JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0001941

Rapid adaptation to non-native speech is impaired in cochlear implant users

Michelle R. Kapolowicz,[1,a)] Vahid Montazeri,[2] Melissa M. Baese-Berk,[3] Fan-Gang Zeng,[1,b)] and Peter F. Assmann[2]

[1]*Center for Hearing Research, University of California, Irvine, California 92697, USA*
[2]*School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas 75080, USA*
[3]*Department of Linguistics, University of Oregon, Eugene, Oregon 97403, USA*
mkapolow@uci.edu, vmontazery@gmail.com, mbaesebe@uoregon.edu, fzeng@hs.uci.edu, assmann@utdallas.edu

Abstract: To examine difficulties experienced by cochlear implant (CI) users when perceiving non-native speech, intelligibility of non-native speech was compared in conditions with single and multiple alternating talkers. Compared to listeners with normal hearing, no rapid talker-dependent adaptation was observed and performance was approximately 40% lower for CI users following increased exposure in both talker conditions. Results suggest that lower performance for CI users may stem from combined effects of limited spectral resolution, which diminishes perceptible differences across accents, and limited access to talker-specific acoustic features of speech, which reduces the ability to adapt to non-native speech in a talker-dependent manner. https://doi.org/10.1121/10.0001941

[Editor: Martin Cooke] Pages: EL267–EL272

Received: 29 June 2020 **Accepted:** 21 August 2020 **Published Online:** 16 September 2020

1. Introduction

Compared to native-accented speech, non-native speech is less intelligible and requires greater listening effort due to segmental and suprasegmental patterns that differ from those listeners are accustomed to in their native language.[1] Listeners can partially overcome this difficulty with increased exposure[2,3] even after just a few sentences.[4] Accent-dependent adaptation is one mechanism shown to facilitate this process, whereby listeners utilize similar acoustic cues across multiple non-native talkers who share the same native language.[3] For example, when native talkers of English are speaking French as a second language, since English does not have the high vowel /y/, they tend to substitute /u/ in place of /y/.[6] Talker-dependent adaptation is another mechanism that occurs after increased exposure to an individual talker, whereby listeners utilize talker-specific cues, such as the fundamental frequency (F0, related to glottal pulse rate) and the spectral envelope (related to vocal tract length), to aid with adaptation.[7,8]

Few studies have characterized the added difficulties that cochlear implant (CI) users experience when attending to non-native speech compared to (NH) listeners. Ji *et al.* found that CI users have poorer than normal speech reception thresholds in non-native speech in noise perception compared to NH listeners, with performance being strongly correlated to ratings of intelligibility and talker accentedness for CI users.[9] The same study also reported greater inter-subject variability for CI users perceiving non-native speech compared to native speech.[9] Tamati and Pisoni found that CI users are less sensitive to differences across accents compared to NH listeners in a task where listeners judged perceived intelligibility of native and non-native speech.[10] To date, only one study has reported whether CI users can overcome perceptual deficits with non-native speech. Waddington *et al.* found that CI users could utilize audiovisual cues to aid with non-native speech perception, though not to the level observed for native speech.[11] The same study found that older CI users were at an increased disadvantage when perceiving non-native speech relative to their younger counterparts, though they could partially overcome the age-related deficit with access to audiovisual cues.[11]

The extent to which CI users can rapidly adapt to non-native speech and how this process may be similar to or different from NH listeners is unknown. CI users have relatively normal temporal envelope processing but have limited access to fine structure cues such as mean F0

[a)]Author to whom correspondence should be addressed. Also at: Department of Otolaryngology—Head and Neck Surgery, University of California, Irvine, CA, USA.
[b)]Also at: Department of Otolaryngology—Head and Neck Surgery, University of California, Irvine, CA, USA.

and spectral envelope,[12] which lead to reduced talker discrimination.[13] These robust voice characteristics may aid NH listeners with perceptual recalibration when adapting to non-native speech in a talker-dependent manner. Although these cues are limited for CI users, they are not entirely absent,[14] and CI users can partially discriminate talkers, individual voice cues, and can learn talkers' voices in native-accented speech.[15] This entails that CI users can obtain some talker-dependent benefit; however, when speech production patterns drastically vary, such as with non-native speech, having only limited access to these talker-specific cues may not suffice. CI users may find it more reliable to utilize other systematic cues in non-native speech that are accessible via temporal envelope processing, such as a reduction of unstressed syllables.[5] This latter mode implies that CI users may benefit more from an accent-dependent mechanism, where stable cues inherent to the accent can be accessed via exposure to several different non-native talkers who share the same native language.[3] But even this method may be limited for CI users regardless of whether certain accent-dependent cues are more aptly transmitted through their device. Specifically, in an easier listening condition such as with native speech, CI users already exhibit a detriment when perceiving speech from several different talkers compared to only a single talker.[16] This suggests that perceiving non-native speech from multiple talkers may be even more difficult since they would need to simultaneously resolve sources of variability in vocal characteristics stemming from different talkers while also reconciling with increased segmental and suprasegmental variability occurring in non-native speech production.

Given that it is unclear whether CI users utilize similar mechanisms as NH listeners to adapt to non-native speech, the present work examined whether listeners exhibit rapid adaptation (defined as an improvement in intelligibility performance with increased exposure) to non-native speech by utilizing either talker-dependent or accent-dependent mechanisms. Listeners were divided into in either single- or multiple-talker conditions. In single-talker conditions, listeners were exposed to speech from the same non-native talker to allow for an assessment of whether listeners could adapt in a talker-dependent manner. In multiple-talker conditions, listeners were exposed to five different interleaved non-native talkers who share the same native language in order to assess abilities to adapt in an accent-dependent manner. Listeners were further divided into three groups: NH listeners perceiving unprocessed sentences, NH listeners perceiving nine-channel vocoded sentences, and CI users perceiving sentences through their implant device. Vocoded conditions were included because talker-specific cues, such as F0 and spectral envelope, are neither well-encoded in CI devices nor in vocoded speech with decreased spectral resolution.[17,18] This provided an assessment of rapid adaptation without additional potential confounds that can vary across CI users, such as the number of active electrodes.

Based on previous results using the same paradigm,[8] it was hypothesized that NH listeners would rapidly adapt to non-native speech in the unprocessed single-talker condition but not in the unprocessed multiple-talker condition, providing evidence that a talker-dependent mechanism aids with rapid adaptation to non-native speech to a greater extent than accent-dependent adaptation for NH listeners. It was also hypothesized that when talker-specific voice cues are limited, as with vocoded speech or with a CI, this talker-dependent adaptation effect would be impaired. In such cases, it was considered whether CI users and listeners in the vocoded condition could benefit in the multiple-talker condition by utilizing systematic accent-dependent cues to aid with adaptation.

2. Method

2.1 Listeners

Thirty-six monolingual ($n = 6$ per listening condition), native talkers of American English participated (NH: $n = 24$, age range: 18–39 years, $M = 23$ years; CI users: $n = 12$, age range: 23–73 years, $M = 53$ years). NH participants reported having no hearing impairments and passed a hearing screening at 20 dB hearing level at octave frequencies from 250 to 8000 Hz. Additional information regarding CI participants can be found in supplementary material.[19] Participants were monetarily compensated for participation.

2.2 Stimuli

Phonetically balanced low-context Harvard sentences (e.g., "The ripe taste of cheese improves with age") recorded from five non-native American English talkers (3 females, 2 males) were root-mean-squared equalized and presented for the experiments. Talkers' accentedness and intelligibility were classified in quiet by 22 NH native talkers of American English who did not participate in this study. Non-native talkers were Mandarin-accented, had resided in Taiwan and the United States, and rated as being relatively heavily non-native-accented ($M = 7$, $SD = 0.66$) using a nine-point Likert scale where 9 corresponded to "heavily non-native-accented." Intelligibility was scored as the percent of correctly typed key words (key words = words other than article

https://doi.org/10.1121/10.0001941

adjectives) to total key words using a customized automatic scoring program created in MATLAB (The MathWorks USA). The program also accounted for commonly misspelled words and homonyms. Non-native talkers' intelligibility scores ranged from 58% to 74% ($M = 68\%$, $SD = 7.0$). Sentences from an additional female native-accented talker (accent rating: $M = 1$, $SD = 0$; intelligibility: $M = 97\%$, $SD = 0.55$) were used for task familiarization. For vocoded conditions, spectral resolution was limited with a nine-channel tone vocoder that included low-pass filtering at 160 Hz and full-wave rectification.[20] Nine channels were chosen for being similar to the effective number of channels when perceiving speech with a CI.[21]

2.3 Group assignment

The experiment was divided into single- and multiple-talker conditions. In *single-talker conditions*, sentences from one of the five talkers in the multiple-talker conditions were randomly selected for each listener, and that same talker was heard for the duration of the experiment. In *multiple-talker conditions*, listeners heard speech from five non-native talkers, sequentially (i.e., talkers were presented one at a time). Two sentences per talker were randomly presented every ten sentences for the multiple-talker conditions. For each of the two talker conditions, listeners were assigned to one of three groups (unprocessed, vocoded, CI). For the *unprocessed* group, NH listeners heard unprocessed stimuli in either the single-talker ($n = 6$, age range: 18–39) or the multiple-talker ($n = 6$, age range: 18–28) condition. For the *vocoded* group, NH listeners heard vocoded stimuli in either the single-talker ($n = 6$, age range: 18–27) or the multiple-talker ($n = 6$, age range: 18–30) condition. For the *CI* group, CI users heard stimuli processed through their devices in either the single-talker ($n = 6$, age range: 23–71) or the multiple-talker condition ($n = 6$, age range: 23–73). Listeners were only assigned to one of the six conditions to control for talker adaptation.

2.4 Procedure

Stimuli were presented in a sound-attenuating booth through a speaker (Grason-Stadler, Inc., Eden Prairie, MN) located 1 m away and directly in front of the listener. The presentation level was constant for NH listeners [72 dB sound pressure level (SPL)] but presented at a comfortable level for each individual CI user [varying from 62 to 83 dB SPL ($M = 72$, $SD = 5.79$)]. CI users were instructed to use their preferred clinical MAP setting and to insert an ear plug into the less dominant implanted ear in bilateral users or into the non-implanted ear in unilateral users to reduce possible residual hearing. NH listeners were instructed to plug their right ears. No check for residual hearing in the dominant ear of CI users was performed nor were listeners screened to assure that plugging their ears prevented them from hearing sounds in the plugged ear.

Listeners were asked to type what they heard for a total of 40 unique sentences to simulate the amount of exposure that could occur over the course of a single conversation (i.e., rapid adaptation). A brief practice session comprising ten sentences spoken by a single native-accented talker was given immediately prior to the experiment. Listeners in the vocoded conditions heard vocoder-processed speech instead of unprocessed speech during the practice session. The talker included in the practice session was the same for all listeners and was not presented for the experimental session. The experiment lasted approximately 30 min. Procedures were approved by the University of California, Irvine Institutional Review Board.

Each talker condition was a 3×2 mixed design: three listening conditions (unprocessed, vocoded, CI; between-subjects) by 2 exposure stages (initial, final; within-subjects). The *initial* and *final* exposure stages correspond to mean scores for the first and last ten sentences, respectively. Adaptation entails an improvement in scores from the initial to the final exposure stage. Intelligibility was scored as the percentage of correctly typed keywords over total keywords using the same scoring program that was used to obtain baseline intelligibility scores for each talker. Scores were transformed into rationalized arcsine units (RAU) for analyses.

3. Results

3.1 Perception of non-native speech from a single talker

Figure 1(a) shows intelligibility scores as a function of listening condition across exposure stages. NH listeners in the unprocessed condition had the highest intelligibility scores ($M = 80.5$ RAU, $SD = 8.529$), followed by those in the vocoded condition ($M = 60.6$, $SD = 12.652$). CI users had the lowest scores ($M = 43.5$, $SD = 14.210$). A mixed design analysis of variance (ANOVA) indicated main effects of listening condition [$F(2,15) = 14.565$, $p < 0.001$, generalized $\eta^2 = 0.588$] and exposure stage [$F(3,45) = 5.420$, $p = 0.003$, generalized $\eta^2 = 0.087$] but no significant interaction of listening condition by exposure stage [$F(6,45) = 1.706$, $p = 0.142$, generalized $\eta^2 = 0.057$]. To test if rapid adaptation occurred after increased exposure, *post hoc* comparisons using Dunnett's tests were made. An improvement from initial exposure to final exposure was observed for the

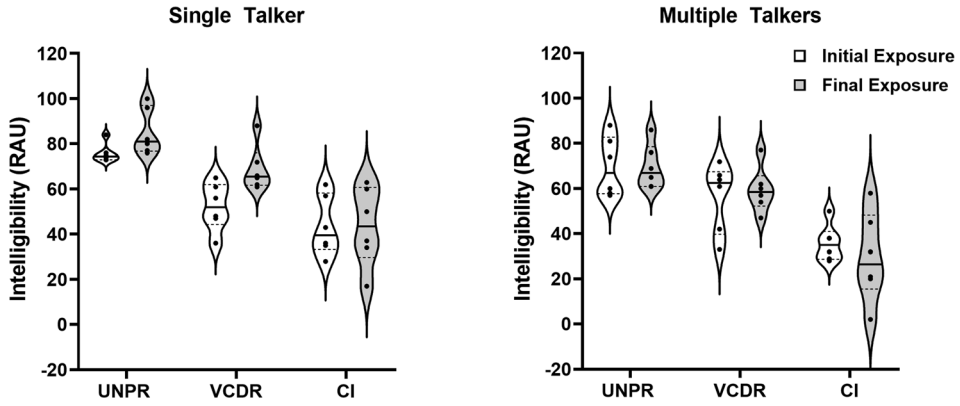

Fig. 1. Intelligibility of non-native speech in single- and multiple-talker conditions. (a) Single-talker conditions (left). The density of RAU-transformed intelligibility scores for each exposure period across listener groups is presented. Solid lines represent the median, dashed lines above indicate the third quartile, and dashed lines below indicate the first quartile. Dots represents individual data points. Abbreviations: UNPR = unprocessed, VCDR = vocoder-processed, CI = cochlear implant. (b) Multiple-talker conditions (right). (b) Follows the same conventions as (a).

unprocessed condition, $p = 0.039$, and for the vocoded condition, $p = 0.010$, but not for CI users, $p = 0.503$.

3.2 Perception of non-native speech from multiple talkers

A mixed design ANOVA indicated a main effect of listening condition [$F(2,15) = 28.250$, $p < 0.001$, generalized $\eta^2 = 0.664$], but no effect of exposure stage [$F(3,45) = 0.609$, $p = 0.613$, generalized $\eta^2 = 0.019$]. No interaction of listening condition by exposure stage was found [$F(6,45) = 0.356$, $p = 0.903$, generalized $\eta^2 = 0.022$]. Figure 1(b) displays performance across groups. *Post hoc* comparisons revealed no significant improvement from initial exposure to final exposure for any group: unprocessed, $p = 0.502$, vocoded, $p = 0.353$, CI, $p = 0.759$.

3.3 Single- versus multiple-talker conditions

Post hoc comparisons testing for differences in single- versus multi-talker conditions across exposure to all 40 sentences revealed that NH listeners in the unprocessed condition had higher intelligibility scores when perceiving speech from only a single non-native talker, $p = 0.049$. No differences between single- and multi-talker conditions were observed for NH listeners perceiving vocoded speech nor for CI users, both $p > 0.05$. Given the greater amount of variability in performance from CI users, further analyses are provided in supplementary material[19] revealing that CI users' performance when perceiving native-accented speech from a single native talker can partially predict outcomes for non-native speech perception.

4. Discussion

Results demonstrated that NH listeners benefited from rapid adaptation to a single non-native talker. Talker-dependent adaptation was also observed with nine-channel vocoded speech, though only to a limited extent. Talker-dependent adaptation was not observed for CI users. No listener group benefited from increased exposure to multiple talkers, suggesting that rapid accent-dependent adaptation is limited across all groups. A greater amount of variability for CI users in the multiple-talker condition revealed that while some performed worse with increased exposure, others improved to levels similar to NH listeners. This suggests that some CI users benefit from rapid accent-dependent adaptation while others are further impaired. Finally, a relationship between how well CI users perceive native-accented speech and how well they can rapidly adapt to non-native speech was also found. This suggests that CI users' performance in easier listening conditions can partially predict non-native speech perception outcomes.

The present work provides a preliminary comparison of potential mechanisms used by CI users and NH listeners to adapt to non-native talkers. Intelligibility of non-native speech was scored initially and after increased exposure in a rapid adaptation paradigm with single- and multiple-interleaved-talker sentence sequences to assess whether a talker- or accent-dependent benefit would be observed for each listener group. The results showed that intelligibility scores were lower for CI users compared to NH listeners in both talker conditions, and unlike for NH listeners, no talker-dependent benefit was observed for CI users after increased exposure. Despite no observed improvement for any CI participant in the single-talker condition, two showed some improvement in the multiple-talker condition. This implies that some CI users can more readily adapt to non-native speech via an accent-dependent mechanism that may utilize systematic

https://doi.org/10.1121/10.0001941

non-native production cues that are better encoded in implant devices, such as duration cues.[5] Since not all CI users adapted in the multiple-talker condition, and some did worse with more exposure, other factors may be limiting performance, such as listener fatigue. Also, listeners who improved in the multiple-talker condition may have also improved if they were in the single-talker condition. Future work could use a repeated-measures design to control for this possibility.

The observed lower intelligibility scores and lack of rapid adaptation experienced by CI users perceiving non-native speech may be a consequence of limited access to talker-specific fine structure cues. This account is further supported by the present results from NH listeners' perception of vocoded non-native speech, which also limits talker identity cues while minimizing the effects of uncontrolled factors associated with electric hearing. Lower intelligibility scores were observed when NH listeners perceived vocoded speech compared to unprocessed speech. Listeners were able to somewhat overcome their initial limitation in the vocoded condition with increased exposure, but not to levels of those in the unprocessed condition. Also, no difference between the two talker conditions was observed.

Although the CI users and those in the vocoded condition may have had some access to talker-specific cues,[14,15] since NH listeners perceiving vocoded speech do not also have the added potential confounds that CI users incur, this may be why listeners perceiving vocoded speech from a single talker could partially adapt. Because of these added confounds, CI users may still be able to adapt but require more processing time than what is needed for NH listeners.[22] Age may have also contributed to the results. In the present work, the CI participants were, on average, older than the NH listeners. A recent study found that older CI listeners were at an increased disadvantage when perceiving non-native speech compared to younger CI users.[11] Although age may have been a factor in the present findings, adaptation in the vocoded single-talker condition never reached levels observed in the unprocessed condition despite listeners in both groups sharing matching demographics. This suggests that reduced spectral resolution may largely, though not fully, explain the deficits shown by CI users and listeners in the vocoded condition. Another important consideration is that the present work did not assess listeners' abilities to adapt to native-accented speech with increased exposure in both talker conditions. It could be that CI users would also exhibit the same difficulties with any unfamiliar talker. However, previous work included both native- and non-native talkers to address this concern using vocoded speech and found that listeners performed near ceiling performance in both native talker conditions, while scores were drastically lower when perceiving non-native speech.[8] Those results suggest the present findings are more specific to non-native speech.

Finally, a moderate relationship was found between CI users' performance when listening to the same native-accented talker and performance when listening to non-native speech. Despite this correlation, even higher performing CI users did not attain scores near those of NH listeners when perceiving non-native speech. This indicates that outcomes of CI users' perception of non-native speech are more complex than their performance abilities in easier listening conditions. Specifically, in more difficult listening situations, there is likely an interplay between device limitations, such as reduced spectral resolution, and cognitive limitations related to attention and processing speed that disrupt rapid adaptation by impeding utilization of top-down processing to enhance acoustic bottom-up information.[12] These limitations can vary extensively across CI users and support the need to develop processing and listening strategies for improved performance under realistic listening environments.

Acknowledgments

We thank our participants and Christina Mai and Danni Yang for research assistance. We also thank Martin Cooke and our anonymous reviewers for their helpful suggestions.

References and links

[1]K. J. Van Engen and J. E. Peelle, "Listening effort and accented speech," Front. Hum. Neurosci. **8**, 1–4 (2014).
[2]M. M. Baese-Berk, A. R. Bradlow, and B. A. Wright, "Accent-independent adaptation to foreign accented speech," J. Acoust. Soc. Am. **133**, EL174–EL180 (2013).
[3]A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," Cognition **106**, 707–729 (2008).
[4]C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," J. Acoust. Soc. Am. **116**, 3647–3658 (2004).
[5]R. E. Baker, M. M. Baese-Berk, L. Bonnassee-Gahot, M. Kim, K. J. Van Engen, and A. R. Bradlow, "Word durations in non-native English," J. Phon. **39**, 1–17 (2011).
[6]B. L. Rochet, "Perception and production of second-language speech sounds by adults," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York, Timonium, MD, 1995), pp. 379–410.
[7]J. Y. Choi, E. R. Hu, and T. K. Perrachione, "Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing," Atten. Percept. Psychophys. **80**, 784–797 (2018).

[8]M. R. Kapolowicz, V. Montazeri, and P. F. Assmann, "Perceiving foreign-accented speech with decreased spectral resolution in single- and multiple-talker conditions," J. Acoust. Soc. Am. **143**, EL99–EL104 (2018).
[9]C. Ji, J. J. Galvin, Y. Chang, A. Xu, and Q.-J. Fu, "Perception of speech produced by native and nonnative talkers by listeners with normal hearing and listeners with cochlear implants," J. Speech Lang. Hear. Res. **57**, 532–554 (2014).
[10]T. N. Tamati and D. B. Pisoni, "The perception of foreign-accented speech by cochlear implant users," in *Proceedings of the 18th International Congress of Phonetic Sciences* (2015).
[11]E. Waddington, B. N. Jaekel, A. R. Tinnemore, S. Gordon-Salant, and M. J. Goupell, "Recognition of accented speech by cochlear-implant listeners," Ear Hear. **41**, 1236–1250 (2020).
[12]D. Başkent, E. Gaudrain, T. N. Tamati, and A. Wagner, "Perception and psychoacoustics of speech in cochlear implant users," in *Scientific Foundations of Audiology: Perspectives from Physics, Biology, Modeling, and Medicine*, edited by A. T. Cocace, E. de Kleine, A. G. Holt, and P. van Dijk (Plural, San Diego, CA, 2016), pp. 285–319.
[13]M. Vongphoe and F.-G. Zeng, "Speaker recognition with temporal cues in acoustic and electric hearing," J. Acoust. Soc. Am. **118**, 1055–1061 (2005).
[14]E. Gaudrain and D. Başkent, "Discrimination of voice pitch and vocal-tract length in cochlear implant users," Ear Hear. **39**, 226–237 (2018).
[15]V. Krull and X. Luo, "Talker-identification training using simulations of binaurally combined electric and acoustic hearing: Generalization to speech and emotion recognition," J. Acoust. Soc. Am. **131**, 3069–3078 (2020).
[16]Y. Chang and Q-J Fu, "Effects of talker variability on vowel recognition in cochlear implants," J. Speech Lang. Hear. Res. **49**, 1331–1341 (2006).
[17]E. Gaudrain and D. Başkent, "Factors limiting vocal-tract length discrimination in cochlear implant simulations," J. Acoust. Soc. Am. **137**, 1298–1308 (2015).
[18]G. S. Stickney, P. F. Assmann, J. Chang, and F.-G. Zeng, "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078 (2007).
[19]See supplementary material at https://doi.org/10.1121/10.0001941 for demographic information about the CI users' statistical analyses and a figure displaying the relationship between CI users' performance when perceiving native-accented speech and when perceiving non-native speech after increased exposure.
[20]M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411 (1997).
[21]B. S. Wilson and M. F. Dorman, "Cochlear implants: A remarkable past and a brilliant future," Hear. Res. **242**, 3–21 (2008).
[22]G. N. L. Smith, D. B. Pisoni, and W. G. Kronenberger, "High-variability sentence recognition in long-term cochlear implant users: Associations with rapid phonological coding and executive functioning," Ear Hear. **40**, 1149–1161 (2019).