, LLC (Madison, Wisconsin) for analysis of spoken language samples. **Figure 1** depicts the process of uploading LAM data from a BCI device into a computer for analysis and reporting.

LAM logfiles include a record of language produced by the AAC speaker but typically do not include language produced by conversation partners. Unless an additional data stream is collected to account for language produced by conversation partners, spontaneous language data logged in a LAM logfile is inherently decontextualized. Reliable transcription of spontaneous LAM data requires transcription procedures that can be used with decontextualized language samples. LAM data have supported the measurement of communication performance for ALS patients using commercially available AAC systems [1].*

Performance related to word-based and utterance-based measures have been used to report the effectiveness of the AAC interventions [33]. For example, Cook and Hill documented how monitoring of selection rate

*Hill H, Rupp T, Hill K, Tucci M. AAC outcomes and persons with ALS and visual problems. National Conference of the American Speech-Language-Hearing Association; 2004 Nov 18–21; Philadelphia, PA; Hill K, Romich B, Cook S. AAC performance: The elements of average communication rate. National Convention of the American Speech-Language-Hearing Association; 2002 Nov 21–24; Atlanta, GA.

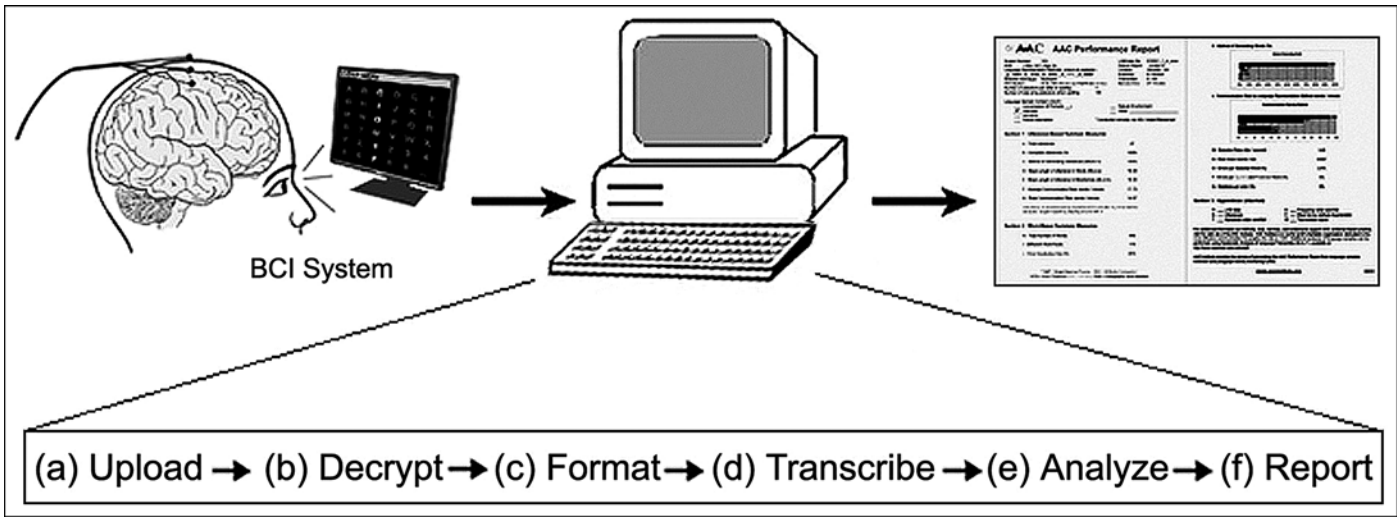

Figure 1.
Language activity monitor (LAM) data analysis and reporting procedures. **(a)** Once encrypted LAM logfiles are uploaded to computer, **(b)** logfiles are decrypted using Department of Veterans Affairs security procedures. **(c)** Next, logfiles are formatted for analysis. **(d)** Logfiles are then transcribed using Performance Report Tool (PeRT). **(e)** PeRT then automatically analyzes transcribed utterances and **(f)** generates summary report of performance measures. BCI = brain-computer interface.

and error frequencies provided evidence for modifying single-switch scanning features [34]. Hill and Jans reported on monitoring changes to AAC system use through end stages of ALS [35]. However, no similar monitoring has been done using BCIs as AAC devices for daily communication before our study. Therefore, establishing transcription procedures that resulted in high reliability testing and results was important for being able to report valid measures as study results and evaluating the overall effectiveness of expressive communication.

METHODS

Subjects

A total of 15 subjects generated at least one language sample while using the email or WordPad (Microsoft Corporation; Redmond, Washington) program on the BCI device for communication. The subjects were adult veterans with an El Escorial "Lab-Supported Probable" or more definite diagnosis of ALS [36] who had lost the ability to communicate either verbally or in writing as indicated by a score of 0 on item 1 or 4 of the ALS Functional Rating Scale, Revised [37]. At the time of enrollment, all subjects had corrected visual acuity of at least 20/80, were able to read and understand 6th grade English text on a computer screen, were able to communicate nonverbally, and were able to give informed consent using their existing communication methods. All subjects identified a system operator who agreed to be trained to set up the BCI device. All subjects and system operators had a life expectancy of at least 1 yr. All subjects lived at home within 100 mi of a participating Department of Veterans Affairs (VA) study site.

During a screening phase, all subjects demonstrated sufficient EEG interaction for the BCI device to operate, which was operationally defined as 70 percent accuracy during a copy-spelling task in a daily calibration period. The system operators were required to demonstrate sufficient skill to manage the daily setup and routine operations needed to support the subject's basic operation of the BCI device.

Data Logging and Data Security

KeyLAM, a built-in LAM feature developed by the AAC Institute, was installed on the BCI device by the Wadsworth Center (Albany, New York). KeyLAM is used to log keystrokes generated using the BCI device for the purposes of language sample transcription and analysis [29,32]. The data logging process was fully transparent. Subjects were aware of ongoing data logging throughout the study. Subjects could optionally toggle into a Privacy Mode at any time to suppress data logging

during text entry. In Privacy Mode, all keystrokes were logged as asterisks. On-screen visual feedback clearly indicated whether data logging was active or the BCI device was in Privacy Mode at all times.

Data generated by the subjects' BCI devices, including raw LAM data, were saved as encrypted logfiles. New LAM logfiles were created each time a program was started on the BCI device. If a subject started the same program multiple times in a day, multiple logfiles were created. These encrypted logfiles were electronically transferred to the Wadsworth Center via a secure internet connection on a daily basis. Logfiles from all subjects were transferred to the AAC Core Laboratory at the Pittsburgh VA Medical Center every 2 wk on encrypted discs. The logfiles were then decrypted and passed through a macro that converted the data to the universal logfile format for transcription and analysis. Each letter was logged as a separate event. When subjects used a word-prediction feature to complete a word that they began spelling, the remaining characters were logged as a series of rapid events with the same time stamp. Examples of formatted LAM data from the study are shown in **Figure 2**, including utterances produced using spelling and word prediction and a block of text generated in Privacy Mode.

Transcription and Analysis of Daily Logfile Samples

For each day that a subject used the email program, a daily email logfile sample was formed by merging all of the individual email logfiles from that day into a single .txt file. The same procedure was used to form daily WordPad logfile samples when the WordPad program was used.

The daily logfile samples were transcribed and analyzed using PeRT. PeRT generates a report of performance measures that are computed based on transcribed utterances. Transcription procedures must be reliable for PeRT to reliably compute performance measures.

Transcription procedures were defined in a procedural manual and followed established SALT standards for spoken language sample transcription and analysis as much as possible. Utterances were parsed using C-unit segmentation, with C-units defined as "an independent clause and its modifiers" [38]. Self-selected utterance terminators such as punctuation marks or the Enter key were also used to mark the ends of utterances if these keystrokes did not appear to be produced as errors, following language sample analysis procedures used with other populations of adult AAC speakers [26]. All analyzed utterances included the subjects' final written output, after self-corrected errors and revisions were accounted for. This is similar to the established practice of excluding false starts, repetitions, and reformulations, also known as maze words, from established measures in spoken language sample analysis [38] unless the researcher is specifically investigating maze words. Additional considerations were made because the subjects were using a text-entry system as the modality for their expressive communication. Such considerations include operationally defining a string of five or more unintelligible characters as an unintelligible utterance and rules for interpreting hot key commands that appeared in the logfiles (e.g., using Ctrl + Backspace to delete a whole word).

The intrarater reliability of this procedural manual was verified with simulated language sample data that were generated using a BCI system in a laboratory setting. Each rater analyzed these simulated logfile samples multiple times to establish intrarater reliability. Transcription and analysis procedures were revised to resolve any points of confusion until intrarater reliability >90 percent was established. This process verified that each of the raters could consistently transcribe and analyze the same sample multiple times and produce the same results. Once a final procedural manual was established, the raters began to analyze batches of daily logfile research samples.

Daily logfile samples were randomly assigned to three raters for data transcription and analysis. Two primary raters independently transcribed and analyzed 60 percent of the daily logfile samples, with 10 percent of the samples overlapping between raters for reliability. A third rater independently transcribed and analyzed 10 percent of the daily logfile samples. A total of 20 percent of the daily logfile samples were independently transcribed and analyzed by two different raters for reliability testing. Daily logfile samples were randomly assigned in batches over the course of the study using these proportions.

Periodic reliability tests were conducted for TNW and total utterances each time a batch of daily logfile samples was analyzed. These global measures were selected for reliability testing because they reflect the overall consistency of transcripts between raters. Both of these measures are automatically reported by PeRT software when an analysis is completed.

A frequency of agreement ratio, *F(agree)* [39], was found for the frequency of words and utterances in each batch of daily logfile samples as a measure of interrater reliability. *F(agree)* was computed using **Equation 1**:

$$F(agree) = (\sum F(\min) / \sum F(\max) \times 100\% \text{ , } \textbf{(1)}$$

(a)

"Wish her happy birthday "

15:35:15 "w"
15:36:06 ","
15:36:57 "Backspace"
15:37:48 "¡"
15:38:39 "s"
15:39:30 "**1**"
15:39:30 "h"
15:39:30 " "
15:40:21 "h"
15:41:12 "e"
15:42:03 "**4**"
15:42:03 "r"
15:42:03 " "
15:42:54 "h"
15:43:45 "a"
15:44:36 "p"
15:45:27 "P"
15:45:27 "!"
15:46:18 "Backspace"
15:47:09 "**2**"
15:47:09 "p"
15:47:09 "y"
15:47:09 " "
15:48:00 "b"
15:48:51 "**2**"
15:48:51 "¡"
15:48:51 "r"
15:48:51 "t"
15:48:51 "h"
15:48:51 "d"
15:48:51 "a"
15:48:51 "y"
15:48:51 " "
15:49:42 "Enter"

(b)

"Write this down for [NAME] "

17:08:45 "w"
17:09:36 "r"
17:10:27 "**1**"
17:10:27 "¡"
17:10:27 "t"
17:10:27 "e"
17:10:27 " "
17:11:18 "t"
17:12:09 "**3**"
17:12:09 "h"
17:12:09 "¡"
17:12:09 "s"
17:12:09 " "
17:13:00 "d"
17:13:51 "o"
17:14:42 "w"
17:15:33 "**1**"
17:15:33 "n"
17:15:33 " "
17:16:24 "f"
17:17:15 "3"
17:17:15 "o"
17:17:15 "r"
17:17:15 " "
17:24:03 "[NAME] "

(c)

"**********"

12:08:01 "PRIVON"
12:08:22 "*"
12:08:43 "*"
12:09:04 "*"
12:09:25 "*"
12:09:46 "*"
12:10:07 "*"
12:10:28 "*"
12:10:28 "*"
12:10:28 "*"
12:10:28 "*"
12:10:28 "*"
12:10:28 "*"
12:10:28 "*"
12:10:49 "*****"
12:11:10 "*****"
12:11:10 "*"
12:11:10 "*"
12:11:10 "*"
12:11:10 "*****"
12:11:10 "*"
12:11:10 "*"
12:11:31 "*"
12:11:52 "*"
12:12:13 "*"
12:12:34 "*"
12:12:55 "*"
12:13:16 "*****"
12:13:37 "*"
12:13:58 "*"
12:14:19 "*"
12:14:40 "*"

Figure 2.

(a)–**(b)** Examples of complete utterances generated by subject using WordPad program on brain-computer interface (BCI) device and **(c)** block of text generated by subject using BCI device in Privacy Mode. **(a)** First utterance contains some self-corrected errors. **(b)** Second utterance is deidentified, with reported time stamp showing last keystroke of [NAME] that was generated. Numbers shown in bold text indicate that subject used word-prediction feature to complete current word.

where *F*(min) is the lowest reported value for TNW or total utterances in a daily logfile sample and *F*(max) is the highest reported value for TNW or total utterances from the same daily logfile sample when these values are reported by two independent raters. *F(agree)* is weighted so that each word or utterance is weighted equally, regardless of the number of words or utterances in the individual daily logfile samples.

Fidelity of transcription procedures was maintained by quickly resolving discrepancies through an interjudge agreement process. After interrater reliability was obtained using *F(Agree)*, 100 percent interjudge agreement was achieved for TNW and total utterances. The raters identified and resolved discrepancies on an as-needed basis to achieve a consensus on the correct transcription of each utterance based on the operational guidelines provided in the procedural manual. Transcripts were adjusted accordingly as needed.

The number of daily logfile samples varied over the course of the study because the number of subjects actively using their BCI devices at any given time varied. A total of 345 daily logfile samples was generated during the first 60 wk of data collection. A total of 68 daily logfile samples from this time period was randomly selected and analyzed by two different raters for reliability testing.

RESULTS

F(agree) pooled across the first 60 wk of data collection was 97.6 percent for TNW and 93.5 percent for total utterances. Interjudge agreement was 100 percent for TNW and total utterances. The **Table** summarizes the number of daily logfile samples and corresponding reliability measures. The number of daily logfile samples and corresponding reliability measures are presented for each of the first five 12 wk periods of data collection (60 wk total).

An error analysis was conducted to characterize the nature of the discrepancies between raters for total utterances. Three main patterns accounted for 84 percent of the discrepancies between raters. The two most common patterns each accounted for 31 percent of the discrepancies. In one pattern, one rater reported a one-word utterance while the second rater included that same word as part of a longer utterance. The second pattern involved errors parsing complex sentences according to operational definitions. Another 23 percent of the discrepancies were related to errors parsing unintelligible strings according to operational definitions. All these discrepancies were easily resolved by establishing interjudge agreement.

DISCUSSION

Adequate resources and sufficient time to prepare and test transcription procedures during the start-up phase of the research were central to our ability to achieve high reliability results. Writing the transcription procedural manual required installing and testing KeyLAM and generating non-research BCI logfiles for practicing analysis. Experience analyzing BCI logfiles with PeRT was needed to verify that our formatted data were compatible with our analysis software. The procedural manual required two revisions: one after we examined our first experience with the full process and the second after we completed our intrarater reliability testing to add detailed operational guidelines for handling more complex language samples, e.g., strings of errors, identifying

Table.
Summary of daily logfile samples and reliability measures for 60 wk of data collection.

Period	Samples(N)	Samples(R)	Total Utterances (%)		TNW (%)	
			F(agree)	Interjudge	*F(agree)*	Interjudge
1	67	13	97.4	100	96.6	100
2	65	12	90.9	100	95.4	100
3	25	6	100.0	100	100.0	100
4	107	22	89.5	100	98.3	100
5	81	15	96.4	100	97.7	100
Total	345	68	93.5	100	97.6	100

F(agree) = frequency of agreement ratio, Interjudge = interjudge agreement, Period = 12 consecutive weeks of data collection, Samples(N) = number of daily logfile samples, Samples(R) = number of samples analyzed by second rater for reliability testing, TNW = total number of words.

utterances when sentence terminators were selected in error. Transcribers were required to achieve intrarater reliability of 90 percent before handling research data. During this stage, any interrater discrepancies were brought to the team for discussion and a determination of how the discrepancy should be transcribed. Thus, the team achieved 100 percent interjudge agreement on practice samples before starting the transcription of the study's logfiles. This attention to detail is highly recommended as research and clinical teams start to use these tools more widely in the future.

This study demonstrates that implementation of LAM and PeRT tools provides an efficient and effective approach to language sample analysis to report language performance measures. Based on our experience with language sample transcription and previous studies reporting difficulty with audio and video transcription [20,40–41], LAM-related tools offer improved efficiency and effectiveness over traditional methods of audio and video recording. To date, transcription and analysis of LAM-formatted language sample data have been completed more than 400 times by a team of three raters. Studies using traditional methods of audio and video transcription and analysis involve smaller numbers of language samples because the transcription process is labor intensive. The study also involved the video recording of language samples of participants using other alternative communication methods, e.g., another speech-generating device (SGD), manual communication boards, or natural speech. These video recordings required much more time to transcribe a few minutes of communication. Many of these video recordings failed to capture the alternative communication methods with enough detail to identify and describe the strategies used and effectively analyze performance. Consequently, we believe that any research investigating SGD use and performance among veterans who cannot use natural speech should incorporate collection of LAM logfiles for reporting results of language performance measures as dependent variables.

Periodic reliability tests were used to maintain fidelity of transcription procedures and quickly resolve questions related to transcription and analysis that arose over the course of the study before systematic discrepancies between raters developed. For example, interrater reliability began to decline during period 4, shortly after two new subjects began to generate a variety of complex sentence types that had not been observed in earlier samples. These subjects began to use their BCI devices at approximately the same time during the 38th and 39th weeks of data collection. When periodic reliability testing revealed lower interrater reliability for total utterances, the raters reviewed the data and identified a systematic pattern of discrepancies between raters related to transcription of complex sentences in samples generated by these two subjects. A review of operational guidelines for C-unit segmentation resolved these systematic discrepancies so that high reliability could be maintained in future analyses.

Crucially, high reliability was maintained for transcription of decontextualized LAM data. Our operational guidelines followed established procedures for utterance segmentation that were based on syntactic features of the language events in the LAM data. We may not have been able to maintain high reliability using guidelines based on pragmatic roles of utterances or other features that are based on conversational context.

Limitations to our study are based on two factors. First, KeyLAM was installed as an add-on application to the BCI system and not as an integrated or built-in LAM data logging feature. KeyLAM was installed as a separate program running simultaneously with the BCI during the use of the email and WordPad programs. This created avoidable analysis problems. Several commercially available high performance SGDs have LAM integrated with the communication software so that the logfile does not require a labor-intensive reformatting process before PeRT can be used for transcription and analysis. Our transcription process required idiosyncratic additional steps in order to correctly identify word-prediction use. An integrated LAM feature would have saved time and allowed for increased efficiency.

The second limitation related to the 2 wk delay we experienced in receiving the daily language samples. This delay prevented us from identifying potential problems and providing timely feedback about BCI use. Our laboratory's transcribers were speech-language pathologists who were able to observe changes in language sample data within subjects. The PeRT results provided us with insights into overall BCI communication performance that could have been used more routinely to support the study's participants.

This study provides insights into potential future research using software tools to support transcription and analysis of SGD performance and outcomes. Based on the usefulness of logfiles for measuring language performance, BCI programmers should consider installing the LAM feature during initial system development phases to

improve efficiency and utility. In addition to the universal LAM format, an optional enhanced LAM format provides for the recording of a sequence of keystroke events leading to an output. This enhanced logfile format will be of value when BCI systems advance to the stage of offering fully functional language application programs that offer multiple encoding methods to represent and generate language. Providing more widespread application of data logging and logfile analysis tools will promote the increased collection of language samples and reporting of performance measures along with outcome measures. While outcome measures support perceptions of effectiveness, performance measures quantify the actual results of using AAC systems to communicate in a range of contexts.

CONCLUSIONS

As BCIs become a consideration as a promising AAC system for individuals with ALS, consistent language sample transcription and analysis procedures are important for reliability and validity of performance results. For the purpose of documenting the reliability of transcription procedures using BCIs for daily communication, we calculated primary communication measures for 15 subjects with ALS while they were using the email or WordPad programs. With the transcription of 345 decontextualized daily logfile samples, three raters showed consistent measures for the words and utterances for point-by-point reliability. These results indicate that the implementation of data logging and logfile analysis tools such as LAM and PeRT provide for an efficient and effective approach to the analysis of BCI language samples for reporting language performance measures. Finally, our study showed that the established transcription procedures resulted in high reliability and fidelity of the process, which supports valid results for the final study reporting.

ACKNOWLEDGMENTS

Author Contributions:
Study concept and design: K. Hill.
Analysis and interpretation of data: K. Hill, T. Kovacs, S. Shin.
Drafting of manuscript: K. Hill, T. Kovacs.
Statistical analysis: K. Hill, T. Kovacs, S. Shin.
Supervision of research: K. Hill.

Financial Disclosures: The authors have declared that no competing interests exist.
Funding/Support: This material was based on work supported by VA Cooperative Studies Program (CSP) 567, Office of Research and Development.
Additional Contributions: The BCI 2000 was developed at the Wadsworth Center, Albany, New York, under the direction of Jonathan Wolpaw, MD, and used in this study. The authors wish to acknowledge and thank the VA CSP 567 study group identified in the **Appendix** (available online only).
Institutional Review: Informed consent for VA CSP 567 was approved through the Centralized Institutional Review Board.
Participant Follow-Up: The authors do not plan to inform participants of the publication of this study because participant data were de-indentified.

REFERENCES

1. Hill K. Advances in augmentative and alternative communication as quality-of-life technology. Phys Med Rehabil Clin N Am. 2010;21(1):43–58. [PMID:19951777] http://dx.doi.org/10.1016/j.pmr.2009.07.007
2. Schlosser RW, Raghavendra P. Evidence-based practice in augmentative and alternative communication. Augment Altern Commun. 2004;20(1):1–21. http://dx.doi.org/10.1080/07434610310001621083
3. Wolpaw JR, Wolpaw EW. Brain-computer interfaces: Something new under the sun. In: Wolpaw JR, Wolpaw EW, editors. Brain-computer interfaces. New York (NY): Oxford University Press; 2012. p. 3–12.
4. Fabiani M, Gratton G, Karis D, Donchin E. Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. Adv Psychophysiol. 1987;2:1–78.
5. Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr Clin Neurophysiol. 1988; 70(6):510–23. [PMID:2461285] http://dx.doi.org/10.1016/0013-4694(88)90149-6
6. Guenther FH, Brumberg JS, Wright EJ, Nieto-Castanon A, Tourville JA, Panko M, Law R, Siebert SA, Bartels JL, Andreasen DS, Ehirim P, Mao H, Kennedy PR. A wireless brain-machine interface for real-time speech synthesis. PLoS ONE. 2009;4(12):e8218. [PMID:20011034] http://dx.doi.org/10.1371/journal.pone.0008218
7. Sheikh H, McFarland DJ, Sarnacki WA, Wolpaw JR. Electroencephalographic(EEG)-based communication: EEG control versus system performance in humans. Neurosci Lett. 2003;345(2):89–92. [PMID:12821178] http://dx.doi.org/10.1016/S0304-3940(03)00470-1
8. Wolpaw JR, Ramoser H, McFarland DJ, Pfurtscheller G. EEG-based communication: Improved accuracy by response

verification. IEEE Trans Rehabil Eng. 1998;6(3):326–33. [PMID:9749910] http://dx.doi.org/10.1109/86.712231

9. Miner LA, McFarland DJ, Wolpaw JR. Answering questions with an electroencephalogram-based brain-computer interface. Arch Phys Med Rehabil. 1998;79(9):1029–33. [PMID:9749678] http://dx.doi.org/10.1016/S0003-9993(98)90165-4
10. Ahi ST, Kambara H, Koike Y. A dictionary-driven P300 speller with a modified interface. IEEE Trans Neural Syst Rehabil Eng. 2011;19(1):6–14. [PMID:20457551] http://dx.doi.org/10.1109/TNSRE.2010.2049373
11. Ryan DB, Frye GE, Townsend G, Berry DR, Mesa-G S, Gates NA, Sellers EW. Predictive spelling with a P300-based brain-computer interface: Increasing the rate of communication. Int J Hum Comput Interact. 2011;27(1):69–84. [PMID:21278858] http://dx.doi.org/10.1080/10447318.2011.535754
12. Acqualagna L, Treder MS, Schreuder M, Blankertz B. A novel brain-computer interface based on the rapid serial visual presentation paradigm. Conf Proc IEEE Eng Med Med Biol Soc. 2010;2010:2686–89. [PMID:21096199]
13. Blankertz B, Schmidt NM, Treder MS. Gaze-independent BCI spellers based on covert attention and feature attention. 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL); 2010 Nov 7–10; Rome, Italy. New York (NY): IEEE; 2010. p. 1–2.
14. Schaeff S, Treder MS, Venthur B, Blankertz B. Exploring motion VEPs for gaze-independent communication. J Neural Eng. 2012;9(4):045006. [PMID:22832017] http://dx.doi.org/10.1088/1741-2560/9/4/045006
15. Krusienski DJ, Sellers EW, Cabestaing F, Bayoudh S, McFarland DJ, Vaughan TM, Wolpaw JR. A comparison of classification techniques for the P300 Speller. J Neural Eng. 2006;3(4):299–305. [PMID:17124334] http://dx.doi.org/10.1088/1741-2560/3/4/007
16. Sellers EW, Krusienski DJ, McFarland DJ, Vaughan TM, Wolpaw JR. A P300 event-related potential brain-computer interface (BCI): The effects of matrix size and inter stimulus interval on performance. Biol Psychol. 2006;73(3): 242–52. [PMID:16860920] http://dx.doi.org/10.1016/j.biopsycho.2006.04.007
17. Bach JR. Amyotrophic lateral sclerosis. Communication status and ventilatory support. Am J Phys Med Rehabil. 1993;72:343–49. [PMID:8260126] http://dx.doi.org/10.1097/00002060-199312000-00002
18. Murphy J. Communication strategies of people with ALS and their partners. Amyotroph Lateral Scler. 2004;5(2): 121–26. http://dx.doi.org/10.1080/14660820410020411
19. Heilmann JJ, Miller JF, Nockerts A. Using language sample databases. Lang Speech Hear Serv Sch. 2010;41(1):84–95. [PMID:20051580] http://dx.doi.org/10.1044/0161-1461(2009/08-0075)
20. Price LH, Hendricks S, Cook C. Incorporating computer-aided language sample analysis into clinical practice. Lang Speech Hear Serv Sch. 2010;41(2):206–22. [PMID:19755638] http://dx.doi.org/10.1044/0161-1461(2009/08-0054)
21. Hill K, Corsi V. The role speech language pathologists in assistive technology assessments. In: Scherer MJ, Federici S, editors. Assistive technology assessment: A handbook for professionals in disability, rehabilitation and health professions. London (England): Taylor & Francis Group; 2012. p. 301–36.
22. Nicholas LE, Brookshire RH. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. J Speech Hear Res. 1993;36(2):338–50. [PMID:8487525]
23. Hula W, McNeil M, Doyle P, Rubinsky H, Fossett T. The inter-rater reliability of the story retell procedure. Aphasiology. 2003;17(5):523–28. http://dx.doi.org/10.1080/02687030344000139
24. Miller J. Assessing language production in children: Experimental procedures. Needham Heights (MA): Allyn and Bacon; 1981. 186 p.
25. Portney LG, Watkins MP. Foundations of clinical research: Applications to practice. 3rd ed. Upper Saddle River (NJ): Prentice-Hall; 2009. 912 p.
26. Hill K. The development of a model for automated performance measurement and the establishment of performance indices for augmented communicators under two sampling conditions. Dissertation Abstr Int. 2001;60(5):2293.
27. Hill K. A case study model for augmentative and alternative communication. Assist Technol Outcomes Benefits. 2006;3(1):53–66.
28. Romski M, Sevcik RA, Adamson LB, Cheslock MA, Smith A, Barker RM, Bakeman R. Randomized comparison of augmented and nonaugmented language interventions for toddlers with developmental delays and their parents. J Speech Lang Hear Res. 2010;53(2):350–64. [PMID:20360461] http://dx.doi.org/10.1044/1092-4388(2009/08-0156)
29. Hill KJ, Romich BA. A language activity monitor for supporting AAC evidence-based clinical practice. Assist Technol. 2001;13(1):12–22. [PMID:12212432] http://dx.doi.org/10.1080/10400435.2001.10132030
30. Smith LE, Higginbotham DJ, Lesher GW, Moulton B, Mathy P. The development of an automated method for analyzing communication rate in augmentative and alternative communication. Assist Technol. 2006;18(1):107–21. [PMID:16796245] http://dx.doi.org/10.1080/10400435.2006.10131910

31. Beukelman DR, Mirenda P. Augmentative and alternative communication: Supporting children and adults with complex communication needs. 4th ed. Baltimore (MD): Paul H. Brookes Publishing Co; 2013. 616 p.
32. Hill K. AAC evidence-based practice and language activity monitoring. Top Lang Disord. 2004;24(1):18–30. http://dx.doi.org/10.1097/00011363-200401000-00004
33. Hill K, Romich B. AAC language monitoring and analysis for clinical intervention and research outcomes. Proceedings of 14th Annual Technology for Persons with Disabilities Conference, California State University, Northridge; 1999 Mar 16–20; Northridge CA.
34. Cook S, Hill K. AAC performance data for an individual with amyotrophic lateral sclerosis. Proceedings of RESNA 26th Annual Conference; 2003 Jun 19–23; Atlanta, Georgia. 2003; Arlington (VA): RESNA.
35. Hill K, Jans D. AAC outcomes and persons with ALS/MND. Proceedings of 11th Biennial ISAAC Conference; 2004 Oct 4–12; Natal, Brazil. Toronto (Canada): ISAAC; 2004.
36. Brooks BR. El Escorial World Federation of Neurology criteria for the diagnosis of amyotrophic lateral sclerosis. Subcommittee on Motor Neuron Diseases/Amyotrophic Lateral Sclerosis of the World Federation of Neurology Research Group on Neuromuscular Diseases and the El Escorial "Clinical limits of amyotrophic lateral sclerosis" workshop contributors. J Neurol Sci. 1994;124(Suppl):96–107. [PMID:7807156] http://dx.doi.org/10.1016/0022-510X(94)90191-0
37. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, Nakanishi A; BDNF ALS Study Group (Phase III). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. J Neurol Sci. 1999;169(1–2):13–21. [PMID:10540002] http://dx.doi.org/10.1016/S0022-510X(99)00210-5
38. Loban W. Language development: Kindergarten through grade twelve. Urbana (IL): National Council of Teachers of English; 1976. 144 p.
39. Hegde MN. Clinical research in communicative disorders: Principles and strategies. 3rd ed. Austin (TX): PRO-ED, Inc; 2003. 585 p.
40. Togher L. Discourse sampling in the 21st century. J Commun Disord. 2001;34(1–2):131–50. [PMID:11322564] http://dx.doi.org/10.1016/S0021-9924(00)00045-9
41. Heilmann J, Nockerts A, Miller JF. Language sampling: Does the length of the transcript matter? Lang Speech Hear Serv Sch. 2010;41(4):393–404. [PMID:20601531] http://dx.doi.org/10.1044/0161-1461(2009/09-0023)

Submitted for publication May 1, 2013. Accepted in revised form December 16, 2013.

This article and any supplementary material should be cited as follows:
Hill K, Kovacs T, Shin S. Reliability of brain-computer interface language sample transcription procedures. J Rehabil Res Dev. 2014;51(4):579–90.
http://dx.doi.org/10.1682/JRRD.2013.05.0102