Journal of Communication Disorders 87 (2020) 106026

Contents lists available at ScienceDirect

Journal of Communication Disorders

journal homepage: www.elsevier.com/locate/jcomdis

# The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech

Jacqueline McKechnie[a,b,*], Beena Ahmed[c,d], Ricardo Gutierrez-Osuna[e], Elizabeth Murray[a], Patricia McCabe[a], Kirrie J. Ballard[a]

[a] Faculty of Health Sciences, The University of Sydney, Lidcombe, NSW, Australia
[b] Faculty of Health, University of Canberra, Bruce, ACT, Australia
[c] Texas A&M University at Qatar, Doha, Qatar
[d] Faculty of Engineering, University of New South Wales, Sydney, NSW, Australia
[e] Texas A & M University, College Station, TX, USA

ARTICLE INFO

*Keywords:*
Childhood apraxia of speech
Mobile technology
Service delivery
Principles of motor learning

ABSTRACT

*Purpose:* One of the key principles of motor learning supports using knowledge of results feedback (KR, i.e., whether a response was correct / incorrect only) during high intensity motor practice, rather than knowledge of performance (KP, i.e., whether and how a response was correct/incorrect). In the future, mobile technology equipped with automatic speech recognition (ASR) could provide KR feedback, enabling this practice to move outside the clinic, supplementing speech pathology sessions and reducing burden on already stretched speech-language pathology resources. Here, we employ a randomized controlled trial design to test the impact of KR vs KP feedback on children's response to the Nuffield Dyspraxia Programme 3, delivered through an android tablet. At the time of testing, ASR was not feasible and so correctness of responses was decided by the treating clinician.

*Method:* Fourteen children with CAS, aged 4–10 years, participated in a parallel group design, matched for age and severity of CAS. Both groups attended a university clinic for 1-hr therapy sessions 4 days a week for 3 weeks. One group received high frequency feedback comprised of both KR and KP, in the style of traditional, face-to-face intensive intervention on all days. The other group received high frequency KR + KP feedback on 1 day per week and high frequency KR feedback on the other 3 days per week, simulating the service delivery model of one clinic session per week supported by tablet-based home practice.

*Results:* Both groups had significantly improved speech outcomes at 4-months post-treatment. Post-hoc comparisons suggested that only the KP group showed a significant change from pre- to immediately post-treatment but the group difference had dissipated by 1-month post-treatment. Heterogeneity in response to intervention within the groups suggests that other factors, not measured here, may be having a substantive influence on response to intervention and feedback type.

*Conclusion:* Mobile technology has the potential to increase motivation and engagement with therapy and to mitigate barriers associated with distance and access to speech pathology services. Further research is needed to explore the influence of type and frequency of feedback on motor learning, optimal timing for transitioning from KP to KR feedback, and how these parameters interact with task, child and context-related factors.

* Corresponding author at: PO Box 5014, University of Canberra LPO, Bruce, ACT, 2617, Australia.
*E-mail address:* Jacqui.McKechnie@canberra.edu.au (J. McKechnie).

https://doi.org/10.1016/j.jcomdis.2020.106026
Received 18 March 2019; Received in revised form 11 May 2020; Accepted 18 May 2020
Available online 08 July 2020
0021-9924/ 

## 1. Introduction

Childhood apraxia of speech (CAS) is a disorder of speech motor control that causes substantial disruption to development of intelligible and natural sounding speech (ASHA, 2007). The speech of children with CAS is notable for substitutions and distortions of speech sounds and altered prosody. As a disorder of speech motor control, it is often recommended that CAS treatment apply principles of motor learning (PML), including high frequency of treatment sessions and high numbers of practice trials per session (Maas et al., 2008; Schmidt & Lee, 2011). However, parents often report difficulty accessing, attending and affording this level of clinical care and a willingness for alternative service delivery methods to alleviate these burdens (Ruggero, McCabe, Ballard, & Munro, 2012). It is here that theory-driven development of mobile technology applications can offer children with CAS access to engaging high intensity speech therapy that follows the best-practice PML. Apps can provide animated platforms in both exercise and game formats, to engage and motivate children to undertake the difficult and often tedious intensive practice. Incorporating state of the art automated speech recognition (ASR) also has potential, in the future, for delivering feedback on accuracy of speech attempts, leading to reduced burden on clinician time and the cost born by families. However, ASR, by nature, can only deliver feedback on correctness (i.e., knowledge of results, KR), and non-speech motor learning literature suggests that people benefit from knowledge of performance (KP) feedback when first acquiring new motor skills (Maas et al., 2008; Schmidt & Lee, 2011). It is not yet known how some of these principles of motor learning apply to children with CAS and whether some children respond differently to these principles than others. Maas and colleagues have reported that different children respond differently to principles of practice schedule (Maas & Farinella, 2012) and feedback frequency (Maas, Butalla, & Farinella, 2012). However, no study has specifically compared the effects of KP and KR on speech motor learning in children with CAS. It is important to know this before we embark on prescribing apps that may force specific principles or aspects of therapeutic method.

Here, we explore the impact of using KP versus KR feedback, (a) to test whether children respond differently to these two types of feedback in a controlled setting, and (b) to inform future development of speech therapy apps as ASR technology comes online. To isolate the effects of feedback type, independent of the performance of any ASR algorithm, clinicians supervised the app-delivered treatment and made perceptual judgments of children's speech for feedback decisions.

### 1.1. Treatment for CAS

There are different approaches to treatment for CAS currently used around the world. These include motor-based approaches, linguistic approaches and multi-modal communication approaches. In a systematic review of the evidence on treatment for CAS, Murray et al. (2014) identified three treatment protocols as having the strongest levels of evidence to support their use in a clinical setting to achieve positive treatment, maintenance and generalization effects. These included Dynamic Temporal and Tactile Cueing [DTTC] (Strand, Stoekel, & Baas, 2006), Rapid Syllable Transition Treatment [ReST] (Ballard, Robin, McCabe, & McDonald, 2010; Murray, McCabe, & Ballard, 2012), and Integrated Phonological Awareness Intervention (Moriarty & Gillon, 2006). There was suggestive evidence for ten other treatment approaches including the Nuffield Dyspraxia Programme – Third Edition (NDP3; Williams & Stephens, 2004), which is commonly used across Australia as best-practice (Gomez, McCabe, & Purcell, 2018). This review then led to the first and, currently, only randomized controlled trial (RCT) of treatment for CAS, comparing the NDP3 and ReST (Murray, McCabe, & Ballard, 2015). Results of the RCT indicated that both NDP3 and ReST treatments resulted in similar positive treatment outcomes, particularly for generalization to real words. The authors reported that NDP3 demonstrated greater immediate gains in speech accuracy and ReST treatment lead to better maintenance of treatment gains and generalization to untreated pseudo-words. However, a subsequent Cochrane review (Morgan, Murray, & Liégeois, 2018) offered a more conservative interpretation of these findings based on a re-analysis. This suggested no reliable difference existed between the two treatment groups on acquisition or maintenance of targets based on small absolute mean differences in accuracy scores between the groups and that both treatment protocols demonstrated a similar, moderate level of evidence (Morgan et al., 2018). Thus, the app-delivered intervention approach employed in this study is modeled on the NDP3 given the moderate level of evidence for positive treatment and maintenance effects as demonstrated by the RCT (Morgan et al., 2018; Murray et al., 2015). In addition, the NDP3 uses real words, which potentially can be analyzed with ASR algorithms built on databases of incidental speech; and its theoretical basis is in motor learning as a complex and hierarchical skill involving sounds, syllable shapes, words and prosody which are developed through repetition, elicitation and the provision of frequent, specific feedback on performance and results (Williams & Stephens, 2004, 2010). This is similar to the pre-practice phase of a motor learning approach.

### 1.2. Principles of motor learning

Much of what we know about principles of motor learning (PML) has come from limb movement studies in non-disordered populations or investigations involving adults with acquired apraxia of speech (AOS) or dysarthria. Limb movement studies have demonstrated that greater long-term learning occurs when practice of motor targets is variable, randomized, and frequent, with delayed feedback provided on an intermittent schedule (see Maas et al., 2008 for a review). However, investigation into adult motor speech disorders revealed that some participants benefited more from low frequency feedback and others from high frequency feedback, with similar mixed results when exploring the effects of delayed versus immediate feedback (Austermann Hula, Robin, Maas, Ballard, & Schmidt, 2008). The type of feedback received also influences acquisition and retention effects. Non-speech motor learning literature has demonstrated that detailed feedback about the motor movement, designed to guide and shape subsequent movements towards a correct response (i.e., Knowledge of Performance, KP; e.g., "straighten your arm more quickly") enhances

acquisition but potentially inhibits maintenance of skill post-treatment. In contrast, feedback on the outcome or correctness of the completed movement (i.e., Knowledge of Results, KR; e.g., "That's right" or "Not that time") leads to greater maintenance of skill (Schmidt & Lee, 2011). However, KR is most effective when the learner has some internal representation of the target movement program and some ability to self-evaluate and self-correct (see Maas et al., 2008 for a review of non-speech and speech motor learning literature). As mentioned previously, the effects of KP and KR has not been directly compared in speech motor learning.

Few studies have explicitly investigated the influence of specific principles of motor learning in CAS. The principles that have been studied include (a) amount of practice, where providing ~ 150 trials per session leads to greater treatment, generalization and maintenance effects than only 30–40 trials per session (Edeal & Gildersleeve-Neumann, 2011); (b) treatment intensity, where twice weekly treatment sessions led to significantly better outcomes than once per week treatment sessions (Namasivayam et al., 2015); (c) practice schedule (i.e., blocked versus random practice; Maas & Farinella, 2012), where findings were mixed across participants; (d) feedback frequency (i.e., low versus high frequency feedback; Maas et al., 2012) where findings were also mixed across participants (see Maas, Gildersleeve-Neumann, Jakielski, & Stoekel, 2014 for a review); and (e) distribution of practice (i.e., closely distributed at four times weekly for three weeks versus less closely distributed at twice weekly for six weeks; Thomas, McCabe, & Ballard, 2014).

In their RCT comparing treatment outcomes from the NDP3 and ReST treatments, Murray et al. (2015) suggested that inherent differences in the type and frequency of feedback provided to children may have influenced children's responses to intervention. Although both groups made significant improvements with treatment, NDP3 treatment, which uses 100 % KP feedback, effected greater improvement on treated targets immediately post-treatment (i.e., greater acquisition) than ReST intervention, which utilizes 50 % KR feedback only (in line with PML). Conversely, the ReST group showed greater maintenance of treatment effects than the NDP3 group. The results from Murray et al. (2015) are consistent with previous work arguing that high frequency KP feedback confers an acquisition advantage, while low frequency KR feedback confers a maintenance advantage (e.g., Maas et al., 2008; Schmidt & Lee, 2011). Despite the more conservative interpretation offered by Morgan et al. (2018) concluding no reliable difference between the two treatments, exploration of the effects of feedback type and feedback frequency is warranted given that others have also reported equivocal effects for feedback parameters such as frequency when treating CAS (Maas et al., 2012). The current study was designed to specifically investigate the influence of the type of feedback received during speech production practice when therapy was delivered using mobile technology that has potential to provide KR feedback only via ASR. KP versus KR feedback has not previously been systematically compared; that is, the study of Murray et al. (2015) varied many components between their two treatment approaches such that the specific effect of feedback type could not be determined. To isolate the effect of feedback type, we maintained feedback frequency at 100 % for both experimental groups and administered the same treatment protocol to both groups.

### 1.3. Service delivery

Despite research consistently demonstrating that best practice intervention frequency for speech sound disorders, including CAS, is between 2 and 4 sessions per week (Murray, McCabe, & Ballard, 2014; Namasivayam et al., 2015; Sugden, Baker, Munro, Williams, & Trivette, 2018; Thomas et al., 2014), these intervention frequencies are uncommon in clinical practice (Gomez et al., 2018; Ruggero et al., 2012; Sugden, Baker, Munro, Williams, & Trivette, 2017). Parent involvement and home practice activities are routinely prescribed as a way to supplement face-to-face therapy sessions with a clinician (Lim, McCabe, & Purcell, 2017; Sugden, Baker, Munro, & Williams, 2016, 2017). Homework can also provide the frequent and regular practice of speech production targets that is needed for children to acquire new skills and habitualize these new movement skills, as well as different but related movement skills, into non-intervention contexts (Gordon-Brannan & Weiss, 2007; McLeod & Baker, 2017; Olswang & Bain, 2013). Effective home practice requires that the child be motivated to engage in their practice activities and that parents or carers can be available to supervise the practice sessions and provide feedback on the accuracy of the child's productions. However, parents and children perceive speech practice as "work" (Thomas, McCabe, & Ballard, 2017; McAllister, McCormack, McLeod, & Harrison, 2011). In addition, research has demonstrated that treatment fidelity (i.e., adherence to the treatment protocol) and difficulty judging the accuracy of their child's speech production attempts have been identified as barriers to the efficacy of parent-implemented approaches (Lim et al., 2017; Thomas et al., 2017; Thomas, McCabe, Ballard, & Bricker-Katz, 2018). It is here that computer-based or app-delivered home practice can be useful for strict protocol delivery as well as increasing a child's engagement and motivation to participate in speech homework (Hair, Monroe, Ahmed, Ballard, & Gutierrez-Osuna, 2018; Nordness & Beukelman, 2010; Toki & Pange, 2010).

### 1.4. Computer-based treatment approaches

Computer software packages designed to act as a virtual speech-language pathologist (SLP) can be effective for a range of pediatric speech disorders (Chen et al., 2016; Furlong, Erickson, & Morris, 2017). Fewer than half of the programs in two recent review papers provided feedback to the user on speech attempts (Chen et al., 2016; Furlong et al., 2017). Program-driven feedback was mostly non-specific (e.g., a visual speech waveform or tracking number of trials completed). Explicit feedback on speech accuracy was experimenter/clinician controlled and judged. None of the included studies in either review included mobile technology.

The efficacy or effectiveness of therapy for pediatric speech sound disorders delivered via tablet and smartphone applications, however, has not been empirically tested (McKechnie, Ahmed, Gutierrez-Osuna, Monroe, McCabe & Ballard, 2018). This may be due in part to the risks in running time- and cost-intensive experimental trials in the fast turnover environment of the app market, compared with the relative low cost, both of producing and purchasing an app, and perceived low risk of the products themselves, given that apps are often frequently similar to printed materials already in use or form only a small part of an overall intervention

approach (Edwards & Dukhovny, 2017). However, recent analysis of the quality and potential therapeutic benefit of mobile applications for children's speech disorders found that less than 3% of more than 5000 identified apps met criteria that would warrant full evaluation (Furlong, Morris, Serry, & Erickson, 2018). Of that 3% (132 unique apps that were appraised), only 19 apps (14 %) were deemed to have therapeutic potential in the sense that engagement with the app offered a positive potential impact on the user's speech sound disorder (Furlong et al., 2018). Positive potential impact was derived from the behavior change scale from the Perceived Impact section of the Mobile App Rating Scale (MARS; Stoyanov, Hides, Kavanagh, Zelenko, Tjondronegoro & Mani, 2015), where perceived impact was defined as potential to improve the user's awareness, knowledge, attitudes, help-seeking behavior as well as the perceived behavioral change from use of the app (Stoyanov et al., 2015). Note that this analysis is not a valid replacement for well-controlled experimental studies of efficacy or effectiveness, given the high risk of bias.

Each app evaluated in Furlong et al. (2018) was rated by two reviewers on a 5-point likert scale from strongly disagree to strongly agree for the statement "the use of this app is likely to increase/decrease [insert target health behavior]". Raters were experienced speech-language pathologists. Inter-rater reliability, calculated using two-way mixed Interclass Correlation, was reported only for the total MARS score which did not include the score for perceived impact. Nonetheless, satisfactory inter-rater reliability was obtained (ICC .72) (Furlong et al., 2018).

The majority of available computer- or app-based intervention tools offer digital stimulus presentation via engaging graphics and sound effects. They typically do not provide the child with explicit feedback on the accuracy of their productions (KP or KR) nor offer remote and/or automated assessment for the SLP to monitor. The lack of integrated, automated feedback is largely due to the challenges involved in developing ASR software which can provide decisions on speech production accuracy that are highly reliable with expert clinician judgements and delivered in a timely manner (see McKechnie et al., 2018 for a review). There has been limited research on computer-based or mobile technology approaches for CAS, perhaps due to the historical challenges in defining a relatively homogeneous group of children for testing and developing computerized approaches that treat the range of CAS features, not just segmental accuracy.

### *1.5. Service delivery for CAS using mobile technology*

There is a mismatch between the need for children with CAS to receive intensive treatment and the reality of service delivery models in Australia and elsewhere. As a result, recent research evidence has emerged exploring the efficacy of ReST intervention delivered via alternative methods including tele-practice (Thomas, McCabe, Ballard, & Lincoln, 2016) and combined clinician- and parent-delivered methods (Thomas et al., 2017). However, there has currently been no well-controlled study published that investigates alternatives to fully clinician-delivered (in person or via video conferencing) speech therapy. In light of the fact that NDP3 is the most frequently used intervention for CAS in Australia (Gomez et al., 2018), it is timely to consider the impact of alternative service delivery methods on treatment efficacy for CAS using NDP3 intervention.

In light of the low number of commercially available apps deemed to have potential for therapeutic impact (Furlong et al., 2018), our group have developed *Tabby Talks*, which is a theory-driven, multi-tiered system for facilitating remote access to speech-language pathology services (Parnandi et al., 2013, 2015). *Tabby Talks* consists of three components: (1) Android platform application running on mobile tablets, (2) server-based learning management software (i.e., Moodle) running a speech analysis engine to evaluate children's speech attempts offline for assessment of progress in therapy, and (3) a clinician interface allowing for the remote management and updating of clients and therapy exercises (see Table 1 and Fig. 1).

The clinician interface allows the clinician to create individual client profiles and input demographic information about each client; to create exercises containing customizable sets of stimuli to target specific speech production goals; and to enroll clients into exercises based on their individual assessment data. The Moodle server houses the speech analysis engine which evaluates the children's speech offline and communicates data back to the clinician interface to allow for remote monitoring of progress through therapy (See Fig. 2A-B). The application running on the tablets requires each client user to log in with an individualized user-name and password. The app then displays the set of exercises into which that client has been enrolled and enables practice of speech targets in either a flash card/swipe through display or a simple memory game format. The user interface has icons for the client to touch, which will either play an audio model of the target word or start an audio recording to capture the client's production attempt. The user interface tested here also has a non-ASR-based scoring system where the clinician can make a perceptual judgement of

**Table 1**
Features available in the *Tabby Talks* multi-tiered system for facilitating remote access to speech pathology services.

| App features (online real-time) | Server features (offline) |
|---|---|
| - Real-speech audio models | - Speech recognition software |
| - Coloured flash cards | - Individual case files |
| - Swipe features and simple memory game | - Access to saved audio recordings of every production trial, for each child |
| - Record and playback function | - Graphs of session by session accuracy |
| - Animated cartoon cat providing motivational feedback | - Bar charts presenting star and medal data for each practiced word or goal. |
| - Star chart and medals for reaching milestones | |
| - [ASR-ready][1] | |

**Note.** [1]ASR = automatic speech recognition / analysis. At the time of this study, the tablet-based app was ASR-ready. ASR was not used here as reliability of the app-compatible algorithms was still being tested.

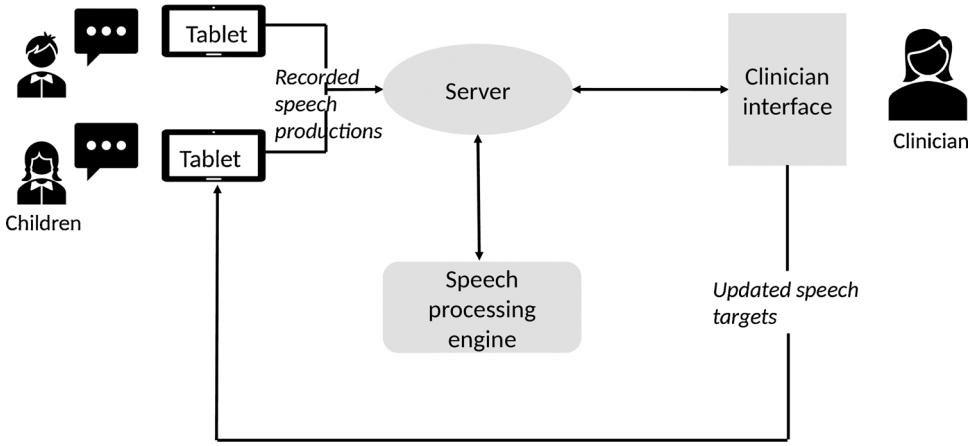


**Fig. 1.** Overview of the *Tabby Talks* system architecture.

speech accuracy by selecting/awarding either a gold (for correct) or silver (for incorrect) star (see Fig. 3A–C). Additional details about the design and development of the *Tabby Talks* infrastructure have been published elsewhere (Parnandi et al., 2013, 2015).

One of the critical considerations, when exploring alternative service delivery options for CAS, is how these options will affect the structure of the treatment protocol and how different PMLs can be incorporated. Given the intention to incorporate ASR into *Tabby Talks*, and its restriction to KR feedback only, the first step in testing our program was to examine the impact of delivering primarily KR feedback (i.e., right / wrong decisions) vs KP feedback. While the PML approach advocates KR feedback during practice for best maintenance of treatment effects, a learner must first be trained in producing the target movement skills accurately through what is referred to as pre-practice. Pre-practice, unlike practice, is where the clinician/trainer provides detailed KP feedback to guide and shape performance so that the learner can experience the sensorimotor consequences of performing the targeted movement(s) correctly. Pre-practice serves to guide the learner in developing an internal reference of correctness that can be accessed later during practice, once KP is removed. This internal reference is needed to guide self-evaluation and self-correction. We propose that *Tabby Talks* may be best used in between the weekly in-clinic pre-practice sessions with the speech-language pathologist, to provide high intensity and frequent practice on speech behaviors that the child has begun to acquire.

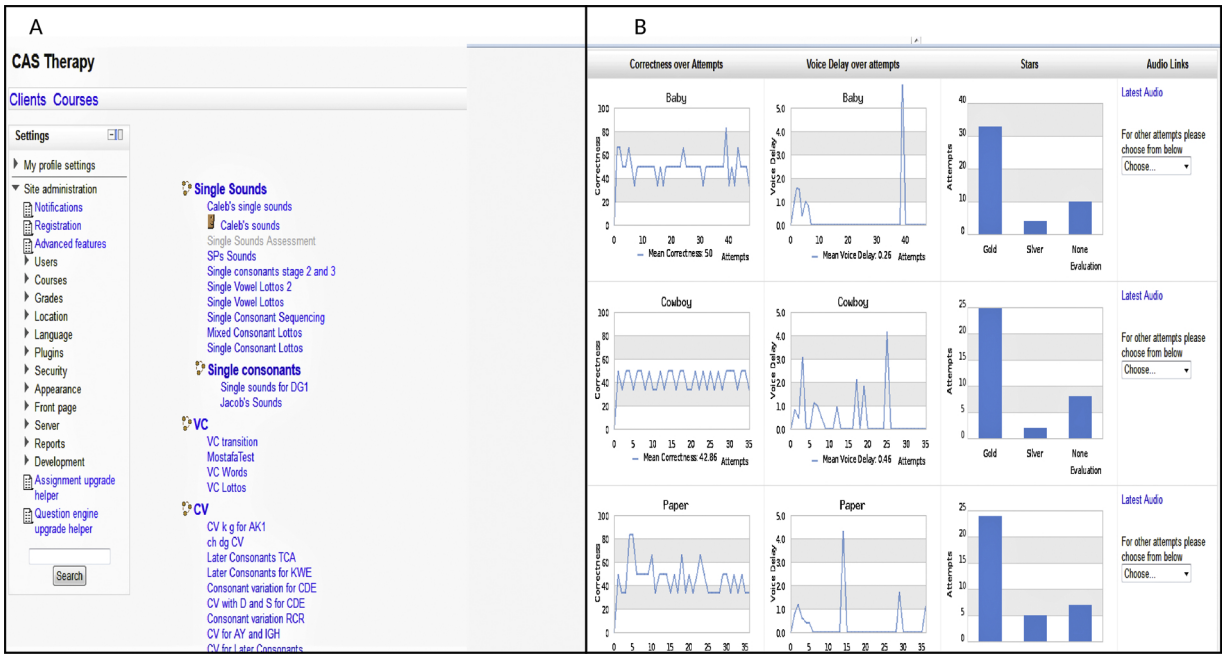


**Fig. 2.** Clinician server/interface showing: A. list of available/created speech exercises; and B. a visual display of client progress towards speech goals.

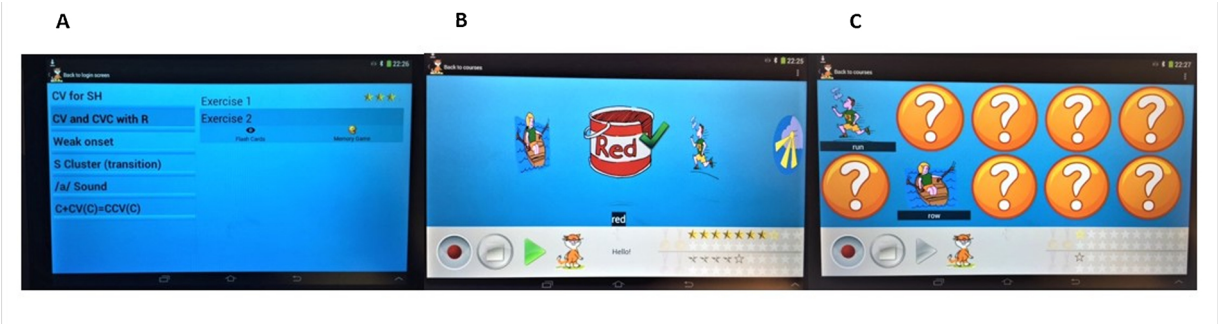


**Fig. 3.** User interface showing: A. all enrolled speech exercises for the child; B. practice exercises and reward stars during flash card mode; C. practice exercises and reward stars during memory game mode.

### *1.6. Purpose*

This study aims to (a) explicitly investigate the influence of type of feedback (primarily KP vs primarily KR) on response to treatment for CAS using the *Tabby Talks* app; and (b) to determine the feasibility for such technology and software to provide an effective supplement to face-to-face intensive treatment. We compared intensive 4 days / week therapy with the clinician (100 % KP) with the more common model of once / week with the clinician (100 % KP) and 3 days / week with the app (100 % KR). Being an efficacy study, all sessions were conducted in the research clinic under supervision of the research team, so that parameters of the treatment protocol, including the frequency and type of feedback, could be carefully controlled. In this way, we simulated the once / week in-clinic model (i.e., the intervening 3 "at-home" sessions were completed in the clinic) and we also simulated the ASR-style KR that the app would provide during home practice sessions. Here, *Tabby Talks* was populated with stimuli from the NDP3 (with permission from the authors, Williams & Stephens, 2004).

All conditions other than feedback type were held constant across the two groups: children in both treatment conditions attended the clinic for all therapy sessions, all sessions were delivered by trained student speech-language pathologists under the supervision of experienced clinicians, all treatment stimuli were delivered via the *Tabby Talks* app, and the student clinicians delivered all feedback verbally.

### *1.7. Research aims and hypotheses*

This study explored the impact of using KP versus KR feedback, (a) to test whether children respond differently to these two types of feedback in a controlled setting, and (b) to inform future development of speech therapy apps as advancements in ASR technology progress towards full online integration into the apps. We also compared the two methods of treatment used here to our historical data for traditional paper-based delivery of NDP3 (Murray et al., 2015). Further, we invited participants to complete a questionnaire exploring satisfaction with the treatment process; motivation and engagement with therapy activities; app features, likes, and dislikes; ease of use; and interest in further treatment using the app. These results will be used to further inform future app development and are available as supplemental materials. We hypothesized that:

(i) Tablet-based delivery of NDP3 using high frequency KP feedback would obtain similar treatment outcomes to Murray et al.'s (2015) traditional paper-based delivery of NDP3.
(ii) Compared to participants in the high frequency KP group and the traditional paper-based NDP3 group, participants in the high frequency KR condition may demonstrate smaller treatment gains immediately post-treatment (i.e., evidence of slower acquisition and generalization) but greater maintenance at 1- and 4- months post-treatment (i.e., evidence of more robust learning).
(iii) The experimental groups would demonstrate at least similar long-term outcomes to Murray et al.'s (2015) traditional NDP3 delivery.

## 2. Method

This study was approved by the Human Research Ethics Committee at the University of Sydney (Protocol number 2013/703). All parents provided written informed consent for their child to participate and children older than 6 years of age provided written assent.

### *2.1. Participants*

Recruitment occurred via university research volunteer websites, advertisement in magazines of relevant professional associations, as well as flyers to community-based SLPs, social media forums for SLPs and special interest groups for CAS.

Inclusion criteria were (1) confirmed clinical diagnosis of CAS by the research team, as described below, (2) aged between 4 and

12 years at the time of treatment, (3) age appropriate receptive language skills, indicated by a standard score of ≥ 85 on the receptive language index of the Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2006) or CELF-Preschool-Second Edition (CELF-P2; Wiig, Semel, & Secord, 2006), (4) normal or adjusted to normal hearing and vision, (5) the child and at least one parent being native English speakers, and (6) no other diagnosed genetic, developmental or acquired diagnosis (e.g., autism spectrum disorder, dysarthria or intellectual disability).

A total of 38 children were referred. Referral sources were first interviewed by phone or via email to rule out potential contraindications to the inclusion criteria above. Comprehensive assessments were carried out in two stages. Assessments to determine eligibility for participation in the study included (1) a case history questionnaire; (2) hearing screening to exclude undiagnosed hearing impairment; (3) Peabody Picture Vocabulary Test – Fourth Edition (PPVT-4) (Dunn & Dunn, 2007) which is highly correlated with psychometric assessments of cognitive functioning and used here to exclude potential intellectual disability; (4) CELF-4 or CELF-P2 Australian Standardizations to exclude delayed receptive language skills and therefore avoid conflating our findings with concomitant intellectual or language comprehension difficulties; and (5) the Oral and Speech Motor Protocol (Robbins & Klee, 1987) to exclude oral-structural or dysarthria diagnoses. In addition, speech samples for perceptually judging the presence and severity of CAS were obtained through administration of (6) The Goldman-Fristoe Test of Articulation – Second Edition (GFTA-2) (Goldman & Fristoe, 2000); (7) the DEAP Inconsistency subtest (Dodd, Hua, Crosbie, Holm, & Ozanne, 2002); (8) Single Word Test of Polysyllables (Gozzard, Baker, & McCabe, 2008; Gozzard, Baker, & McCabe, 2004); and (9) NDP3 assessment (Williams & Stephens, 2004). Three experienced SLPs (first, fifth and sixth authors) independently confirmed diagnosis of CAS based on the presence of the three consensus-based features of CAS: (1) inconsistent errors on consonants and vowels, (2) difficulty transitioning between sounds and syllables; and (3) prosodic difficulties (ASHA, 2007); and any four of the 10 features in Strand's 10-point checklist (Shriberg, Potter, & Strand, 2009) over at least three assessment tasks. There was no disagreement between authors. A flowchart demonstrating the outcome at each stage of the referral and screening/eligibility process for each of the 38 referred children is shown in Fig. 4.

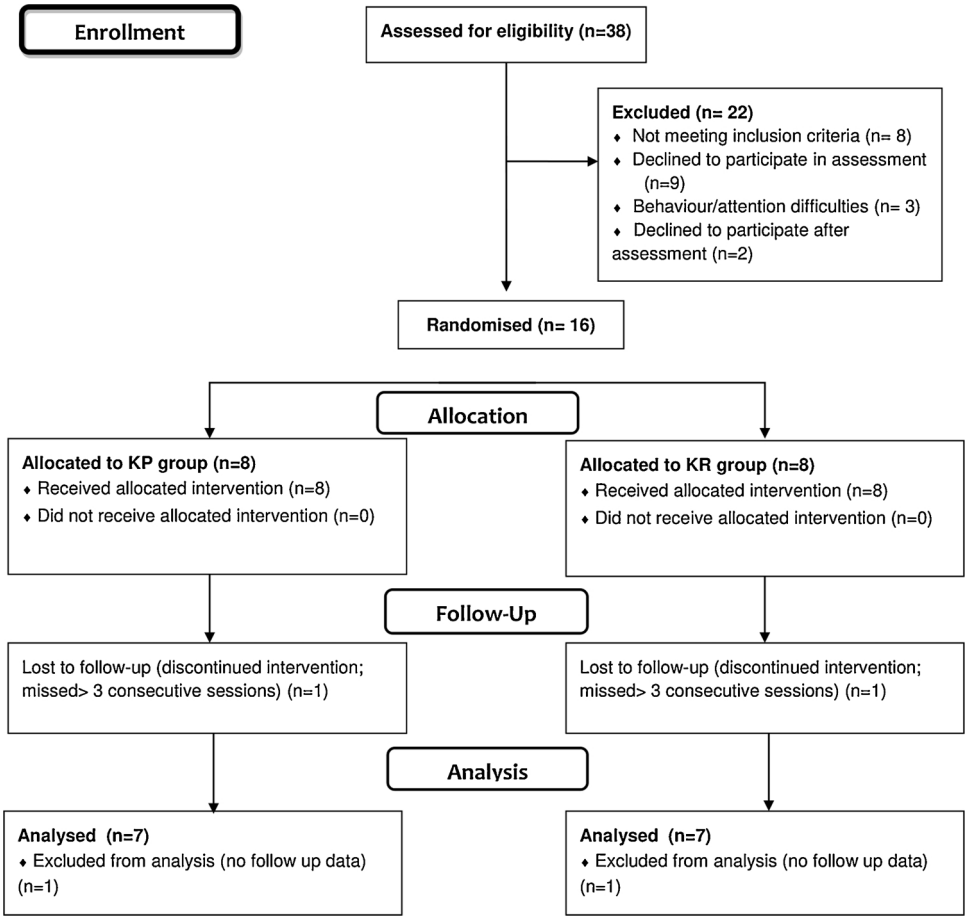


**Fig. 4.** CONSORT flow diagram of participant recruitment.

Fourteen children were included in the study: 13 males and 1 female aged between 4 and 11 years, with a mean age of 6;7 years (SD = 2;5; range of 4;1–10;10 years). Two sets of twins participated. Severity of CAS, ranged from mild to severe, as measured by Percent Consonants Correct (PCC; Shriberg, Aram, & Kwiatkowski, 1997; Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997) for the

Single Word Test of Polysyllables. Inter-rater reliability was > 85 % for point-to-point transcription reliability on both these tests (Kratochwill et al., 2010). Demographic data are presented in Table 2. There were no significant differences between the two groups on any of the baseline variables (i.e., age, primary and secondary outcome measures or CAS severity; see Table 2).

**Table 2**
Comparison of pretreatment variables by group for children with apraxia of speech assigned to either the Knowledge of Performance (KP) or Knowledge of Results (KR) feedback group.

| | KP group (n = 7) | | KR group (n = 7) | | | |
|---|---|---|---|---|---|---|
| Variable assessed | *M (SD)* | Range | *M (SD)* | Range | t | *p* |
| **Demographic** | | | | | | |
| Age in months | 81.7 (32.3) | 50 - 129 | 83.6 (33.7) | 54 – 131 | −0.11 | .92 |
| Sex | 7 Male | | 6 Male<br>1 Female | | | |
| Had previous speech treatment? | 7/7 | | 7/7 | | | |
| **Primary outcome measures at baseline** | | | | | | |
| Accuracy on items treated | 18.6 (15.2) | 0 - 42.3 | 20.5 (13.0) | 0 – 36.4 | −0.25 | .81 |
| Accuracy on items expected to generalize | 55.2 (12.5) | 41.8 -77.3 | 51.6 (20.6) | 24.5 -76.1 | 0.40 | .70 |
| **Secondary outcome measures at baseline** | | | | | | |
| *DEAP Inconsistency* | 48.0 (20.0) | 16 - 64 | 46.3 (15.6) | 24 – 68 | 0.18 | .86 |
| *Single Word Test of Polysyllables* | | | | | | |
| PPC | 68.9 (19.9) | 37 -89 | 62.8 (29.8) | 24 - 92 | 0.49 | .66 |
| PVC | 72.7 (16.8) | 45 - 91 | 70.0 (23.2) | 39 - 92 | 0.26 | .80 |
| PCC | 66.0 (23.2) | 32 - 96 | 57.4 (34.9) | 12 - 93 | 0.54 | .60 |
| Percent lexical stress matches | 62.5 (20.6) | 34 - 90 | 55.9 (28.3) | 24 – 88 | 0.50 | .63 |
| *GFTA-2* | | | | | | |
| Standard score | 73.7 (24.3) | 51 -109 | 72.3 (22.0) | 40 - 102 | 0.15 | .88 |
| PPC | 75.0 (14.0) | 52 - 91 | 69.8 (26.8) | 36 - 97 | 0.45 | .66 |
| PVC | 82.6 (8.9) | 65 - 91 | 81.8 (16.2) | 61 - 99 | 0.11 | .91 |
| PCC | 70.8 (17.8) | 45 - 95 | 62.9 (33.4) | 17 - 97 | 0.55 | .59 |
| *Speech disorder severity* | | | | | | |
| Severe (< 50 %) | *n* = 2 | | *n* = 3 | | | |
| Moderate-severe (50 – 65%) | *n* = 1 | | *n* = 1 | | | |
| Mild-moderate (65 – 85% | *n* = 3 | | *n* = 0 | | | |
| Mild (> 85 %) | *n* = 1 | | *n* = 3 | | | |
| *CELF-P2 / CELF-4* | | | | | | |
| Receptive language score | 97.3 (13.3) | 82 - 121 | 90.1 (7.6) | 81 - 106 | 1.23 | .24 |
| Expressive language score | 84.7 (14.5) | 66 - 107 | 85 (18.6) | 63 - 112 | 0.03 | .98 |

Note: DEAP = Diagnostic Evaluation of Articulation and Phonology (Dodd et al., 2002); Single Word Test of Polysyllables (Gozzard et al., 2004, 2008); PPC = percent phonemes correct; PVC = percent vowels correct; PCC = percent consonants correct; GFTA-2 = Goldman-Fristoe Test of Articulation – Second Edition (Goldman & Fristoe, 2000); Speech disorder severity was based on PCC from the Single Word Test of Polysyllables; CELF-P2 = Clinical Evaluation of Language Fundamentals – Preschool – Second Edition (Semel et al., 2006); CELF-4 = Clinical Evaluation of Language Fundamentals – Fourth Edition (Wiig et al., 2006).

### 2.2. Design

The study used a parallel-group design with groups matched by age and severity of disorder. Stratified randomization was employed to assign pairs of children, age- and severity- matched, to each treatment condition; that is, one child from each pair was randomly assigned to one treatment group and the matched pair assigned to the other group. In this way, each child within the sets of twins was randomized to a different group. The groups differed only by the type of feedback received (see *2.4 Feedback Conditions* below). All other components of the protocol were identical across the groups. Fig. 5 provides an overview of the assessment and treatment timeline of the experiment.

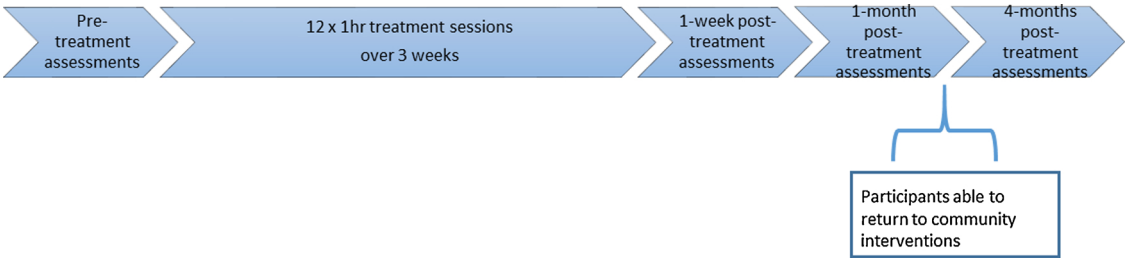


**Fig. 5.** Assessment and intervention timeline.

### 2.3. Intervention

The NDP3 was implemented as described by Williams and Stephens (2004, 2010) and operationalized by Murray and colleagues (Murray et al., 2012, 2015). Each child had three individualized speech production goals determined based on their pre-treatment assessment results. Goals were selected to include new speech sounds as single sounds or in known syllable shapes, new syllable structures using known sounds, and prosodic accuracy (i.e., lexical or phrasal stress). Five NDP3 stimulus words or phrases were selected per goal. Whereas the children in the original RCT (Murray et al., 2015) completed their speech production practice within 18-minute blocks using play-based activities, the nature of using app-delivered intervention required some adjustments to be made. Here, each goal was targeted in a 16-minute block using list-based exercises (i.e., swiping through the set of words/phrases and producing each target) and/or a memory game within the *Tabby Talks* app, with 2 min of free play between each goal. The total number of production trials per session was kept consistent with the protocol of Murray et al. (2012; 2015). Children needed to achieve 90 % spontaneous accuracy on each target item before new stimuli were introduced into the goal. Once all five stimuli within a goal reached criterion accuracy, the child was stepped up to the next level in the NDP3 hierarchy (see Appendix A). Immediate feedback was provided on 100 % of production attempts throughout the sessions; however, the two groups differed in the type of feedback received during their treatment sessions.

### 2.4. Feedback conditions

The KP Group received KP feedback (i.e., specific, performance-based information about articulators/voicing/timing and how to adapt or change their production for next time) which included KR feedback (i.e., on outcome accuracy on all production attempts (i.e., 100 % KR + KP feedback) on all four days per week, following the protocol of Murray et al. (2015). Teaching and cueing were provided as needed through verbal instructions, articulation placement cues, visual-tactile cues, metaphors, analogies and modeling.

The KR Group received 100 % KP feedback (which included KR feedback) on one day per week, as described above; and 100 % KR only feedback on the other three days per week. For children experiencing a high degree of difficulty (5 sequential incorrect responses), a brief period of KP feedback was introduced in order to establish a correct production before resuming with high frequency KR only (McCabe, Macdonald-D'Silva, van Rees, Ballard, & Arciuli, 2014). While this departs from the goal of 100 % KR feedback, this threshold for number of sequential errors is easily implemented in the app (Hair et al., 2018) and is necessary for duty of care. Clinicians collected data on the type of feedback provided to these children and engaged in continuous real-time monitoring of feedback type to ensure that a ratio of 80 % KR to 20 % KP was maintained for items on which a child was experiencing significant difficulty.

For both groups, when a production was correct, the child was instructed to repeat the response three times, with KR feedback provided by the clinician. This procedure is consistent with the NDP3 manual and the protocol developed by Murray et al. (2012). To maintain experimental control, all sessions were delivered in a University clinic and clinicians delivered all feedback.

Dose was controlled across both treatment groups. Treatment was delivered over 12 1-hr sessions, four days per week for 3 weeks during school vacation periods. Children in the KP group received an average of 156.2 response trials per session (SD = 44.9) and the KR children an average of 142.5 (SD = 36.6), and these dose levels were not statistically different ($t = 0.73$, $p = 0.49$).

Student speech-language pathologists in their third year of a four-year training program (intermediate level, previous experience with pediatric cases) provided the treatment under the supervision of the first, fourth and last authors. The student clinicians received two days of training in providing treatment, transcription and data collection. Inter-rater reliability after this training was ≥ 85 %. To avoid potential clinician effects, each clinician was randomly allocated to one child from each group and delivered two sessions per day – one child in the KP treatment condition and the other child in the KR feedback condition. The clinicians treated the same children for the 3-week block of treatment. The clinicians were, therefore, aware that treatment involved a comparison of two types of feedback, however, they remained blinded to the research hypotheses. To ensure adherence to the treatment protocol and avoid interference from one feedback condition to the other, treatment fidelity was measured in every session.

Caregivers were informed that their child would be treated using the NDP3 but were blinded to the feedback condition their child was receiving. Caregivers were able to observe treatment via one-way mirrors and could speak to other caregivers in the waiting room. Two of the participating families included twins who were paired with one another and consequently allocated to different treatment groups; therefore, the caregivers from these two families were aware that the nature of the experiment involved manipulation of the feedback conditions. All caregivers remained blinded to the experimental hypotheses and were instructed that no home practice should be done during the study. Reports containing detailed descriptions of the children's treatment condition, goals, progress, beneficial cues and strategies and recommendations for further treatment were provided to the caregivers after the 1-week post-treatment follow up assessment. No stimuli were provided to families and they were requested to refrain from practicing or resuming treatment until after the 1-month post-treatment assessment, which matched Murray et al.'s RCT (2015).

### 2.5. Outcomes

All children completed an individualized experimental probe immediately prior to commencing treatment. Probes varied in length from 116 to 176 items (M = 148, SD = 15.3) and consisted of (a) treated NDP3 items, to test for a treatment effect; and (b) untreated items from the NDP3 Assessment, to test for generalization of any treatment effect. The untreated items represented a range of difficulty in the NDP3 hierarchy from one level below the lowest level of treatment complexity to two levels above the highest level of treatment complexity (see Appendix A). These untreated items were analyzed as a set and not by difficulty level.

Post-treatment assessments were conducted at 1-week, 1-month and 4-months post-treatment as per Murray et al. (2015). At each of these time points, the children completed their experimental probe and the DEAP Inconsistency subtest as an additional measure of generalization. In addition, each child and their caregiver completed a user-experience questionnaire at 1-week post-treatment. At the 1-month post-treatment time point, the GFTA-2 and Single Word Test of Polysyllables were also re-administered. All caregivers reported that their child had received no additional SLP input between the commencement of treatment and the 1-month post-treatment evaluation. Four children in each group reported resuming regular SLP services between 1-month and 4-months post-treatment.

#### 2.5.1. Primary outcome measures

The primary dependent variable was percent accuracy of responses on experimental probe stimuli, judged perceptually. To be judged correct and scored as 1, each word or phrase was required to have: (a) all phonemes produced accurately, including no phonetic distortions, (b) smooth transitions between sounds and syllables (i.e., no syllable segregations or within word groping), and (c) accurate prosody (i.e., lexical or phrasal stress) across syllables. If any error was perceived on sounds, transitions, or prosody, the item was judged incorrect and scored as 0.

#### 2.5.2. Secondary outcome measures

A secondary outcome measure to further explore generalization effects was the score on the Inconsistency subtest of the DEAP. Lower inconsistency scores would indicate improved speech motor control. In addition, responses on the Single Word Test of Polysyllables and GFTA-2 were analyzed to explore potential changes to percent phonemes correct (PPC), percent consonants correct (PCC), percent vowels correct (PVC) and prosodic accuracy (i.e., percent lexical stress match) of untreated single words. Here, higher percentages would indicate improved segmental and prosodic accuracy.

### 2.6. Recording equipment

All treatment sessions were audio- and video-recorded using the Cinde 88 audiovisual system (Cinde, Melbourne, Australia) and the Bosch Video Management System (Bosch Sicherheitssysteme GmbH, Grasbrunn, Germany). In addition, treatment sessions were audio-recorded using within-room digital voice recorders such as the Olympus VN-732PC or Sony Stereo ICD-UX200 F digital voice recorder to enable off-line calculation of treatment fidelity and intra- and inter-rater reliability on the dependent variables. All pre- and post-treatment evaluations were audio- and video-recorded as above as well as audio-recorded using Roland Quad-Capture UA-55 [Roland, Los Angeles, CA] or Avid M-Track Audio [Avid, Burlington, MA] via an adjustable head-worn microphone (AKG C520, AKG Acoustics, Vienna, Austria) at 5 cm mouth-to-microphone distance.

### 2.7. Reliability and treatment fidelity

#### 2.7.1. Treatment sessions

Reliability for judgments of correct/incorrect on response trials was recorded for 25 % of each treatment session. Mean inter-rater reliability was 88 % (SD = 10.3). Treatment sessions were also closely monitored to ensure adherence to the treatment protocol. Data were collected on transcription accuracy, judgements of correct/incorrect, provision of appropriate feedback according to children's allocated treatment group, provision of teaching/cueing where appropriate and eliciting three repetitions of a correctly produced treatment target. These data were compiled to generate an overall measure of treatment fidelity. Mean fidelity was 84.7 % (SD = 9.5).

#### 2.7.2. Experimental probes

Twenty-five percent of each probe assessment was re-rated to determine intra- and inter-rater reliability on primary outcome measures. For point-by-point transcription, mean intra-rater reliability was 89 % (SD = 5.4) and mean inter-rater reliability was 84 % (SD = 6.2). For judgments of correct/incorrect, mean intra-rater reliability was 92 % (SD = 6.1) and mean inter-rater reliability was 87 % (SD = 6.3).

Reliability for point-by-point transcription accuracy was also calculated on 20 % of the secondary outcome data. This included broad transcription of the DEAP inconsistency subtest and phonetic transcription (with diacritics for errors) on the GFTA-2 and Single Word Test of Polysyllables. Mean inter-rater reliability was 85 % (SD = 9.8).

### 2.8. Statistical analysis

All statistical analyses were run using IBM SPSS Statistics 24 for Windows (IBM Corp., 2016). A series of linear mixed effects models were run to test the effects of group (KP, KR), time (pre- and 1-week, 1-month and 4-months post-treatment) and their interaction on (a) treated items, exploring the treatment effect, and (b) untreated but related items, exploring generalization of any treatment effect. First order autoregressive and unstructured models were tested with and without the covariates of age and baseline severity (i.e., PPC score for the Single Word Test of Polysyllables), using Sidak adjustment for multiple comparisons for post hoc testing.

Linear mixed effects modeling with Sidak adjustment was also used to test the effects of group and time on the DEAP Inconsistency scores, the only secondary outcome to be measured at all four time points. To assess for further treatment-related changes in the secondary outcome measures from the Single Word Test of Polysyllables and GFTA-2, repeated measures analysis of variance (ANOVA) was used. This analysis included the between-subjects factor of group (KP, KR) and two-level within-subjects

factor of time (pre, 1-month post) with 95 % confidence intervals and alpha set at .05.

Independent samples t-tests were used to compare average gain (in percentage points) for treatment and generalization words immediately post-treatment between the KP group in this study and the traditional NDP3 group from Murray et al. (2015). Repeated measures ANOVA using group and time (pre- and 1-week, 1-month and 4-months post-treatment) was used to compare long-term outcomes between the two experimental groups here (KP, KR) and the historical comparison group (TRAD) group.

### 2.9. Questionnaire

A 16-item questionnaire was developed using a combination of yes/no, multiple choice, Likert scale and open-ended response types (see Supplemental materials A).

## 3. Results

To assess for treatment and generalization effects, first order autoregressive and unstructured linear mixed effects models were tested with and without the covariates of age and baseline speech disorder severity (i.e., PPC score for the Single Word Test of Polysyllables). In all cases, except for age for the treated items, both covariates were significant. For all dependent variables, the unstructured model including the covariate of severity provided the best fit, with residuals being normally distributed. However, the findings were the same when either covariate was included in the model; note that age and severity were highly correlated in this sample (Pearson $r = .68$, $p < 0.01$). Results for the unstructured models, covarying for severity, are reported here.

### 3.1. Primary outcomes

Performance on treated words across all four time points for the two experimental groups and also for the historical comparison group from Murray et al. (2015) is shown in Fig. 6A. Performance on untreated but related items is shown in Fig. 6B. Means and standard deviations for all measures made over four time points are presented in Table 3. Individual data for all 14 participants for change in percent correct from pre- to immediately post-treatment is also graphed in Fig. 7, for transparency.

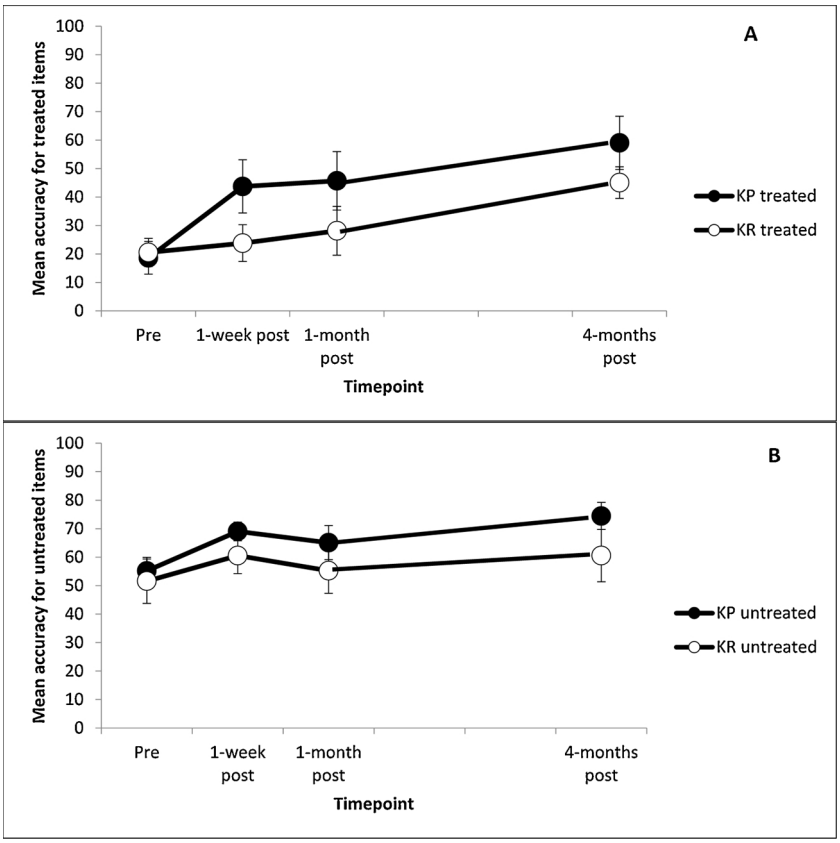


**Fig. 6.** Mean performance at pre-treatment, 1-week, 1-month and 4-months post-treatment for: A. treated items; B. untreated items.
Note: KP = 100 % knowledge of results and performance feedback for all 4 sessions each week; KR = 100 % knowledge of results and performance feedback on session 1 and 100 % knowledge of results feedback on sessions 2–4 each week. Error bars represent standard error.

**Table 3**
Mean treatment and generalization measures across the four test points for children with apraxia of speech assigned to either the Knowledge of Performance (KP) or Knowledge of Results (KR) feedback group.

| | Pre-treatment | | Post-treatment | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 1-week | | 1-month | | 4-months | |
| Treatment group | *KP* *M (SD)* | *KR* *M (SD)* | *KP* *M (SD)* | *KR* *M (SD)* | *KP* *M (SD)* | *KR* *M (SD)* | *KP* *M (SD)* | *KR* *M (SD)* |
| **Primary outcomes** | | | | | | | | |
| Treated items [1] | 18.6 (15.2) | 20.5 (13.0) | 43.7 (24.7) | 23.8 (17.1) | 45.7 (27.2) | 28.1 (22.7) | 59.0 (24.7) | 45.1 (14.6) |
| Generalization items [1] | 55.2 (12.5) | 51.6 (20.6) | 69.0 (8.7) | 60.54 (16.8) | 65.1 (15.9) | 55.2 (20.9) | 74.5 (12.6) | 60.5 (24.3) |
| **Secondary outcomes** | | | | | | | | |
| *DEAP Inconsistency* | 48 (20) | 46.3 (15.6) | 44 (17.4) | 41.7 (13.2) | 39.4 (20.2) | 43.4 (17.5) | 33.1 (16.1) | 38.3 (24.9) |
| *Single-word Test of Polysyllables* | | | | | | | | |
| PPC | 68.9 (19.9) | 62.79 (29.8) | _ | _ | 78.0 (11.5) | 66.0 (23.3) | _ | _ |
| PVC | 72.7 (16.8) | 70.0 (23.2) | _ | _ | 78.4 (9.1) | 67.1 (19.2) | _ | _ |
| PCC | 66.0 (23.2) | 57.4 (34.9) | _ | _ | 77.6 (14.4) | 66.5 (30.1) | _ | _ |
| Percent lexical stress matches | 62.5 (20.6) | 55.9 (28.3) | _ | _ | 61.7 (11.5) | 46.6 (29.7) | _ | _ |
| *GFTA-2* | | | | | | | | |
| Standard score | 73.7 (24.3) | 72.3 (22.0) | | | 78.6 (25.6) | 72.4 (25.8) | | |
| PPC | 75.0 (14.0) | 69.8 (26.8) | _ | _ | 79.9 (11.1) | 72.2 (24.5) | _ | _ |
| PVC | 82.6 (8.9) | 81.8 (16.2) | _ | _ | 83.9 (7.1) | 79.6 (17.2) | _ | _ |
| PCC | 70.8 (17.8) | 77.7 (14.9) | _ | _ | 62.9 (33.4) | 68.1 (30.4) | _ | _ |

Note: [1]Percent correct; DEAP = DEAP = Diagnostic Evaluation of Articulation and Phonology (Dodd et al., 2002); PPC = percent phonemes correct; PVC = percent vowels correct; PCC = percent consonants correct; GFTA-2 = Goldman-Fristoe Test of Articulation – Second Edition (Goldman & Fristoe, 2000).

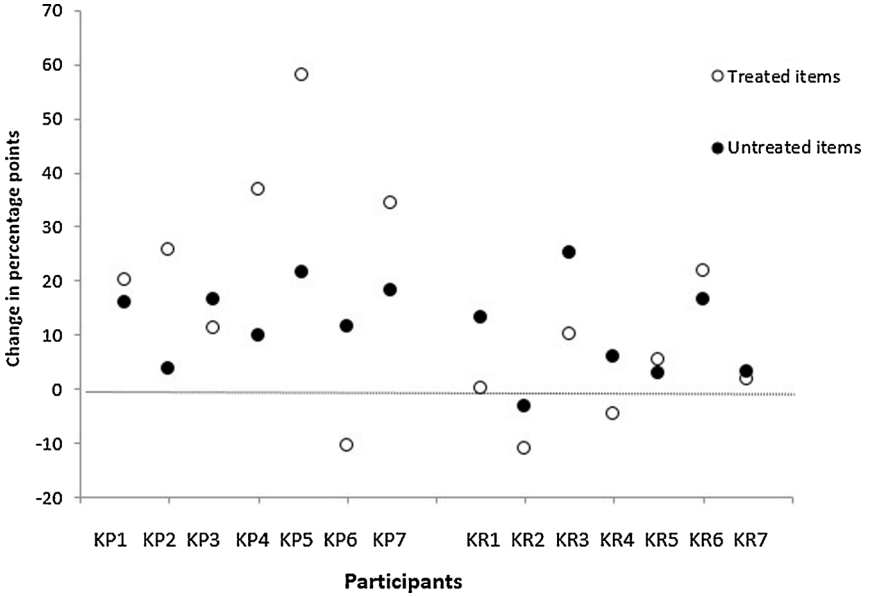


**Fig. 7.** Individual percent change from pre-treatment to 1-week post-treatment for treated items and untreated items.

### *3.1.1. Treatment effects*

There was no statistically significant difference (p = .11, mean difference = 14.7, SE = 8.8) when comparing average percent improvement from baseline for the KP group here (N = 7, mean = 25.1, SD = 21.6) and the traditional NDP3 group (N = 13, mean = 39.8, SD = 17.3) from Murray et al. (2015).

For the two experimental groups here, adjusting for severity, the main effect of time was highly significant; however, the effect of group and the group by time interaction did not reach significance (see Table 4 and Fig. 6A). Due to the exploratory nature of this study, with a relatively small participant sample, post hoc comparisons were explored. For the KP group, there was a significant improvement from pre- to 1-week post-treatment (p = .01, mean difference = 25.1, SE = 6.4), the difference from pre- to 1-month post-treatment approached significance (p = .06, M = 27.0, SE = 9.0) but was robust for the pre- to 4-months comparison

**Table 4**
Type III Tests of Fixed Effects for the dependent measure of treated items produced correctly in the one pre- and three post-treatment probes.

| Source | Numerator df | Denominator df | F | Sig. |
|---|---|---|---|---|
| Intercept | 1 | 12.12 | 2.27 | .16 |
| Group | 1 | 11.53 | 1.69 | .22 |
| Time | 3 | 12 | 19.27 | < .01 |
| Severity | 1 | 11 | 5.03 | < .05 |
| Group * Time | 3 | 12 | 2.08 | .16 |

($p < 0.01$, M = 40.4, SE = 6.2). For the KR group, only the pre- to 4-month comparison reached significance (pre- to 1-week: $p = .99$, M = 3.3, SE = 6.4; pre- to 1-month: $p = .96$, M = 7.6, SE = 9.0; pre- to 4-month: $p = .01$, M = 24.5, SE = 6.2). As shown in Fig. 7, the effect for the KP group was driven by three participants who improved more than 30 percentage points from pre- to 1-week post-treatment.

To explore the issue of statistical power, we conducted a power analysis. First, the effect size (partial eta squared) for the group by time interaction was estimated using a traditional repeated measures ANOVA with group (KP, KR) and time (the first three time points only, free of influence from recommencement of community-based therapy). This yielded an effect size of $\eta_p^2 = 0.18$. To achieve a statistically significant interaction with this effect size, the sample size would need to be 26 per group (total sample size of 52; with alpha 0.05, power 0.8, 2 groups, 3 measurement time points, using G*Power v3.1.9.2). Conversely, with the current total sample size of 14, the effect size would have needed to be 0.36 to reach significance.

Long term outcomes for treated items between the two experimental groups in this study and the historical comparison group from Murray et al. (2015) was explored using repeated measures ANOVA with group (KP, KR, TRAD) and time. There were no significant differences between groups at the 4-months post-treatment time point ($F = 0.46$, $p = .51$).

### *3.1.2. Generalization effect*

Average gain (i.e., percent improvement from baseline) on items expected to generalize was similar ($p = .35$, mean difference = 3.6, SE = 3.7) between the KP group here (N = 7, mean = 13.8, SD = 5.9) and the traditional NDP3 group (N = 13, mean = 10.3, SD = 8.6) from Murray et al. (2015).

Considering the two experimental groups in this study, the first analysis considered the untreated word set. Adjusting for severity, the main effect of time was highly significant; however, the effect of group and the group by time interaction did not reach significance (see Table 5 and Fig. 6B). Again, due to the exploratory nature of the study, post hoc tests with Sidak adjustment for multiple comparisons were examined. Pre-treatment performance was compared to each of the three post-treatment time points for the two groups (see Figs. 6B and 7). For the KP group, there was a significant improvement from pre- to 1-week post-treatment ($p < 0.01$, mean difference = 13.9, SE = 3.0), the pre- to 1-month post-treatment comparison was not significant ($p = .17$, M = 9.9, SE = 4.1), but the pre- to 4-months post-treatment was significant ($p < 0.01$, M = 19.3, SE = 4.0). For the KR group, the pre- to 1-week post-treatment approached significance ($p = .07$, mean difference = 9.0, SE = 3.0), and no other comparisons were significant (pre- to 1-month: $p = .95$, M = 3.6, SE = 4.1; pre- to 4-month: $p = .25$, M = 9.0, SE = 4.0).

Long term outcomes for items expected to generalize were compared between the two experimental groups in this study and the historical comparison group from Murray et al. (2015) using repeated measures ANOVA with group (KP, KR, TRAD) and time. There were no significant differences between groups at the 4-months post-treatment time point ($F = 1.87$, $p = .18$).

### *3.2. Secondary outcome measures: generalization effects*

Outcomes on the DEAP inconsistency subtest demonstrated no significant effects at any time point after adjusting for severity (See Table 6).

Statistical analysis of the four outcome measures derived from the Single Word Test of Polysyllables (PCC, PVC, PPC and percent lexical stress match) and GFTA-2 (PCC, PVC, PPC and Standard Score) demonstrated no group or interaction effect for any measure in either test (see Table 7). For the Single-Word Test of Polysyllables only, there was a large significant main effect of time (pre-treatment to 1-month post-treatment) for PCC ($p < 0.01$, $\eta_p^2 = 0.49$) and a large significant main effect of time for PPC ($p = .04$, $\eta_p^2 = 0.31$) (Cohen, 1969).

**Table 5**
Type III Tests of Fixed Effects for the dependent measure of untreated related items produced correctly in the one pre- and three post-treatment probes.

| Source | Numerator df | Denominator df | F | Sig. |
|---|---|---|---|---|
| Intercept | 1 | 11.61 | 34.66 | < .01 |
| Group | 1 | 11.36 | 3.38 | .09 |
| Time | 3 | 12 | 14.23 | < .01 |
| Severity | 1 | 11 | 108.35 | < .01 |
| Group * Time | 3 | 12 | 1.29 | .32 |

**Table 6**
Type III Tests of Fixed Effects for the dependent measure of Inconsistency Score on the Diagnostic Evaluation of Articulation and Phonology (Dodd et al., 2002) at the pre- and the three post-treatment probes.

| Source | Numerator df | Denominator df | F | Sig. |
|---|---|---|---|---|
| Intercept | 1 | 11.31 | 164.86 | < .01 |
| Group | 1 | 11.15 | .18 | .68 |
| Time | 3 | 12 | 2.92 | .08 |
| Severity | 1 | 11 | 43.32 | < .01 |
| Group * Time | 3 | 12 | .88 | .48 |

**Table 7**
Results of statistical comparisons made for the secondary outcomes measured at only two time points between the Knowledge of Performance (KP) and Knowledge of Results (KR) feedback groups.

| | Pre-treatment to 1-month post-treatment | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Single word test of polysyllables* | **PPC** | | | **PVC** | | | **PCC** | | | **% LS match** | | |
| | *F* | *p* | *Effect* | *F* | *p* | *Effect* | *F* | *p* | *Effect* | *F* | *p* | *Effect* |
| Group | 0.61 | .45 | .05 | 0.61 | .45 | .05 | 0.50 | .49 | .04 | 0.81 | .39 | .06 |
| Time | 5.36 | **.04*** | .31 | 0.21 | .65 | .02 | 11.42 | < **0.01*** | .49 | 1.77 | .21 | .13 |
| Group * Time | 1.24 | .29 | .09 | 1.95 | .19 | .14 | 0.18 | .18 | .02 | 1.28 | .28 | .10 |

| *GFTA-2* | **PPC** | | | **PVC** | | | **PCC** | | | **Standard score** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *F* | *p* | *Effect* | *F* | *p* | *Effect* | *F* | *p* | *Effect* | *F* | *p* | *Effect* |
| Group | 0.38 | .55 | .03 | 0.16 | .70 | .01 | 0.44 | .52 | .04 | 0.10 | .76 | .01 |
| Time | 1.80 | .20 | .13 | 0.03 | .88 | < 0.01 | 3.49 | .09 | .23 | 0.46 | .51 | .04 |
| Group * Time | 0.22 | .65 | .02 | 0.35 | .57 | .03 | 0.08 | .79 | < 0.01 | 0.40 | .54 | .03 |

*Note*. Effect is partial eta squared with .01 representing a small effect, .06 representing a medium effect and .14 representing a large effect. * denotes significant at $p < .05$. GFTA-2 = Goldman-Fristoe Test of Articulation – Second Edition (Goldman & Fristoe, 2000); PPC = percent phonemes correct; PVC = percent vowels correct; PCC = percent consonants correct; LS = lexical stress.

## 4. Discussion

This study compared two methods of feedback during tablet-delivered NDP3 treatment. This investigation is a necessary first step towards determining whether app-delivered right/wrong (KR) feedback during intensive at-home practice of new motor speech targets can effectively facilitate acquisition and maintenance of new segmental and suprasegmental speech patterns. Such technology has the potential to bridge the gap between optimal service delivery intensity in CAS and current service delivery models in Australia.

We hypothesized that (i) tablet-based delivery of NDP3 using high frequency KP feedback would obtain similar treatment and generalization outcomes to Murray et al.'s (2015) traditional paper-based delivery of NDP3, (ii) participants in the high frequency KR condition may demonstrate smaller gains immediately post-treatment (i.e., evidence of slower acquisition and generalization), compared with the KP group, but greater maintenance at 1- and 4- months post-treatment (i.e., evidence of more robust learning), and (iii) the experimental groups would demonstrate similar long-term maintenance of any treatment and generalization effects to Murray et al.'s (2015) traditional NDP3 delivery.

### *4.1. Treatment effects*

Our first hypothesis was confirmed in that the KP group made statistically significant gains in treated and untreated word accuracy, which were similar in magnitude to the traditional NDP3 treatment group from Murray et al. (2015). Our second hypothesis was partially supported. Overall, for both treated and untreated words, no group effect was detected but the effect of time was highly significant. This suggests that children in both experimental groups responded to the treatment, with positive gains in treated and untreated words over time, regardless of the feedback condition. However, on closer examination of the data from the individual children, it was noted that 6/7 children in the KP group made substantive gains on treated words of 10 or more percentage points from pre- to 1-week post-treatment, while only 2/7 children in the KR group did; and 3/7 K P children but no KR children improved > 30 percentage points. Similarly, 6/7 K P children and only 3/7 KR children showed a 10 or more percentage point gain for untreated words, indicative of generalization. It is likely that the small sample size in this study meant insufficient power to detect a significant group by time interaction effect. Our power analysis suggested that a group size of 26 was needed to achieve a significant interaction for the treated words, or else a larger effect size of 0.36. To date, this is the only study that has examined the influence of

feedback type on speech intervention in CAS. These data suggest that the influence of KP vs KR feedback needs to be further explored in a larger sample, to determine whether there is indeed an effect of feedback type or whether the differences observed are driven by other factors such as age, severity of CAS, or self-evaluation ability.

The lack of significant improvement for the KR group immediately post-treatment appears consistent with the tendency for slower improvement with KR than KP. Although the KP group's accuracy on both treated and untreated but related items at 1-week post-treatment reached significance, while the KR groups did not, there were no significant differences between the groups at any time point. This is likely because variability within groups was large, as shown in Fig. 7.

### 4.2. Maintenance of treatment effects

Regarding the third hypothesis, both tablet-delivered treatment groups had made similar long-term gains at 4-months post-treatment that were statistically significant compared to pre-treatment performance level and similar to the gains made by the traditional NDP3 group in Murray et al. (2015). However, this also suggests that evidence of a significant treatment effect for the KP group here should be interpreted with caution. If treatment was responsible for accelerated changes in the KP group's speech production skills, one might expect that their progression over time should remain accelerated when compared with the KR group. This was not the case. Instead, the KR group demonstrated similar achievements in speech production skills at the 4-month follow up assessment. This finding may be confounded by (i) the return to community-based treatment for 4/7 children in both groups and (ii) that community clinicians were likely to have been providing KP feedback, although we do not have any evidence to support this suggestion. The lack of significant treatment effect for the KR group, also makes it difficult to attribute the improved performance at the 4-month follow up to 'maintenance'.

### 4.3. Generalization effects

The overall trend in improvement on both treated words and generalization words differed between this study and the historical comparison study (Murray et al., 2015). Whereas, the traditional NDP3 group showed large improvement on treated words immediately post-treatment with a tendency towards loss of skill at follow up due to 1/13 clients having poor maintenance (Murray, McKechnie, & Williams, 2017) both groups here continued an upward trajectory during the follow up period. Reasons for this are not clear but may be due to factors to do with the use of the tablet for stimulus presentation, audio recording or self-evaluation, or the reinstatement of community-based therapy for some children. In contrast, performance on generalization words showed the opposite effect. Where the traditional NDP3 group showed a continuous upward trend in performance accuracy, the two experimental groups here showed similar gains in untreated real words immediately post-treatment, with a trend towards deterioration of skill at 1-month post-treatment, followed by continued improvement from 1-month to 4-months post-treatment. Given that children were able to return to their regular speech-language pathology treatments following the 1-month post-treatment evaluation, this could explain the continued long-term improvements on all items. However, approximately half of all children did not resume treatment in this period and so it is likely additional but unidentified factors contributed to the trend for continued longer-term improvement. This is a desirable trend warranting further investigation into which child-related or treatment-related factors may have contributed to this observation.

These results echo those of previous studies in CAS and AOS that have demonstrated that responses to different feedback types and frequencies vary across participants (Maas et al., 2012; Austermann Hula et al., 2008). Variation in response to different feedback types and frequency may be influenced by strength of internal representation of the specific speech behaviors targeted and/or pre-treatment level of proficiency. Target selection was individualized for each participant, resulting in some treatment and/or generalization targets being relatively more difficult than others. Stimulus selection may therefore have served as a confound within and between participants (Maas et al., 2012; Wambaugh et al., 2017). This confound is almost impossible to avoid in these studies as treatment must address the sounds in error for each individual child. This is mitigated in part by limiting the sounds to those that are stimulable for a correct response and selection of three goals crossing different levels of proficiency (e.g., single sound to word level). Nonetheless, the children still vary in their ability to self-evaluate, ease in production, ability to attend and comply with the training context, and their motivation.

### 4.4. Limitations and future directions

The sample size of the study was small and within-group variability was large, thus limiting the power of our statistical analyses. CAS is relatively rare (Shriberg, Aram et al., 1997; Shriberg, Austin et al., 1997) but much larger sample sizes may be possible with multi-center collaboration. Our power analysis suggests that a sample size of about 26 per group is desirable and this would also allow exploration of other child-related factors that might influence or predict response to intervention. Alternatively, larger scale analyses may be possible through meta-analyses of studies which have used similar outcome measures. In addition, we acknowledge that the use of nonverbal measures of cognitive capacity would have provided a better assessment of cognitive skills than the PPVT-4 and more comprehensive cognitive testing is warranted in future to assess potential factors influencing treatment response.

Future research should explore alternative feedback type and frequency conditions and combinations. The feedback frequency and schedule used for the KR group in this study involved 100 % pre-practice with KR + KP on day 1 and 100 % practice with KR only on days 2–4 and was designed to mimic the common Australian service delivery model of once per week face-to-face with a clinician with a home-practice program with less rich feedback from an app or a parent. This model deviates from the schedule used

in our previous work with PML, wherein a period of pre-practice with KR + KP is provided at the beginning of *every* session, and the child only progresses to practice with KR alone when they reach a predetermined threshold of success (Ballard et al., 2010; Iuzzini & Forrest, 2010; McCabe et al., 2014). It is possible that the children in the KR group here did not receive sufficient pre-practice to develop a stable internal reference of correctness. This could explain why predominantly KR feedback appeared less effective than KR + KP feedback in stimulating improvement at 1-week post-treatment in this study.

It is also possible that the effects of feedback type were mediated by the frequency of feedback. High frequency feedback was used here, even though low frequency feedback has been recommended in the PML approach (Schmidt & Lee, 2011). This was in order to examine the effect of KR versus KP feedback types without the potential confound by potentially positive effects of low frequency feedback. However, high frequency feedback has been demonstrated to increase response variability if participants continually change their performance in different ways each time they are presented with feedback on error (Wulf & Shea, 2004). The within-group variability observed in this study may have been related to the high frequency feedback schedule. There was some suggestion that other aspects of the guidance hypothesis were supported, however, in that high frequency feedback guides the individual towards the correct response and that performance accuracy decreases when feedback is withdrawn (Salmoni, Schmidt, & Walter, 1984). That is, the drop in average accuracy such that performance was no longer significantly higher than pre-treatment across the entire sample may have been related to removal of feedback post-treatment. In contrast, the findings here reflect findings from some motor learning studies in children where higher frequency feedback has been shown to lead to maintenance of skill post-treatment (Sullivan, Kantak, & Burtner, 2008) and longer-term retention (Chiviacowsky, Wulf, de Medeiros, Kaefer, & Wally, 2008). Clearly, the influence of type and frequency of feedback on motor learning in children and how these principles may interact with specific task and child factors, is still not entirely clear.

Evidence from Iuzzini and Forrest (2010), who demonstrated variable reinforcement schedules only effected changes in accuracy during the third week of treatment, even after establishing a threshold of success, suggests that the KR group in this study may have benefited from a longer treatment period in order to establish acquisition of targets; or a longer period of pre-practice (Miller, Plante, Ballard, & Robin, 2018). Future research could explore whether the KR-based practice needs to be delivered for longer duration, or for more trials, to obtain a similar level of acquisition to KR + KP-based practice, and consequently to greater maintenance and generalization of these gains. Alternatively, a more gradual progression from predominantly KR + KP feedback into predominantly KR feedback (see Strand et al., 2006), gradual transition from immediate high-frequency feedback to delayed and reduced frequency feedback (Ballard et al., 2010; McCabe et al., 2014; Schmidt & Lee, 2011), or feedback fading based on successful execution of speech targets may be beneficial. One suggestion is to structure the feedback schedule beginning with three sessions of KR + KP and one session of KR in the first week, gradually progressing to one session of KR + KP and three sessions of KR in the third week of treatment. A similar gradual shift was employed by Thomas et al. (2017) who explored parent-training in ReST treatment as a method of achieving recommended intervention frequency for children with CAS, albeit with limited success.

Another factor that may have influenced the findings here was that clinicians were instructed to shift back to KR + KP feedback when children in the KR group produced five sequential incorrect productions of their selected treatment words. This was necessary in order to uphold our ethical duty of care for the children involved in the treatment, as extended intensive practice of incorrect motor plans could be harmful for learning as well as for motivation and engagement. This was monitored so that the ratio of KR and KP feedback over the study for these children was maintained at 80 % KR to 20 % KP trials. In clinical practice, such apps would typically be recommended for supervised use in the home environment. Clinicians would engage in progress monitoring and intervene, when required, in order to either provide coaching for the parent to assist their child to achieve more difficult speech production targets or to schedule a clinic visit in order to provide some additional pre-practice and KP-style feedback. For example, we have now implemented a threshold system where the therapy app discontinues delivery of a specific stimulus after a set number of incorrect responses (Hair et al., 2018), allowing a parent or clinician to step in and provide additional coaching with KP. Future research is needed to explore within-participant factors in order to determine which children would be most suited to intensive practice with high frequency KR-style home practice conditions as delivered in this study.

The home practice condition was simulated in this study, as clinicians delivered all feedback. This was done for two main reasons. First, this maintained experimental control. Secondly, while automated speech analysis algorithms running offline on computers are becoming more accurate at identifying errors in children's speech (Shahin, Ji, & Ahmed, 2018), software that can run in real-time on a tablet is less sophisticated. Our speech analysis software for the tablet had not yet been sufficiently developed to meet clinically acceptable levels of reliability with human perceptual judgment and so was not incorporated into the tablet when this study was conducted. In response to the participants' feedback about the need for greater interactivity and variety of games, the research team is continuing to develop a wider range of games and alternative ASR algorithms in order to improve the gaming quality of an app designed for speech behavior change. The team are currently trialing the effectiveness of a new app using integrated ASR to determine the effectiveness of tablet-delivered treatment and ASR-generated feedback in a real home setting.

## 5. Conclusions

Mobile technology has the potential to increase the engagement and motivation of clients and to facilitate intensive practice of speech production targets (e.g., Hair et al., 2018). Combined with less frequent direct clinical contact via face-to-face sessions or telehealth, it can also mitigate barriers of distance and access to services for rural and remote families. With continued advancements in technology and the development and integration of accurate and reliable ASR software, mobile games are likely to become an effective supplement to face-to-face intervention. This has particular benefit for older children who can then practice independently and take greater responsibility for their remediation. It may also be helpful for some parents who find it difficult to provide reliable

feedback on their child's productions (Thomas et al., 2017, 2018). However, further research is required to understand how the parameters of therapy can change, and therefore the effectiveness of that therapy, with app-based exercises and with ASR versus parent or clinician generated feedback. The current study suggests that provision of predominantly KR feedback on speech accuracy yielded small and perhaps negligible gain compared to primarily KP feedback; however, for the 3-week block of therapy, gains under the primarily KP feedback were not well-maintained. In building apps, it is important to build in flexibility so that practice can adhere to appropriate motor learning principles that may vary depending on the age and skill level of the child and that stimulate optimal long-term learning in a time and cost-effective manner. Additional research is required to develop algorithms for prescribing these variations in practice and feedback conditions for children with CAS.

## CRediT authorship contribution statement

**Jacqueline McKechnie:** Software, Investigation, Data curation, Visualization, Supervision, Formal analysis, Writing - original draft, Writing - review & editing. **Beena Ahmed:** Conceptualization, Software, Resources, Writing - review & editing, Funding acquisition. **Ricardo Gutierrez-Osuna:** Conceptualization, Software, Resources, Writing - review & editing, Funding acquisition. **Elizabeth Murray:** Investigation, Supervision, Writing - review & editing. **Patricia McCabe:** Methodology, Investigation, Data curation, Writing - review & editing, Supervision. **Kirrie J. Ballard:** Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing - review & editing, Visualization, Supervision, Funding acquisition.

## Acknowledgements

This research was made possible by NPRP Grant # 8-293-2-124 and 4-638-2-236 (Ahmed, Gutierrez, Ballard) from the Qatar National Research Fund (a member of the Qatar Foundation) and an Australian Postgraduate Award (McKechnie). During this project, Ballard was supported by an Australian Research Council Future FellowshipFT120100355. The statements made herein are solely the responsibility of the authors. The authors wish to thank the participants and families for their commitment to the treatment program during school vacation periods. Thanks also to the student clinicians who delivered the intervention and research assistants who contributed to data collection.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jcomdis.2020.106026.

## References

American Speech-Language-Hearing Association (2007). *Childhood apraxia of speech [Technical report]*Retrieved from Rockville (MD)http://www.asha.org/policy/TR2007-00278.

Austermann Hula, S. N., Robin, D. A., Maas, E., Ballard, K. J., & Schmidt, R. A. (2008). Effects of feedback frequency and timing on acquisition, retention and transfer of speech skills in acquired apraxia of speech. *Journal of Speech Language and Hearing Research, 51*, 1088–1113.

Ballard, K. J., Robin, D. A., McCabe, P., & McDonald, J. (2010). A treatment for dysprosody in childhood apraxia of speech. *Journal of Speech Language and Hearing Research, 53*(3), 1227–1245.

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., & Morris, M. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language, 37*, 98–128.

Chiviacowsky, S., Wulf, G., de Medeiros, F. L., Kaefer, A., & Wally, R. (2008). Self-controlled feedback in 10-year-old children: Higher feedback frequencies enhance learning. *Research Quarterly for Exercise and Sport, 79*(1), 122–127.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences.* New York; USA: Academic Press.

Dodd, B. L., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2002). *Diagnostic evaluation of articulation and phonology.* London: England: The Psychological Corporation.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th. ed.). New York, USA: Pearson Inc.

Edeal, D. M., & Gildersleeve-Neumann, C. E. (2011). The importance of production frequency in therapy for childhood apraxia of speech. *American Journal of Speech-language Pathology, 20*, 95–110.

Edwards, J., & Dukhovny, E. (2017). Technology training in Speech-Language Pathology: A focus on tablets and apps. *Perspectives of the ASHA Special Interest Groups, SIG 10, 2*(Part 1), 33–48.

Furlong, L., Erickson, S., & Morris, M. (2017). Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders, 68*, 50–69.

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. *PloS One, 13*(8), Article e0201513. https://doi.org/10.1371/journal.pone.0201513.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe test of articulation* (second ed.). Minneapolis, MN: Pearson.

Gomez, M., McCabe, P., & Purcell, A. (2018). Clinical management of childhood apraxia of speech: A survey of speech-language pathologists. *Paper Presented at the Speech Pathology Australia National Conference*.

Gordon-Brannan, M. E., & Weiss, C. E. (2007). *Clinical management of articulatory and phonologic disorders.* Baltimore, MD: Lippincott Williams & Wilkins.

Gozzard, H., Baker, E., & McCabe, P. (2004). *Single word test of polysyllables.* Unpublished work.

Gozzard, H., Baker, E., & McCabe, P. (2008). Requests for clarification and children's speech responses: Changing "pasghetti" to "spaghetti". *Child Language Teaching and Therapy, 24*, 249–263.

Hair, A., Monroe, P., Ahmed, B., Ballard, K. J., & Gutierrez-Osuna, R. (2018). Apraxia World: A speech therapy game for children with speech sound disorders. *Paper Presented at the Interaction Design and Children (IDC)*.

IBM Corp (2016). *IBM SPSS statistics for windows (Version 24.0).* Armonk, NY: IBM Corp.

Iuzzini, J., & Forrest, K. (2010). Evaluation of a combined treatment approach for childhood apraxia of speech. *Clinical Linguistics & Phonetics, 24*(4-5), 335–345.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., ... Shadish, W. R. (2010). *Single-case designs technical documentation.* Retrieved fromhttp://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Lim, J. M., McCabe, P., & Purcell, A. (2017). Challenges and solutions in speech-language pathology service delivery across Australia and Canada. *European Journal for Person Centred Healthcare, 5*(1), 120–128.

Maas, E., & Farinella, K. A. (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech Language and Hearing Research, 55*(2), 561–578.
Maas, E., Butalla, C. E., & Farinella, K. A. (2012). Feedback frequency in treatment for childhood apraxia of speech. *American Journal of Speech-language Pathology, 21*, 239–247.
Maas, E., Gildersleeve-Neumann, C. E., Jakielski, K. J., & Stoekel, R. (2014). Motor-based intervention protocols in treatment of childhood apraxia of speech. *Current Developmental Disorders Reports, 1*(3), 197–206.
Maas, E., Robin, D. A., Austerman Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., ... Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-language Pathology, 17*, 277–298.
McAllister, L., McCormack, J., McLeod, S., & Harrison, L. J. (2011). Expectations and experiences of accessing and participating in services for childhood speech impairment. *International Journal of Speech-language Pathology, 13*(3), 251–267.
McCabe, P., Macdonald-D'Silva, A., van Rees, L., Ballard, K. J., & Arciuli, J. (2014). Orthographically sensitive treatment for dysprosody in children with Childhood Apraxia of Speech using ReST intervention. *Developmental Neurorehabilitation, 17*(2), 137–146.
McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-language Pathology,* 1–17. https://doi.org/10.1080/17549507.2018.1477991.
McLeod, S., & Baker, E. (2017). *Children's speech: An evidence-based approach to assessment and intervention.* Boston, MA: Pearson.
Miller, H., Plante, A., Ballard, K. J., & Robin, D. (2018). Treatment of lexical stress, segmentation, and sound distortions in childhood apraxia of speech. *Paper Presented at the International Conference on Motor Speech*.
Morgan, A. T., Murray, E., & Liégeois, F. J. (2018). Interventions for childhood apraxia of speech. *The Cochrane Database of Systematic Reviews*(5), https://doi.org/10.1002/14651858.CD006278.pub3.
Moriarty, B. C., & Gillon, G. (2006). Phonological awareness intervention for children with childhood apraxia of speech. *International Journal of Language & Communication Disorders, 41*(6), 713–734.
Murray, E., McCabe, P., & Ballard, K. J. (2012). A comparison of two treatments for childhood apraxia of speech: Methods and treatment protocol for a parallel group randomised control trial. *BMC Pediatrics, 12*, 112–120.
Murray, E., McCabe, P., & Ballard, K. J. (2014). A systematic review of treatment outcomes for children with childhood apraxia of speech. *American Journal of Speech-language Pathology, 23*, 486–504.
Murray, E., McCabe, P., & Ballard, K. J. (2015). A randomized controlled trial for children with childhood apraxia of speech comparing Rapid Syllable Transition Treatment and the Nuffield Dyspraxia Programme - Third Edition. *Journal of Speech Language and Hearing Research, 58*, 669–686.
Murray, E., McKechnie, J., & Williams, P. (2017). Exploring factors for treatment success in childhood apraxia of speech using the nuffield dyspraxia programme - 3rd edition. *Paper Presented at the Speech Pathology Australia National Conference*.
Namasivayam, A. K., Pukonen, M., Goshulak, D., Hard, J., Rudzicz, F., Rietveld, T., ... van Lieshout, P. (2015). Treatment intensity and childhood apraxia of speech. *International Journal of Language & Communication Disorders, 50*, 529–546.
Nordness, A. S., & Beukelman, D. R. (2010). Speech practice patterns of children with speech sound disorders: The impact of parental record keeping and computer-led practice. (Report). *Journal of Medical Speech-language Pathology, 18*(4), 104–108.
Olswang, L. B., & Bain, B. A. (2013). Treatment research. In L. A. C. Golper, & C. Frattali (Eds.). *Measuring outcomes in speech-language pathology* (pp. 245–264). (2nd ed.). New York, N Y: Thieme Medical.
Parnandi, A., Karappa, V., Lan, T., Shahin, M., McKechnie, J., Ballard, K., & Gutierrez-Osuna, R. (2015). Development of a remote therapy tool for childhood apraxia of speech. *ACM Transactions on Accessible Computing, 7*(3), https://doi.org/10.1145/2776895.
Parnandi, A., Karappa, V., Son, Y., Shahin, M., McKechnie, J., Ballard, K., & Gutierrez-Osuna, R. (2013). Architecture of an automated therapy tool for childhood apraxia of speech. *Paper Presented at the Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*.
Robbins, J., & Klee, T. (1987). Clinical assessment of oropharyngeal motor development in young children. *The Journal of Speech and Hearing Disorders, 52*, 271–277.
Ruggero, L., McCabe, P., Ballard, K. J., & Munro, N. (2012). Paediatric speech-language pathology service delivery: An exploratory survey of Australian parents. *International Journal of Speech-language Pathology, 14*(4), 338–350.
Salmoni, A., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin, 95*, 355–386.
Schmidt, R. A., & Lee, T. D. (2011). *Motor control and learning: A behavioural emphasis* (fifth ed.). Champaign, IL: Human Kinetics.
Semel, E., Wiig, E., & Secord, W. (2006). *) clinical evaluation of language fundamentals, Australian standardised* (4th ed.). Sydney, Australia: Pearson.
Shahin, M., Ji, J. X., & Ahmed, B. (2018). One-class SVMs based pronunciation verification approach. *Paper Presented at the International Conference on Pattern Recognition*.
Shriberg, L. D., Potter, N. L., & Strand, E. A. (2009). *Childhood apraxia of speech in children and adolescents with glacatosemia*. November *Paper Presented Ar the American Speech-Language-Hearing Association National Convention*.
Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997). Developmental apraxia of speech: I. Descriptive and theoretical perspectives. *Journal of Speech Language and Hearing Research, 40*(2), 273–285.
Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech Language and Hearing Research, 40*(4), 708–722.
Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile app rating scale: A new tool for assessing the quality of mobile health apps. *JMIR MHealth and UHealth, 3*(1), e27.
Strand, E. A., Stoekel, R., & Baas, B. (2006). Treatment of severe childhood apraxia of speech: A treatment efficacy study. *Journal of Medical Speech-language Pathology, 14*(4), 297–307.
Sugden, E., Baker, E., Munro, N., & Williams, A. L. (2016). Involvement of parents in intervention for childhood speech sound disorders: A review of the evidence. *International Journal of Language & Communication Disorders, 51*(6), 597–625. https://doi.org/10.1111/1460-6984.12247.
Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2017). An Australian survey of parent involvement in intervention for childhood speech sound disorders. *International Journal of Speech-language Pathology,* 1–13. https://doi.org/10.1080/17549507.2017.1356936.
Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *International Journal of Language & Communication Disorders, 53*(4), 718–734. https://doi.org/10.1111/1460-6984.12399.
Sullivan, K. J., Kantak, S. S., & Burtner, P. A. (2008). Motor learning in children: Feedback effects on skill acquisition. *Physical Therapy, 88*, 720–732.
Thomas, D. C., McCabe, P., & Ballard, K. J. (2014). Rapid Syllable Transitions (ReST) treatment for Childhood Apraxia of Speech: The effect off lower dose-Frequency. *Journal of Communication Disorders, 51*, 29–42.
Thomas, D. C., McCabe, P., & Ballard, K. J. (2017). Combined clinician-parent delivery of rapid syllable transition (ReST) treatment for childhood apraxia of speech. *International Journal of Speech-language Pathology,* 1–16. https://doi.org/10.1080/17549507.2017.1316423.
Thomas, D. C., McCabe, P., Ballard, K. J., & Bricker-Katz, G. (2018). Parent experiences of variations in service delivery of Rapid Syllable Transition (ReST) treatment for childhood apraxia of speech. *Developmental Neurorehabilitation, 21*(6), 391–401. https://doi.org/10.1080/17518423.2017.1323971.
Thomas, D. C., McCabe, P., Ballard, K. J., & Lincoln, M. (2016). Telehealth deliery of Rapid Syllable Transitions (ReST) treatment for childhood apraxia of speech. *International Journal of Language & Communication Disorders, 51*(6), 654–671.
Toki, E. I., & Pange, J. (2010). E-learning activities for articulation in speech language therapy and learning for preschool children. *Procedia Social and Behavioural Sciences, 2*, 4274–4278.
Wambaugh, J. L., Nessler, C., Wright, S., Mauszycki, S. C., DeLong, C., Berggren, K., ... Bailey, D. J. (2017). Effects of blocked and random practice schedule on outcomes of Sound Production Treatment for acquired Apraxia of Speech: Results of a group investigation. *Journal of Speech Language and Hearing Research, 60*, 1739–1751.
Wiig, E., Semel, E., & Secord, W. (2006). *Clinical evaluation of language fundamentals preschool, Australian and New Zealand standardised* (2nd ed.). Sydney, Australia: Pearson.

Williams, P., & Stephens, H. (2004). *The nuffield dyspraxia programme* – (third edition). Windsor, England: The Miracle Factory.
Williams, P., & Stephens, H. (2010). The nuffield Centre dyspraxia programme. In A. L. Williams, S. McLeod, & R. J. McCauley (Eds.). *Interventions for speech sound disorders in children* (pp. 159–177). Baltimore, MD: Brookes.
Wulf, G., & Shea, C. (2004). Understanding the role of augmented feedback: The good, the bad and the ugly. In M. Williams, N. Hodges, & M. Scott (Eds.). *Skill acquisition in sport: Research, theory and practice*. Florence, KY, USA: Routledge.