Visual Discrimination of Cantonese Tone by Tonal but Non-Cantonese Speakers, and by Non-Tonal Language Speakers[1]

Denis Burnham, Susanna Lau, Helen Tam, and Colin Schoknecht

Macarthur Auditory Research Centre, Sydney (MARCS), University of Western Sydney, Australia

d.burnham@uws.edu.au; http://www.uws.edu.au/marcs/

Abstract

A previous study by the first two authors suggests there is visual information for tone perception: under certain conditions Cantonese speakers are able to identify spoken words as one of six Cantonese words differing only in tone on the basis of lip and face movements at a rate better than chance [1]. Here, non-native (tonal, Thai, and non-tonal, English) language speakers were tested on a discrimination version of this task in three modes: auditory-visual (AV), auditory only (AO), and visual only (VO). Auditory stimuli were presented either in the clear or accompanied by noise. In all conditions, even VO, performance was significantly better than chance. In the clear, both English and Thai perceivers performed better in the AO and AV conditions than in the VO condition. With auditory noise added, Thai perceivers performed better in the AV than the AO condition. The results support the existence of visual information for tone, and show that this is available in the absence of experience with the language in question, and even in the absence of experience with the lexical use of tone.

1. Introduction

It is now established that speech is an auditory-visual event: whenever visual (lip, face, head, neck movement) information is available, humans use it. A powerful example is the classic McGurk effect, in which English speakers perceive auditory [ba] paired with visual [ga] as "da" or "tha" [2]. So accurate information about *visemes* as well as *phonemes* is important in speech perception. Put another way, convergent information about the *speech source* is important, and is better provided by auditory-visual than by auditory information alone [3].

In what can be called the language-influenced McGurk effect [4], the degree of incorporation of visual information in auditory-visual speech perception differs over languages. For example, it has been found that Japanese perceivers use visual information less than do English perceivers [5], and furthermore that Cantonese perceivers incorporate visual information even less than do Japanese speakers [6,7].

This complex of results has been suggested to be due to the relative prevalence of linguistically relevant lexical pitch in the three languages [6,7]: English only uses pitch information to distinguish lexical items in the relatively rare stress contrasts, e.g., 'project vs pro'ject; Japanese has two pitch-accents which distinguish meaning on two-syllable words, high-low, e.g., ka⌐ki [oyster], and low-high, e.g., ka⌊ki [persimmon]; while Cantonese is tone-rich – it has six lexical tones, high, low-mid/high-rising, mid, low-mid/low-falling, low-mid/mid-rising, and low-mid. It may be argued that as the degree of lexically-relevant pitch variation in a language increases, the reliance on visual information declines. This argument is consistent with the obtained cross-language data, and may well be true. However, it does not necessarily imply that there is no visual information for lexical tone.

In our previous study we tested Cantonese language participants who were either phonetically-trained or phonetically-naïve. A native Cantonese speaker presented words differing only in tone, not segments. Participants were required to choose the correct word from six written words on the screen in one of three modes: auditory-only (AO), visual-only (VO), or auditory-visual (AV). There were three findings of interest with regard to the VO condition: (1) phonetically-naïve observers performed significantly above chance and significantly better than their phonetically-trained counterparts; (2) irrespective of phonetic training, VO performance was significantly above chance for words containing monophthongs (/fu/ and /fan/), and significantly better for these than for those for words containing diphthongs (/soej/ and /hau/); and (3) irrespective of phonetic training, VO performance was significantly above chance if the target words were presented in a sentence and significantly better for these than for words presented in isolation.

These results suggest that there is visual information in the face for tone, although the exact nature of this information is yet to be determined. Some indication of the source of this information may be gleaned from what is known about visual information for *intonation*, which occurs across sentences, rather than within words. A significant correlation has been found for French speakers between speech intonation, and their eyebrow movements [8]. In more comprehensive studies using the OPTOTRAK system, strong correlations

between head movements and F_0 have been found [9,10].

The question of interest here is whether the visual perception of tone is due to language-specific learned associations, or whether the information is generally available to observer not *au fait* with the target language. To test this, adults with two levels of naiveté of Cantonese were employed – Thai speakers, who were familiar with the lexical use of tone (Thai has 5 lexical tones) but unfamiliar with Cantonese, and Australian English speakers who were unfamiliar both with lexical tone and with Cantonese.

To establish the design of this study, the results of our previous study [1] were of interest. Previously we found that performance was better for phonetically-naïve participants. Only phonetically-naïve participants were tested here.

Our previous study found that performance was better for words containing monophthongs than words containing diphthongs. Two words, one containing a monophthong, /fan/, and one containing a diphthong, /soej/, were used here.

In our previous study, performance was generally better for contour tones in which some pitch movement occurred than for level tones in which minimal pitch movement occurs. The contour tones were low-mid/high-rising (2-5), low-mid/low-falling, (2-1), and low-mid/mid-rising (2-3), while the level tones are high (5-5), mid (3-3), and low-mid (2-2), where the numbers refer to the relative level of the pitches involved. In this study all possible pairs were included.

Finally, in our previous study we found that performance was better for words in sentences than words in isolation. Unfortunately we could not investigate words in sentences here due to the need to use a discrimination rather than an identification paradigm. In the identification paradigm single words (or sentences) were presented and the participants are asked to identify (either by open verbal response or by choosing one of the possible alternatives) which of the words was presented. This could not be done here because the participants could not understand the target language, Cantonese, and so could not report or recognise the correct alternative. Thus a same-different AX paradigm was used, in which two words are presented in succession and participants must indicate whether they are the same or different.

Two inter-stimulus intervals (ISIs) were used in the current experiment, 500 msec and 1500 msec. The decision to use these was based upon studies with auditorally-presented stimuli [11,12], which have shown that at 500-msec ISI listeners are able to make same-different discriminations for two speech sounds on the basis of phonetic (language non-specific) information, whereas at 1500-msec, due to memory constraints, participants make use of language-specific phonemic categories to make speech sound discriminations.

The final variable of interest in this study was the inclusion of auditory noise. This was not manipulated in our previous study. There, performance was consistently equivalent and relatively high in the AO and the AV conditions. To allow any augmentation of auditory tone information by visual tone information to be shown here, auditory noise was introduced in half the experiments.

Four separate experiments were conducted - with Thai and Australian English speakers, and with clear auditory presentations and added auditory noise. Due to their tone language background, Thai speakers were expected to perform better than Australian English speakers in all conditions, AO AV, and V-only. Performance in the no-noise (clear) conditions was expected to be generally better than in the auditory noise conditions. Within each of the experiments the following hypotheses were entertained:

Visual-only: Performance in the VO condition was expected to be above chance in both the noise and no noise experiments for both the Thai and Australian English speakers (see [1]).

Visual Augmentation: (a) With clear auditory presentations it was expected that AV and AO performance should be statistically equivalent, and significantly better than in the VO condition. (b) In auditory noise it was expected that performance in the AV condition should be significantly better than AO performance, i.e., that there should be augmentation of performance due to the addition of visual information (see [1]).

Vowels: Performance on the monophthong word, /fan/, was expected to be better than on the diphthong word, /soej/ (see [1]).

Tone Type: Performance was expected to be better to the extent that contour tones were involved, i.e., best for contour-contour discrimination pairs and then better for contour-level pairs than level-level tone pairs (see [1]).

ISI: Especially as participants (Thai and Australian English) did not speak the target language (Cantonese), performance was expected to be better for short ISI of 500 ms (presumably allowing phonetic processing), than the longer ISI of 1500 ms (at which phonemic processing must presumably be involved) (see [11, 12]).

2. Method

2.1 Design

In four separate experiments, native Thai speakers and native Australian English speakers were tested in an auditory clear experiment, and in an auditory noise experiment. In each experiment a 2 x 2 x (3 x 15 x 4) design was employed. The first between-subjects factor was the *vowel type* of the stimulus

words, either a monophthong, in the minimal tone sextuplets of /fan/, or a diphthong, in the minimal tone sextuplets of /soej/. The other between subject factor was the duration of inter-stimulus interval, 500ms or 1500ms.

The within-subjects variables were *mode of presentation* - auditory-only (AO), visual-only (VO), auditory-visual (AV); *tone pair* - the 15 possible pairings of the six Cantonese tones; and *repetitions* - with each tone pair being presented four times. Given that there are three *level* (high (5-5), mid (3-3), and low-mid (2-2)), and three *contour* (low-mid/high-rising (2-5), low-mid/low-falling (2-1), and low-mid/mid-rising (2-3)), Cantonese tones, of the 15 possible tone pairings there are 3 level-level tone pairs, 3 contour-contour tone pairs, and 9 level-contour tone pairs. The repetition factor was included so that order and same/different pairings were controlled. For example, given a pair of tone words A and B, 4 pairs of words associated with this tone pair were presented, two different trials (AB, BA), and two same trials (AA, BB) trials. In this way, a discrimination index (similar to d' in signal detection terms) could be calculated for each of the 15 tone pairs. Thus the dependent variable was discrimination index (DI) given by DI = *hits* {number of correct responses ("different" responses on AB or BA trials, and "same" responses on AA or BB trials)} minus *false positives* {number of incorrect responses ("different" responses on AA or BB trials, and "same" responses on AB or BA trials)} / 4. This yields a maximum score of +1 (perfect responding), and a minimum of –1, with chance level being 0.

2.2 Stimulus Materials

Stimuli consisted of two Cantonese tone sextuplets, one with a monophthong vowel, /fan/, and one with a diphthong vowel, /soej/. Each of these phonetic strings has a lexical meaning for each of the six Cantonese tones, and therefore each comprises a Cantonese tone sextuplet. Each of these 12 words was presented in AO, VO, or AV modes. The stimuli were spoken by a 23-year-old native Cantonese female, and recorded on a digital video-recorder. These were then digitally edited into digital video files (Apple Quicktime) using Adobe Premiere. For the auditory noise experiments multi-talker babble was added with a signal-to-noise ratio of –0.6 dB.

In each word condition, /fan/, /soej/, there was a total of 180 test trials (AO/VO/AV x 15 tones pairs x 4 repetitions) in six 30-trial test blocks. The 6 blocks were split into three presentation modes: 2 AO blocks, 2 VO blocks, 2 AV blocks. The first three blocks always included each of the three modes, AV, AO, and VO, and the order of these was counterbalanced between subjects. In the final three blocks the three presentation modes were presented in the same order. As there were a total of 60 trials for each presentation mode, 30 trials were presented in each block, and were counterbalanced with respect to tone type, repetition type, and order.

Twelve training trials were included at the start of testing to allow participants to become familiar with the testing procedure and the requirements of the experiment. All 12 training trials were presented in the auditory only modality, and without background noise both in the auditory clear and auditory noise experiments. There were 6 same (AA or BB) trials, and 6 different (AB or BA) trials, with equal representation of the six tones across the practice trials and appropriate order counterbalancing.

Following the 12 training trials, participants were familiarised with the three modes of presentation by giving them 2 AO, 2 VO and 2 AV practice trials in succession. (For participants in the auditory noise experiments, the practice trials included background noise.) After these 6 practice trials, the 6 blocks of 30 test trials began. At the beginning of each block, there were 3 buffer items, the responses for which were discarded in the data analyses.

2.3 Procedure

Participants were tested individually in a sound-attenuated room on a PC running the DMDX experimental software [13]. Participants were instructed to press the right shift key on the keyboard of the PC if they perceived the two tones to be the same, and the left shift key if they perceived the two tones to be different. Participants were asked to make their responses as quickly and as accurately as possible. Both response type (correct and incorrect) and reaction time data were collected. Only correct response data are presented here, converted into discrimination indices (see above).

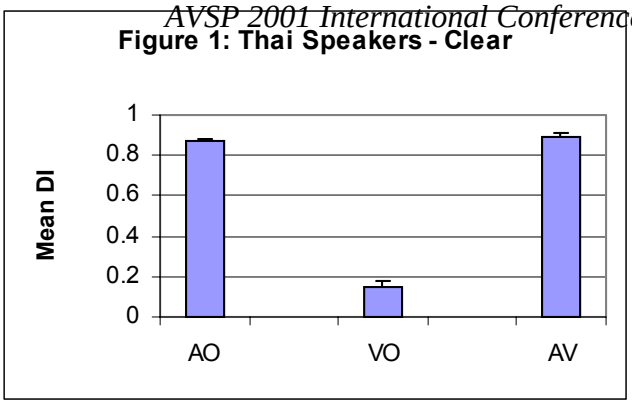
3. Experiment 1: Thai Speakers – Clear Audio

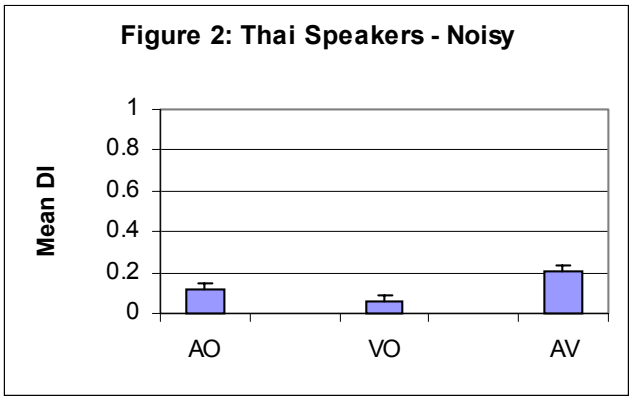
3.1 Participants

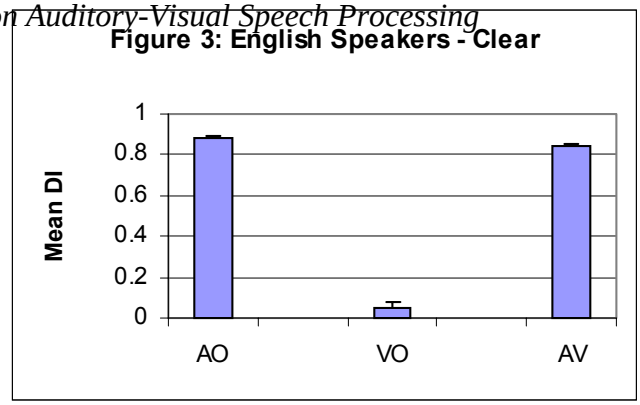
24 native Thai speakers, 17 females and 7 males, all in their early to mid 20s were recruited from Chulalongkorn University in Bangkok Thailand. Of these 12 served in the /fan/ condition, and 12 in the /soej/ condition, and 6 in each word condition served in 500 msec ISI condition and 6 in the 1500 msec ISI condition.

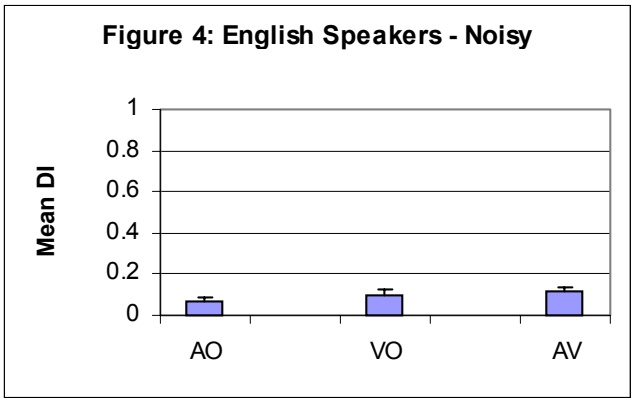
3.2 Results

The results for the three modes of presentation are shown in Figure 1. Performance was above chance level in all three conditions, AO, VO, and AV, $t(360) = 54.99, 5.43, 76.93$, respectively, $p<.01$ for









all tests. Thus, even in the VO condition, discrimination performance was above chance. However there was no augmentation of performance in the AV condition compared with the AO condition. There were no overall effects of vowel or tone type but these two factors interacted significantly, $F(1,20) = 13.41$: for the monophthong /fan/, performance was better on level-level (Mean = .69) than contour-contour (Mean = .54) tone pairings, whereas for the diphthong /soej/, performance was slightly better on contour-contour (Mean = .64) than level-level (Mean = .61) tone pairings. However, as this did not interact with mode of presentation, it is not of great interest here. There were no significant effects due to ISI level.

4. Experiment 2: Thai Speakers – Noisy Audio

4.1 Participants

24 native Thai speakers from Chulalongkorn University, 18 females and 6 males, all in their early to mid 20s were allocated to word and ISI conditions as in Experiment 1.

4.2 Results

The results for the three modes of presentation are shown in Figure 2. Participants performed above chance in all three modes, $t(360) = 4.94$, 2.50, and 8.65 for AO, VO, and AV, $p<01$, .05, .01, respectively. In addition there was augmentation due to addition of visual information, as can be seen by the comparison of AV and AO conditions in Figure 2, $F(1,20) = 11.60$. There was general augmentation of AV over AO across the board, but for /fan/ the augmentation was particularly marked at 500 msec ISI, while for /soej/ there was no augmentation at 500 msec, and augmentation was particularly marked for 1500 msec ISI, $F(1,20) = 11.60$. Apart from this effect, there were no further effects of ISI, or vowel type. There were no significant effects of tone type.

5. Experiment 3: Australian English Speakers – Clear Audio

5.1 Participants

24 native Australian English speakers, 16 females and 8 males, all in their early to mid 20s were recruited from the University of Western Sydney, Sydney, Australia. Of these 12 served in the /fan/ condition, and 12 in the /soej/ condition, and 6 in each word condition served in 500 msec ISI condition and 6 in the 1500 msec ISI condition.

5.2 Results

The results for the three modes of presentation are shown in Figure 3. Performance was again above chance in all three modes, $t(360)_{AO} = 67.70$, $p<.01$, $t(360)_{VO} = 2.05$, $p<.05$, $t(360)_{AV} = 58.06$, $p<.01$. However, there was no visual augmentation effect, and in fact there was a significant effect in the opposite direction, $F(1,20) = 7.89$, i.e., participants performed better with just the auditory information than when visual information was also provided (see AO vs AV in Figure 3). There were no effects due to ISI, vowel, or tone type.

6. Experiment 4: Australian English Speakers – Noisy Audio

6.1 Participants

24 native Australian English speakers from the University of Western Sydney, 19 females and 5 males, all in their early to mid 20s were allocated to word and ISI conditions as in Experiment 3.

6.2 Results

The results for the three modes of presentation are shown in Figure 4. Again performance was significantly above chance in all three conditions, $t(360)_{AO} = 2.82$, $t(360)_{VO} = 3.87$, $t(360)_{AV} = 4.43$, $p<.01$ for all three. Although some visual augmentation can be seen in Figure 4 for AV vs AO, this effect was not significant, $F(1,20) = 1.31$. Neither were there significant effects due to vowel type, tone type, or ISI.

7. General Discussion and Conclusions

As can be seen in Figures 1-4, performance was much better in the clear conditions than in the auditory noise conditions. An overall analysis across language groups revealed better performance by the Thai speakers than Australian English speakers in the clear audio condition, $F(1,40) = 4.49$; and this difference approached significance in the noisy audio condition, $F(1,40) = 4.05$, $F_{critical} = 4.09$. However, the degree of superiority was not as great as might have been expected; the Australian English speakers performed quite well, given their unfamiliarity with the lexical use of tone (see Figures 1 and 2 versus Figures 3 and 4).

Across the four experiments there were some effects of ISI (500 vs 1500 msec), type of tone pair (level-level, contour-contour, and level-contour), and type of vowel in the word (monophthong, diphthong) for the Thai speakers, but these did not support the hypotheses, and did not detract from the effects of main interest here, those regarding the visual perception of lexical tone.

In all four experiments performance was significantly better than chance in all three modes, auditory-only, auditory-visual, *and* visual-only. The latter is notable - both tonal language (Thai), and non-tonal language (Australian English) speakers were able to discriminate between tone pairs presented only in the visual modality, even when the language was unfamiliar to them.

In addition, in the noise added condition, Thai speakers performed better in the AV than the AO condition, showing that under noisy conditions, visual information augments the perception of tone. For the Australian English speakers, this effect, while in the correct direction did not reach significance. It is possible that this is due to a floor effect – scores were much lower for Australian English than Thai speakers. Perhaps the level of noise, added to the unfamiliarity of the task for non-tonal language speakers, supressed performance to such an extent that the additional visual information was not useful. A further study with Australian English speakers in which a lower relative level of noise is used may resolve this issue.

In the clear audio experiment, Australian English speakers performed better in the AO than the AV condition, the opposite to the expected result. Thus, despite an ability to perceive visual information for tone (see VO versus chance results) Australian English speakers are apparently reluctant to use this visual information when other sources are available and reliable. It is possible that, given the novelty of the task of discriminating tones is speech, Australian English speakers found the addition of visual to the auditory information distracting. However, to the extent that the Thai speakers also performed better in AO than AV conditions, this effect cannot be due to the unfamiliarity of the tones *per se*. Rather, it seems it may be due to the unfamiliarity of the language. This issue may be resolved if data from Cantonese speakers were collected. This is currently being done.

It is interesting to note that despite any bias towards disregarding visual information evident when clear audio AV vs. AO conditions are compared, participants still perform above chance in the VO conditions. Thus there *is* visual information for tone, and perceivers use this when it is the only source of information available. Participants, at least the Thai speakers here, also use visual tone information when auditory noise is present.

The results confirm those of our previous identification study with Cantonese speakers [1], and extends them to perceivers who know no Cantonese, and who have no experience with tonal languages. In both this and our previous study [1], we have now found a small but significant effect of the visual perception of tone. In the earlier study we found better performance by phonetically-naïve than phonetically-trained perceivers, suggesting that the visual perception of tone may be amenable to attentional and learning processes. Here we have found visual augmentation of tone perception when the auditory information is degraded. These findings together suggest that hearing-impaired perceivers in a tone language should be especially sensitive to the visual information for tone.

8. References and Notes

[1] Burnham, D., Ciocca, V., Lauw, C., Lau, S., & Stokes, S (2000) Perception of visual information for Cantonese tones. In M. Barlow & P. Rose (Eds) *Proceedings of the Eighth Australian International Conference on Speech*

Science and Technology, Australian Speech Science and Technology Association, Canberra, 2000, pp 86-91.

[2] McGurk, H., & MacDonald, J. Hearing lips and seeing voices. *Nature*, 264, 746-748, 1976.

[3] Vatikiotis-Bateson, E., Kuratate, T., Munhall, K. G., & Yehia, H. C. The production and perception of a realistic talking face. In O. Fujimura, B. D. Joseph, & B. Palek (Eds.), *Proceedings, LP'98, Item order in language & speech 2* (439-460). Prague: Charles University (Karolinum Press), 2000.

[4] Burnham, D., & Sekiyama, K. (in preparation) Investigating auditory-visual speech perception development using the ontogenetic and differential language methods. In E. Vatikiotis-Bateson, P. Perrier, & G. Bailly, G. (Eds.). Advances in auditory-visual speech processing. Cambridge: MIT Press.

[5] Sekiyama, K., & Tohkura, Y. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Amer., 90*, 1797-1805, 1991.

[6] Hayashi, Y., & Sekiyama, K. Native-foreign language effect in the McGurk effect: a test with Chinese and Japanese. *Proceedings of the International Conference on Auditory-Visual Speech Processing. Sydney,* 61-66, 1998.

[7] Sekiyama K. Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, *59*, 73-80, 1997.

[8] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espressor, R. About the relationship between eyebrow movements and F_0 variations. In In T. Bunnell & W. Idsardi (Eds) *Proceedings of the 4th Internat. Conf. Spoken Language Processing. 4,* 2175-78, 1998.

[9] Vatikiotis-Bateson, E., & Yehia, H. Physiological modeling of facial motion during speech. *Trans. Tech. Comm. Psychol. Physiol. Acoustics*, H-96-65, 1-8, 1996.

[10] Yehia, H., Kuratate, T., & Vatikiotis-Bateson, E. Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics*, in press.

[11] Werker, J.F., & Logan, J.S. 1985. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37, 35-44.

[12] Werker, J.F., & Tees, R.C. 1984b. Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of* America, 75, 1866-1878.

[13] To download DMDX go to:

http://www.u.arizona.edu/~jforster/dmdx.htm

[1] The assistance of Ms Susanna Lau in recording stimuli and writing the DMDX program, Mr Colin Schoknecht in augmenting the DMDX program and assisting with data collection, Ms Helen Lam in data collection and data analysis and Kuhn Phanintra and Kuhn Sorabud in data collection is greatly appreciated.