Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Research review

Systematic review: online crowdsourcing to assess perceptual speech outcomes

Anne M. Sescleifer, BS,[a] *Caitlin A. Francoisse, MD,*[a] *and Alexander Y. Lin, MD*[a,b,*]

[a] *Division of Plastic Surgery, Saint Louis University School of Medicine, St. Louis, Missouri*
[b] *SSM Health Cardinal Glennon Children's Hospital at SLU, St. Louis, Missouri*

ARTICLE INFO

Article history:
Received 27 February 2018
Received in revised form
27 February 2018
Accepted 11 June 2018
Available online xxx

Keywords:
Speech
Outcomes
Crowdsourcing
Systematic review

ABSTRACT

Background: Speech is integral for human interaction and development. Speech assessments are critical in the growing child, especially in the surgical evaluation of patients undergoing cleft palate and speech surgeries. Online crowdsourcing enables layperson raters, allowing rapid and large-scale data collection. This systematic review analyzes the utility of online crowdsourcing to evaluate perceptual speech outcomes.

Methods: Terms related to "crowdsourcing" and "speech" were searched on PubMed, Scopus, CINAHL, Cochrane CENTRAL, and PsycINFO on August 16, 2017, returning 2812 unique articles. Inclusion and exclusion criteria concentrated on online crowdsourcing of perceptual speech outcomes: titles led to 140 abstracts that yielded 35 full-text articles, of which eight articles met criteria for analysis.

Results: All studies used Amazon Mechanical Turk for online crowd raters, and one used an additional crowdsourcing site (CrowdFlower). Disordered speech was provided by 376 speakers, for which 2203 crowd workers produced over 700,000 unique ratings. Five studies compared crowdsourced assessments to gold standards and found high concordances. Data collection time ranged from 59 min to 23 h, with worker payments ranging from $0.05 to $2.00 per task. Studies examined child pronunciation of the /r/ sound, dysarthria in Parkinson's speech, and articulation of English words produced by non-English speakers learning English.

Conclusions: Online crowdsourcing for perceptual speech outcomes provides high-quality data consistent with previous speech-assessment standards in a rapid, cost-effective manner. This novel methodology incorporates lay perspective of speech intelligibility and has the potential to revolutionize surgical speech outcome assessments, including cleft palate and speech surgery.

Portions of this research were accepted for abstract presentation at (1) American Cleft Palate-Craniofacial Association (ACPA) 75th Annual Meeting, Pittsburgh, PA, April 10-14, 2018 and (2) Plastic Surgery Research Council (PSRC) 63rd Annual Meeting, Birmingham, AL, May 17-20, 2018.

* *Corresponding author.* Division of Plastic Surgery, Associate Professor and Section Chief of Pediatric Plastic Surgery, Saint Louis University School of Medicine, St. Louis Cleft-Craniofacial Center, SSM Health Cardinal Glennon Children's Hospital at SLU, 1465 South Grand Boulevard, St. Louis, MO 63104-1095. Tel.: +314 678 2171; fax: +1 314 678 2172.
E-mail address: alexander.lin@health.slu.edu (A.Y. Lin).

0022-4804/$ – see front matter
https://doi.org/10.1016/j.jss.2018.06.032

Introduction

Normal intelligible speech is essential for human interaction and normal pediatric development, which is why perceptual speech analysis is critical to assess and treat the child born with cleft palate and to monitor the postoperative course after cleft palate and speech surgery. Clinical assessment of speech is usually performed by the speech language pathologist (SLP) as part of the team care of these patients.[1] It is widely recognized that the demand for SLPs exceeds their availability,[2] and the high cost and limited accessibility of speech experts can serve as a barrier to care and research for patients with cleft palate. Online crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk), offer a new alternative to assess speech outcomes in both clinical and research settings. MTurk works as an online marketplace in which researchers post tasks that laypeople complete for online payments, and therefore, researchers can amass large numbers of raters.

Crowdsourcing refers to large-scale online data collection from lay raters. Several studies have examined the reliability of MTurk as a research instrument and found it to be a source of rapid, low-cost, and high-quality data.[3,4] In the past, crowdsourcing has been validated for use in a wide range of medically related capacities, including evaluation of surgical skill,[5] assessment of pictograph recognition for hospital discharge instructions,[6] quantification of malaria parasites,[7] and classification of histologic findings, including colonic polyps and immunohistochemical stains of breast cancer specimens.[8,9] In addition, crowdsourcing has been validated in a variety of speech tasks, from the collection and transcription of speech samples[10,11] to the assessment of speech perception and intelligibility.

To better understand the potential for crowdsourced evaluations to aid in speech assessment of patients with cleft palate and other speech disorders, a systematic review of the literature was conducted. The aim of this systematic review is to identify the current uses, costs, and validity of online crowdsourcing to assess perceptual speech outcomes.

Methods

The medical literature published in five databases (MEDLINE, Scopus, PsycINFO, CINAHL, and COCHRANE Central) was searched for articles that included terms relating to "crowdsourcing" and "speech." The Medical Subject Headings terms "crowdsourcing" and "speech" were included, as well as manual wildcard asterisked terms to systematically review available literature. A gray literature search on ClinicalTrials.gov did not return any results.

This full database search was completed on August 16, 2017. Results were combined, and duplicates were removed. The search delivered 2812 unique articles. A title weed was conducted and 2672 articles were eliminated based on inclusion and exclusion criteria. An abstract weed was conducted by two independent reviewers on the 140 remaining articles. Kappa statistic showed very good correlation (Cohen's kappa = 0.92), excluding 105 articles. The full text of the remaining 35 articles was obtained, and eight articles were deemed eligible for inclusion in this systematic review. The attrition flowchart is shown in Figure.

The clinical focus of this review was online crowdsourcing to assess perceptual speech outcomes; thus the inclusion criteria were as follows: (1) articles must report primary data involving an assessment of human speech, and (2) articles must use online crowdsourcing to evaluate perceptual speech outcomes (including speech intelligibility, pronunciation, articulation, and nasality). For the purposes of this study, crowdsourcing was defined as any study that uses laypeople as raters. Both case studies and case series were included. Exclusion criteria were as follows: (1) reviews, technique articles that did not actually analyze speech, or editorials, (2) articles that did not use crowdsourced ratings, (3) articles that did not include primary data on perceptual speech outcomes (e.g., natural language processing studies; articles analyzing singing, vocal tone, or emotion; studies investigating pragmatics or semantics; or use of an interpreter in clinical practice), (4) articles analyzing nonhuman speech (such as robots, cars, and Siri), (5) assessments conducted exclusively in traditional (not online) laboratory settings, (6) articles assessing the speech rate and rhythm, and (7) transcription tasks of nondisordered speech.

The following data were extracted when explicitly stated by the article: speaker characteristics (Table 1; speech disorder, number of speakers, number of samples per speaker, speaker gender, speaker age range, follow-up time), online survey characteristics (Table 2; platform, description of task, number of unique tasks, task length, duration of

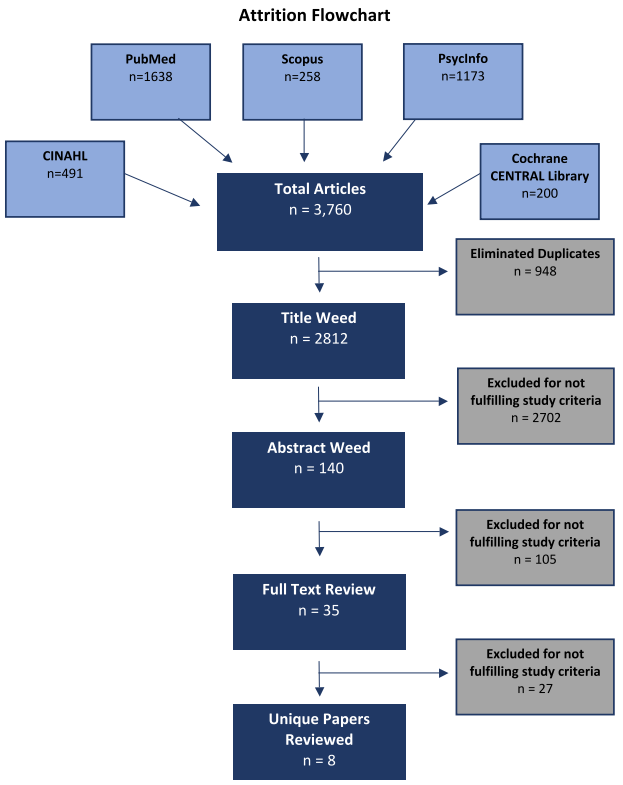

Fig – Attrition flowchart showing the algorithmic elimination of studies through inclusion and exclusion criteria. (Color version of figure is available online.)