The short and long-term effects of pronunciation instruction

GRAEME COUPER – Auckland University of Technology

ABSTRACT
There has been surprisingly little exploration of the effectiveness of pronunciation teaching by researchers. This classroom-based study makes an initial step towards addressing this gap by determining the immediate effect of instruction on specific forms in second language (L2) pronunciation and the extent to which gains were retained over time and integrated into phonological competence. The subjects were New Zealand (NZ) immigrants, largely of Asian origin, attending a high-intermediate level English class. The study focused on epenthesis (the addition of an extra sound, usually a schwa, after a consonant) and absence (the inappropriate dropping of a consonant sound). The instruction was explicit and involved a series of short input and practice sessions interspersed amongst the regular teaching over a period of two weeks. Dramatic gains were achieved: the average error rate dropped from 19.9% to 5.5% in the immediate post-test, and rose slightly to 7.5% in the delayed post-test. A more general test conducted at the same time as the delayed post-test showed the error rate had halved, indicating that these gains had been largely integrated into learners’ phonological competence. By way of comparison, similar groups of students, who did not receive the instruction, achieved no gains in this area of pronunciation during the course of one semester. This suggests that appropriately focused instruction can lead to changes in learners’ phonological interlanguage even where this may appear to have fossilised.

Introduction
The study aims to provide evidence to support earlier findings (Couper 2003) that pronunciation teaching does work. In this earlier study, a systematic explicit pronunciation syllabus was incorporated into the teaching program of a full-time English as a Second Other Language (ESOL) class. It was found to be both popular and successful, but there was insufficient empirical evidence to make claims about exactly what had worked and why.

This study attempts to address these shortcomings in three ways. First, it focuses on just one aspect of pronunciation, that is epenthesis and absence, which makes it easier to measure how pronunciation has changed. Second,
the rationale and type of instruction have been defined more precisely. This is an important first step in trying to generalise from results and to allow others to reproduce the study. Finally, the effects have been measured immediately and over time, and baseline data has been collected to enable a comparison with changes to the interlanguage of similar learners. I have chosen to use the term ‘baseline data’ rather than ‘control group’, as I believe it is more appropriate in the context of this study. The results are then used to draw practical conclusions for classroom teaching.

One of the difficulties is that research to date has given teachers little guidance as to how they should teach pronunciation, or indeed if they should teach it at all. As Derwing and Munro (2005: 382) point out, pronunciation ‘remains a very marginalised topic in Applied Linguistics’. This has in turn meant that there have been very few studies that focus on the effectiveness of pronunciation teaching (Derwing and Munro 2005: 387). This lack of research and consequent lack of training has meant that some teachers have serious misgivings about the effectiveness of teaching pronunciation at all, while others lack the skills and confidence to tackle it in their classes (Macdonald 2002; Breitkreutz, Derwing and Rossiter 2002). It would be far too ambitious to expect to develop a set of principles and guidelines for teachers as a result of a single study, but it is hoped that the findings reported here might contribute to the development of a set of generalisable principles. This is why I have gone to some effort to describe the teaching approach used in this study. The actual teaching practices can then be taken into account in discussing the effectiveness of the teaching.

Teachers will have observed that in some cases students will still get pronunciation wrong even after listening and repeating an item or feature many times. So the first goal is to see if learners can be taught specific items and if they can retain this learning. The second goal is to see if they can then transfer this learning to other items contextualised at the sentence level. At a further stage, one would also want to see if they can also produce them correctly in less-controlled situations, but that is beyond the scope of this study. Therefore the study poses these research questions:

1. Is instruction effective in teaching specific items?
2. Are gains retained over time?
3. Can these gains be transferred to other items in controlled sentence-level contexts?
Literature review

There have been very few studies into the effectiveness of pronunciation instruction (Derwing and Munro 2005) despite its obvious importance for successful communication. Pennington (1998) and Fraser (2000) argue that phonology both can and should be taught to adult learners and call for more research.

There have been some classroom-based studies. Derwing, Munro and Wiebe found a positive effect for instruction, which ‘focused on general speaking habits as opposed to a concentration on individual segments’ (1997: 217). Derwing, Munro and Wiebe (1998) also found that both instruction in segmental accuracy, and instruction in general speaking habits and prosodic features, led to improved pronunciation. Couper (2003) also found pronunciation instruction to be effective, but was unable to provide evidence as to the causes of this improvement.

Earlier studies failed to provide convincing evidence. For example, Suter (1976) was not able to find a positive effect for instruction. However, type of instruction was ignored, even though it is clearly a key factor in its effectiveness. While Acton (1984) reports in detail on a program of instruction focusing on the link between pronunciation, affect, personality and social context, which was designed to help learners whose pronunciation had fossilised, no empirical evidence of its success is offered. Macdonald, Yule and Powers (1994) attempted to measure the short-term effects of different types of instruction, but found that learner differences introduced a very important variable, which was hard to control for, and they were not able to draw clear conclusions.

However, short-term effectiveness of pronunciation instruction has been established in tightly controlled situations where a number of studies have been able to show positive effects on the production of individual phonemes, prosody or overall fluency (e.g. Neufeld 1977; Murakawa 1981; de Bot and Mailfert 1982; de Bot 1983; Leather 1990; Champagne-Muzar, Schneiderman and Bourdages 1993). Studies in speech perception (for example Strange 1995) have explored the role of training to help learners adjust their perception of the L2 phonology, which may lead to improvements in production.

The aspect of pronunciation looked at in this research, epenthesis and absence, has been the subject of considerable study in terms of the influence of a universal grammar and first language (L1) transfer on L2 development. Here it has played a key role in the development of a number of theories of phonological acquisition (Leather 1999). However, there have only been a few studies that have attempted to describe changes to this aspect of
interlanguage over time (Hansen 2001; Hansen 2004). The literature relevant to this aspect of L2 acquisition will not be explored in depth here, as the focus is on the effectiveness of instruction. Epenthesis and absence are merely useful examples taken for this study. A detailed review is provided by Hansen (2004). I have not been able to find any studies that have considered the effect of teaching on epenthesis and absence.

Although there has been little research in pronunciation teaching, one can reasonably transfer some of the suggestions from other areas of research. Ideas that teachers should consider are the role of consciousness- and awareness-raising in L2 acquisition (for example Sharwood Smith 1981); the effectiveness of explicit instruction (Spada 1997); the relationship of L2 pronunciation cognitive factors rather than physical ones (Fraser 2000); the difficulty in changing the way in which the adult brain has established ways of categorising sounds (Leather 1999); and listening and comparing one's own production with a model can help to re-set phonological boundaries (Lively and Pisoni 1995).

The study

METHOD

Initially, the treatment group was given a General Diagnostic Test to determine the most suitable area of pronunciation to focus on (see Appendix A). A few days later, this was followed by a test with a specific focus on epenthesis and absence for both speaking and listening (see Appendixes B and C). There were then a number of short teaching sessions, spread over two weeks, followed by the Specific Test again. This test was given a third time, 12 weeks later, at the end of the semester, when the general diagnostic test was also given again. The baseline group sat both the general diagnostic and the speaking tests early in the semester and then again 12 weeks later at the end of the semester. The focus of this paper is on the results of the speaking tests, as they provided the most reliable data.

THE PARTICIPANTS

The participants in the treatment group were 21 ESOL learners attending a full-time English language class in Auckland, NZ. This class ran from July to November 2003. They were all immigrants to NZ, with L1s as follows: Chinese, 14; Korean, 1; and the remainder with a wide range of non-East Asian languages. On average, they had been in the country for two and a half years (ranging from zero to eight) and were 32.5 years of age (ranging from 20 to 57). As they were not all available for all of the tests, the comparative data is based only on those who did all the tests: the Specific
Test was taken all three times by 17 students, the General Diagnostic Test was taken both times by 14 students, and the Listening Discrimination Test was taken all three times by 15 students.

The baseline data was collected from four high-intermediate classes, which ran one year later, from July to November 2004. A total of 50 learners completed both tests at the two different times. Again, the majority was from China (26) and Korea (13), with two further East Asian speakers, one from Thailand and the other from Japan. The remainder came from a wide range of non-East Asian countries. The average age was 34 years (ranging from 19 to 54) and they had been in the country for an average of three years (ranging from zero to twelve) at the beginning of the course.

The participants from both groups were all judged to be at the same overall level of proficiency, as they had been assigned to a high-intermediate level class (approximately 4.5 to 5.0 on an IELTS [International English Language Testing System] scale). Although the treatment group and the baseline group are similar, given the large number of uncontrollable variables in any classroom setting, I have preferred the term baseline rather than control group.

THE TESTS

The aim of this study was to find evidence for the effectiveness of instruction. The focus of this instruction was determined by the General Diagnostic Test (Appendix A) administered at the beginning of the course. The results of this test were explained to learners to give them some guidance as to their areas of difficulty with pronunciation. They were also used to determine which difficulties were most pervasive, and accordingly which would be most appropriate for instruction and testing of potential short-term and long-term gains.

The results of the diagnostic test were analysed for a number of different types of errors. At the phoneme level, errors were found throughout the whole spectrum of sounds. The phonemes that caused the largest number of learner difficulties (about 80 % or more) were the nasals (alveolar versus velar), voiceless fricatives (alveolar versus dental) and the voiced alveolar stop (/d/) versus the voiced dental fricative.

On the suprasegmental level, there were often problems with sentence stress and rhythm. In particular, the subjects’ speech often sounded very disjointed or broken. This level was where most of the learners had some difficulty. Unfortunately, these sorts of errors are quite difficult to quantify. However, it is relatively easy to count the number of times an extra sound has been inappropriately added and the number of times a sound is missing
when it should not be. Adding these up showed that all of the learners had at least some difficulty with this. While these features may sometimes seem to affect the segmental level only, they have major repercussions on the suprasegmental level, as they impact on the number of syllables, cause words not to be combined fluently, and interfere with the timing and rhythm of English. As difficulties of this type were represented to at least some degree in all respondents, and as they affect intelligibility, epenthesis and absence of final consonants was chosen as the focus of this study.

The results of the General Diagnostic Test were then used to construct a more thorough test of epenthesis and absence. The types of errors made were noted to provide guidance in the writing of the test. It soon became clear that the majority of difficulties involved word boundaries. In other words, it was a matter of sounds in connected speech as much as consonant clusters within a single word.

As can be seen from the speaking test, Specific Test (Appendix B), items 1 to 8 involved pronunciation of past tenses, items 9 to 16 involved pronunciation of third person and plural ‘s’, and items 17 to 24 involved consonant clusters with /nd/, or /ld/ and particularly with ‘and’. The other questions attempted to cover a variety of consonant clusters and were taken from language that had been used in class in the previous week.

The Specific Test was carried out in the language laboratory. Learners were given a cassette and the text and asked to record it. They were expected to be familiar with all the words, so they were not practised beforehand.

The Listening Discrimination Test (Appendix C) was recorded on a tape. Learners were only able to listen once. It was devised to cover the same sort of range of difficulties and phonological contexts and attempted to use language in the distracters, which could also make sense in the right context.

THE TEACHING

As the teacher/researcher, my beliefs about the teaching and learning of pronunciation inevitably exerted a significant influence on how the classes were conducted. These beliefs could be summarised as follows:

• Many learners are not aware of their pronunciation problems. Making these clear to learners is a very important first step to solving them.

• Many learners cannot hear the difference between their pronunciation and that of the target language. They need to learn to hear the language more like the way native speakers do. This requires a great deal of repetitive practice to help learners establish precisely where those
phonological boundaries are. It also involves learners recording themselves and playing back this recording so they can compare their voices with the model.

- Explanations can be helpful, but they are not always effective, as it is often very difficult to get learners to focus on the right thing. In other words, teachers must recognise that learners’ phonological concepts will be different from those of the native speaker, and that it is therefore critical that teachers make sure their students understand the metalanguage they are using.

- Sometimes getting learners to discover patterns or rules can be helpful. In this case there are some useful rules, such as the pronunciation of plural or third person ‘s’ and regular past tenses.

- Learners need a lot of practice and feedback to improve.

There were a number of different kinds of lessons in my instruction program. In total there were 12 sessions of about 30 minutes each. (A detailed schedule is provided in Appendix D.) The first lesson involved returning the tests and explaining them. This was done, for example, by writing ‘asked a questions’ and ‘asked questions’ on the board, and then ‘intermedia students’ and ‘intermediate students’. Learners were then asked to say the two versions side by side. I then modelled them and the learners repeated them. The individual learners would say one and I would point at the one I had heard. I always explained this in terms of ‘this is how it sounds to me’.

Once it was becoming clear to the students what the problem was, they then listened to their tapes. For this activity I had cut and pasted the learners’ recordings from the test so that I had examples from four or five different students for each item on the test. So, I chose one who had said it correctly, one who had left out consonants and one who had added a schwa. I also included one or two others who had made various other combinations of errors with epenthesis and absence. As a whole class, we then listened to each item and the learners attempted to work out which was the best version. They also tried to discern what was wrong with the other versions. Even with my guidance, they found this difficult. Having done this, they then practised saying the phrases in pairs, and trying to get them right.

They also had the chance to listen and repeat, and to record themselves in the language laboratory. Here they used the materials from the listening test. They were able to hear all of the possible answers and record themselves after each one. They could then play it back as many times as they wanted to. Some still said they could not hear the difference between ‘baked fish’ and ‘baked a fish’.
A number of approaches were used to try to explain how to pronounce the target language correctly. Some time was spent explaining how to join words so that some of the problems of epenthesis could be avoided. Some things, such as ‘and’, appeared to be very easy to explain. Many learners felt they had to say ‘and’ in its strong form and added epenthesis in the process. By modelling ‘and’ in different contexts, they were able to hear that they did not have to say the final ‘d’, and that it was then much easier to say the following word, too. Also, in modelling final consonants, especially when the coda consisted of two–three consonants, learners were also able to start to hear the difference. They would often refer to making the consonant shorter or lighter. Those who appeared to be dropping the final consonant often felt they were not, and, in terms of their concepts, talked about making the final consonant stronger or longer.

Time was also spent on getting learners to discover the rules for past tense endings and plural and third person endings. Here, exercises from Headway were used (Soars, Soars, Bowler et al 1990: 1, 84–85).

Finally, an attempt was made to explain the concept of syllables. Learners were asked to work out the possible numbers of consonants in the onset and coda and find examples of different patterns. This approach to teaching the concept of syllables is similar to that described in Celce-Murcia, Brinton and Goodwin (1996). I found that for many learners, this concept was not really understood. It is easy to forget that the concept of what does and does not constitute a syllable varies widely across languages, and is consequently quite difficult to understand.

The effects of teaching: analysis of data

The results have been converted from the number of errors to a percentage error rate, as this is generally considered to provide a better basis for comparison purposes. Theoretically it is not possible to say precisely how many errors could be made, so percentages have been arrived at by estimating the number of syllables that could conceivably involve errors of epenthesis and/or absence. This was set at 100 for the General Diagnostic Test and 110 for the Specific Test. The number of errors was then converted to the percentage of actual over potential errors. In the first instance, data was analysed for changes in individual performance. This provides a comparison for each individual learner’s performance at three points in time: prior to instruction, immediately after instruction and three months after instruction. This analysis was done on the Specific Test for speaking and the listening test.
Table 1: Reliability of testing instruments

<table>
<thead>
<tr>
<th>Pair</th>
<th>Test Description</th>
<th>N</th>
<th>Correlation</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1</td>
<td>Pre-test: General Diagnostic with Specific Test</td>
<td>19</td>
<td>.891</td>
<td>.000</td>
</tr>
<tr>
<td>Pair 2</td>
<td>General Diagnostic Test: Pre-test and Delayed.</td>
<td>14</td>
<td>.912</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 2: Effect of instruction on speaking for each student, as measured by Specific Test

Specific Test: Speaking

Error rate (percentages) by student (S) at Time 1 (T1), Time 2 (T2) and Time 3 (T3).

<table>
<thead>
<tr>
<th>S</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>45</td>
<td>29</td>
<td>27</td>
<td>26</td>
<td>25</td>
<td>23</td>
<td>22</td>
<td>22</td>
<td>18</td>
<td>18</td>
<td>16</td>
<td>16</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>T2</td>
<td>11</td>
<td>9</td>
<td>3</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T3</td>
<td>17</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>15</td>
<td>6</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3: Effect of instruction on listening, for each student

Error rate for listening by student at Time 1, 2 and 3 (percentages)

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>38</td>
<td>14</td>
<td>33</td>
<td>29</td>
<td>38</td>
<td>29</td>
<td>29</td>
<td>14</td>
<td>33</td>
<td>24</td>
<td>29</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Time 2</td>
<td>14</td>
<td>14</td>
<td>38</td>
<td>24</td>
<td>19</td>
<td>14</td>
<td>19</td>
<td>19</td>
<td>5</td>
<td>0</td>
<td>14</td>
<td>19</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Time 3</td>
<td>14</td>
<td>19</td>
<td>19</td>
<td>29</td>
<td>0</td>
<td>19</td>
<td>19</td>
<td>10</td>
<td>10</td>
<td>29</td>
<td>33</td>
<td>29</td>
<td>10</td>
<td>5</td>
<td>19</td>
</tr>
</tbody>
</table>

The second stage of analysis involved comparing mean error rates on the Specific Test over the same three points in time, and on the General Diagnostic Test at times one and three. These results were then compared with the baseline data collected later. As a further step, the mean error rate was also analysed for those learners who made more than the median number errors on the Specific Test at time 1.

Reliability and validity

As is shown in Table 1, paired samples correlations between the General Diagnostic and the Specific Test, and between the General Diagnostic Test at the beginning and end of the semester, show significant correlations of about 0.9, suggesting that we can be confident both instruments were test-
ing the same things and the performance variability was not a significant issue.

Tests were marked without reference to earlier test results to avoid influencing the objectivity of the marker.

The second marking of a selection of the tests provided a further check on the reliability of the marking. A sample of the pre-tests, immediate post-tests and delayed post-tests were transferred from tape to computer, to make it easier for a colleague to listen to them and mark them. This was done for all three Specific Tests for eight of the subjects, ones ranging from average to the worst, and the colleague was asked to choose at random the sound files for five of the eight students. A Pearson correlation of 0.981 (N=15) was found, significant at the 0.01 level. This suggests one can be reasonably confident in the uniformity of marking, even though there is an inherent subjectivity in relying on an auditory analysis.

Findings

CHANGES IN INDIVIDUAL PERFORMANCE

As can be seen in Table 2, all learners made substantially fewer speaking errors after instruction, and even though the error rate increased slightly again over time, it was still well below the rate prior to instruction.

Learners also made fewer mistakes in listening, with the mean error rate dropping from 24.1% to 14.0% in the immediate post-test. A paired samples t-test shows that this was significant (p < 0.05, t = 3.872, df = 14, sig = 0.002). Much of this improvement was lost by the end of the semester when the error rate increased again to 18.1%. Compared with the pre-test, this only just makes the significance level at the 95% level (p < 0.05, t = 2.179, df = 14, sig = 0.047). However, as can be seen in Table 3, this improvement was not consistent on the individual level, either for the immediate post-test or the delayed post-test. Support for this observation is found in the lack of a significant correlation in the paired samples correlations for the t-test. Between the pre-test and immediate post-test, the correlation was 0.486 (sig = 0.067), pre-test and delayed post-test it was 0.342 (sig = 0.212), and between the immediate post-test and the delayed post-test it was 0.181 (sig = 0.518).

CHANGES IN OVERALL PERFORMANCE IN COMPARISON WITH BASELINE DATA

As can be seen in Figure 1, the mean error rate for the treatment group on the Specific Test dropped from 19.9% to 5.5% immediately after instruction. A paired samples t-test shows this to be a significant decrease (p < 0.05, t = 8.069, df = 16, sig = 0.000). The error rate had increased
again to 7.5% by the end of the semester, three months later. However, since this is still less than half the pre-instruction rate, this was still a significant decrease in the error rate over the pre-test results \((p < 0.05, t = 7.631, df = 16, \text{sig} = 0.000)\). The increase in errors between the immediate post-test and the delayed post-test was also significant \((p < 0.05, t = -3.082, df = 16, \text{sig} = 0.007)\). To determine just how meaningful these changes are, an effect size has been calculated. This has been done by using a ‘d’ statistic, which is arrived at by dividing the mean difference found on the paired samples test by the standard deviation. An effect of 0.2 is considered small, 0.5 and above is considered medium, and 0.8 and above is considered large (Green and Salking 2005). The effect size, \(d = 1.95\), was greatest between the pre-test and the immediate post-test. Also, comparing the pre-test with the delayed post-test produced \(d = 1.85\). Clearly, both of these results imply extremely large effect sizes. The increase in the error rate between the immediate and delayed post-tests at \(d = 0.74\) was also quite large. By way of comparison, there was a slight increase in the rate of epenthesis and absence over a one-semester period for the baseline group, from 14.1% to 15.2%. A paired samples t-test found this was significant \((p < 0.05, t = -2.510, df = 49, \text{sig} = 0.015)\), but a further test for the effect size at \(d = 0.35\) also suggests that for all practical purposes this was not very meaningful.

The results of the General Diagnostic Test found a similar improvement for the treatment group, as the error rate almost halved from 13.9% to 7.5%. This was also significantly different on a paired samples t-test \((p < 0.05, t = 4.643, df = 13, \text{sig} = 0.000)\). The effect size was also very large at \(d = 1.24\). By way of comparison, the error rate for the baseline group on the General Diagnostic Test went from 10.7% at the beginning of the semester to 10.5% at the end of the semester, which was not significantly different \((p < 0.05, t = 0.583, df = 49, \text{sig} = 0.562)\). The effect size at \(d = 0.08\) was very small.

A further test was carried out on a subset of the data, that is, on those who had error rates above the median (14.5%) for the baseline group on the Specific pre-test. Here, it was found those in the treatment group with major difficulties (that is, above the median) made and retained significant improvements, as the error rate dropped from 24.6% before instruction to only 10.2% at the end of the semester, three months after instruction \((p < 0.05, t = 8.192, df = 12, \text{sig} = 0.000, d = 2.27)\). The L1s of this group were as follows: Chinese, 10; Korean, 1; Somali, 1; and Arabic, 1 \((N = 13)\). By comparison, the baseline data showed only a slight and insignificant change \((p < 0.05, t = -0.977, df = 18, \text{sig} = 0.341, d = 0.22)\).
moving from an error rate of 22.3% at the beginning of the semester to 23% at the end. The L1s of this group were as follows: Chinese, 11; Korean, 6; Thai, 1; and Japanese, 1 (N = 19).

![Graph showing error rate comparison](image.png)

**Figure 1:** Comparison of error rate (percentages) of the treatment group with baseline data on the General Diagnostic Test and the Specific Test.

Note that the only data at Time 2 is for the Specific Test on the Treatment Group.

**Discussion**

**THE LISTENING DISCRIMINATION TEST**

Improvement in the Listening Discrimination Test was not very convincing, at least over time. There were some difficulties with this test. The first one was that there were only 21 items, which proved to be too small a number to confidently establish significant changes. This was exacerbated by the fact that only a small number of the 21 questions in this test posed major difficulties. Those that were difficult, such as ‘baked fish’ and ‘asked a question’, were difficult for many learners, who told me informally that they still could not really hear the difference. The listening task was also more challenging due to the transitory nature of listening and, of course, they only got to hear each item once. This makes it difficult to draw conclusions about the effect of instruction on perception, but it may be that it is more difficult to change perception than it is to change production. It may also be that perception is not the only cause of pronunciation errors. In other words, learners may learn to apply some patterns or knowledge about pronunciation to their speaking, if they have some time to plan or rehearse, but still not be able to apply it to their listening.
THE SPEAKING TESTS

In looking at the results for individuals (Table 2) it is notable that they all made remarkable progress. In setting up the study, I had planned to look for individual differences but it was not possible to discern any patterns, as it appeared that all learners had improved. Looking at this progress, then, made it clear that some baseline data would help to show that this improvement was the result of instruction.

Now, in comparing the results for the treatment and baseline groups, there is clear evidence that this instruction was effective and that these gains were retained over time. The most telling indicator is that the improvement was also translated to the General Diagnostic Test, as these items were not taught and involved complete sentences. This suggests that the learning has been integrated into their linguistic competence, at least within a careful style.

The treatment group did have a greater average number of errors at the beginning than the baseline group. This was probably because of the way the sample was chosen. As has already been explained, epenthesis and absence were chosen specifically because the treatment group was found to have major problems in this area. Although the average for the group appears to be above that for the entire population, the standard deviations did overlap with the results for the other group (treatment group, 19.9 +/- 9.3; baseline, 14.1 +/- 7.9). This suggests that both groups are still part of the same population. Since the baseline data was collected a year later, there may well have been some shift in some of the characteristics of the population, therefore changes for those whose error rate was above the median (14.5%) in the Specific pre-test were compared. First, this makes the two groups being compared more similar and, second, it is appropriate because our main concern is with those who have major difficulties with this aspect of pronunciation. These are speakers who drop so many consonants, or add so many schwas, that it seriously affects their intelligibility. The treatment group did have a relatively high number of learners with major difficulties, with 13 out of 21 in that category. The baseline group had only 19 of 50 learners in that category. At the end of the semester this had changed markedly, with only two learners from the treatment group still in the above-median group, while the scores of the 19 learners from the baseline group were unchanged and they were still in the above-median group.

The analysis of this subset provides further evidence for the significance of the improvement made, as it was made by those who most needed to improve. In conclusion, then, one can say that this course of instruction led to significant gains.
Implications for teaching and further research

In returning to the research questions, these results have shown that this particular type of teaching was effective, that a significant proportion of the gains made were retained over time, and that these gains can be transferred to other contexts. Thus, one can infer that effective pronunciation teaching involves:

- making learners aware that there is a difference between what they say and what native speakers say
- helping learners to hear the difference and practise it
- finding the right metalanguage
- helping learners to discover useful patterns and rules
- giving feedback and providing opportunities for further practice.

There were at least two deficiencies in this teaching approach, which were observed during teaching. One was that the explanation of syllables was clearly not understood. The reasons for this would be worthy of further investigation. The second difficulty was that the practice activities were not exactly communicative, except that they did involve genuine cross-cultural communication about pronunciation. The learners’ interest and motivation to know about pronunciation overcame the repetitive nature of the teaching. However, in the future I would like to devise ways to make this more communicative and interesting.

Although the gains were transferred to a different context at sentence level, this still involved a reading task. Ideally, one would like to see that learners have also been able to transfer learning to less-controlled situations when using a vernacular style, that is, in free speech. This study aimed to first show if it is possible to change learners’ pronunciation when working within a careful style, that is, when learners have time to focus on pronunciation as the text is provided. As a follow on, it would be important to establish whether this then leads to changes in the vernacular style.

Having found clear evidence for the effectiveness of pronunciation teaching, there are still questions as to how and why it worked. This requires more research to gauge at what point learners understood phonological concepts and how these were reinforced through practice. There is a need to understand more about the processes learners go through and why they make the mistakes they do. Further, teachers would like to know more specifically how they can include pronunciation in their regular syllabus.
Although this teaching has been shown to be effective, one obvious question that remains is, ‘How important is it?’ The baseline data showed that this is an area of pronunciation that often becomes fossilised. As this study has shown, it can be overcome to a point where it may interfere less with intelligibility, and this is a reasonable goal for pronunciation teaching. In this particular case, I observed that those who had an error rate of less than about 10% on the Specific Test did not suffer from any noticeable impact on their intelligibility. Conversely, those who had an error rate of over 20% were typically difficult to understand. There have been attempts to describe what teachers should focus on in terms of what is important for intelligibility (Fraser 2000). However, there has been little empirical research on the effect of various mistakes on intelligibility. Although there is a need for more research into exactly which aspects of pronunciation are most important in achieving intelligibility (Derwing and Munro 2005), Anderson-Hsieh, Johnson and Koehler (1992) did find a significant correlation between a global assessment of intelligibility and syllable structure errors for East Asian learners, and Zielinski (2004 and this volume) investigates this issue. However, from my experience I suggest that epenthesis and absence impact on intelligibility (see Zielinski, this volume) and learners often do not seem to master it without outside help.

In conclusion, teachers should be encouraged to consider ways of making learners aware of their own pronunciation, and closing the gap between this and their target pronunciation. This will involve a combination of practice in listening, repeating and comparing, finding the right metalanguage to focus learners on salient phonological areas, helping learners to discover useful patterns and rules, giving feedback and providing opportunities for further practice. In terms of practical advice for the teacher, there is still some way to go in getting down to the fine detail. One example, though, is in the use of metalanguage. In the case of absence, it does not seem to help to tell learners they are not saying the final sound, as they often think that they are saying it and we simply do not hear it. Therefore, I suggest it is more useful to get them to use their own terms to describe the changes they should be making, that is, to make the consonant ‘stronger’ or ‘longer’. Similarly, in the case of epenthesis, telling learners not to add an extra schwa is not particularly helpful, as they do not think that they are doing this. They are more likely to understand the issue in terms of making the final consonant ‘shorter’ or ‘lighter’. In the end, as much as teachers like practical tips they can use immediately in the classroom, the way forward is to distil a set of generalisable principles. This study is one step towards gathering empirical evidence for the development of a framework to assist teachers more specifically in the integration of pronunciation into the syllabus.
AUTHOR’S NOTE
This study was made possible through an Auckland University of Technology (AUT) Arts Faculty research grant. Further support was provided by the Centre for Communications Research, AUT. I would like to thank my colleague, Annette Sachtleben, who generously gave up her time to second mark a selection of tests, and Stuart Young for all his guidance and support with the various statistical analyses involved. I would also like to thank Dr Helen Fraser, University of New England, for her suggestions and encouragement. Finally, I would like to thank the students at AUT who agreed to participate.

APPENDIXES

APPENDIX A: General Diagnostic Test

HIGH INTERMEDIATE STUDY PRONUNCIATION: DIAGNOSTIC TEST 1.
BEGINNING OF SEMESTER.

Name: ______________________________________

Part 1: Record these sentences on the tape.
It’s good to eat oranges and lemons in winter.
He lit lots of cigarettes because he had to wait for a long time.
Victoria walked all the way even though she wasn’t feeling very well.
He speaks Spanish with a strong accent.
We’ve got hot and cold baked potatoes for breakfast.
Heat the beans and put them on a plate.
I’ve been waiting since five to seven and now it’s ten past twelve.
Many kind men and women live on the slopes of Mount Eden.
Haven’t you ever seen girls and boys cheering and shouting for joy?
He loves sailing through storms.
The lawns in Green Lane are great.
She threw her perfect birthday jumper in the birdbath.

Part 2: Introduce yourself. You have five minutes to prepare. Say your name and where you are from. Then you could tell me something about what you did before you came to New Zealand, your plans for the future, your family or anything else you want to talk about. One or two minutes will be long enough. You can make a few notes first but don’t write down everything you plan to say.
APPENDIX B: Specific Test

Record these phrases on your tape.

1. He asked questions.
2. They thanked him.
3. She told stories.
4. We discovered perfection.
5. Saved money.
7. He tempted fate.
8. She waited patiently.
9. He loves secrets.
10. She builds houses.
11. It works well.
12. He thinks slowly.
13. She jumps.
14. He races dogs.
15. He wants victory.
16. He finds happiness.
17. Find a job.
19. Told off.
20. Lemons and apples.
21. Here and now.
22. Now and then.
23. Push and shove.
24. Pictures and texts.
25. Wrong count.
26. It’s difficult.
27. A grand plan.
28. Fern fronds.
29. Times tables.
30. Five lives.
31. Perfect timing.
32. Experienced climbers.
33. Intermediate students.
34. Film strip.
35. Headlights.
36. He wasn’t thankful.
37. Silk stockings and tights.
38. Wildlife sanctuary.
39. Tourist attractions.
40. Fine lines.
# APPENDIX C: Listening Discrimination Test

Listen to these sentences and circle the one you hear.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>Catch a tram</td>
<td>11</td>
<td>a</td>
<td>Know your mine</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Catch a tramp</td>
<td></td>
<td>b</td>
<td>Know your mind</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Catch a tramer</td>
<td></td>
<td>c</td>
<td>Know your minder</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>It’s coal</td>
<td>12</td>
<td>a</td>
<td>Give away</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>It’s cold</td>
<td></td>
<td>b</td>
<td>Give a wave</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>It’s colder</td>
<td></td>
<td>c</td>
<td>Give a waiver</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>A grey plan</td>
<td>13</td>
<td>a</td>
<td>bake fish</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>A great plan</td>
<td></td>
<td>b</td>
<td>baked fish</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>greater plan</td>
<td></td>
<td>c</td>
<td>baked a fish</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>Is true</td>
<td>14</td>
<td>a</td>
<td>way up</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>It’s true</td>
<td></td>
<td>b</td>
<td>wake up</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>It’s a true</td>
<td></td>
<td>c</td>
<td>wakes up</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>It true</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>move a house</td>
<td>15</td>
<td>a</td>
<td>Thank Mill.</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>moved house</td>
<td></td>
<td>b</td>
<td>Thanks Mill.</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>moved a house</td>
<td></td>
<td>c</td>
<td>Thanks a mill.</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>silk tie</td>
<td>16</td>
<td>a</td>
<td>It’s a sport.</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>silk tight</td>
<td></td>
<td>b</td>
<td>It’s sport.</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>silk tights</td>
<td></td>
<td>c</td>
<td>It’s support.</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>silk ties</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>a</td>
<td>were</td>
<td>17</td>
<td>a</td>
<td>Trains are late.</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>work</td>
<td></td>
<td>b</td>
<td>Train’s late.</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>worker</td>
<td></td>
<td>c</td>
<td>Train late.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>d</td>
<td>Tray’s late.</td>
</tr>
<tr>
<td>8</td>
<td>a</td>
<td>fie</td>
<td>18</td>
<td>a</td>
<td>ask question</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>five</td>
<td></td>
<td>b</td>
<td>asked question</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>fiver</td>
<td></td>
<td>c</td>
<td>asked a question</td>
</tr>
<tr>
<td>9</td>
<td>a</td>
<td>fine</td>
<td>19</td>
<td>a</td>
<td>He has bees</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>find</td>
<td></td>
<td>b</td>
<td>He has beans</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>finder</td>
<td></td>
<td>c</td>
<td>He has been</td>
</tr>
<tr>
<td>10</td>
<td>a</td>
<td>I like law</td>
<td>20</td>
<td>a</td>
<td>tracks of wolves</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>I like lawn</td>
<td></td>
<td>b</td>
<td>tracks wolves</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>I like laws</td>
<td></td>
<td>c</td>
<td>track wolves</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>I like lawns</td>
<td>21</td>
<td>a</td>
<td>Thai table</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Prospect Vol. 21, No. 1 April 2006 63*
APPENDIX D: Teaching Schedule

Week 4
Tuesday: pre-test.
Thursday: Session 1; 40 mins – return of tests and explanation. Listening to peers and guided help in perceiving differences 1;
Friday: Session 2; 40 minutes – past tense endings (Headway) plus practice.

Week 5
Tuesday: Session 3; 20 mins – Explanation of syllables in English and possible syllable structure.
Tuesday: Session 4; 30 mins – Listening to all versions of the original listening test. Students listen, repeat and record themselves. (Language Lab.)
Thursday: Session 5; 20 mins – explanation of pron of 3rd person ‘s’, rule discovery.
Thursday: Session 6; 30mins – work with syllables (consonants and joining), explanation and practice.
Thursday: Session 7; 50 mins – Listening to peers and guided help in perceiving differences 2; incl. Lang. lab; feedback and discussion.
Friday: Session 8; 30 mins – plural and 3rd person ‘s’ endings (Headway).
Friday: Session 9; 30 mins – Listening to peers and guided help in perceiving differences 3; (Pairs; listen and evaluate).

Week 6
Tuesday: Session 10; 30 mins – Listening to peers and guided help in perceiving differences; 4 & 5.
Tuesday: Session 11; 30 mins – Listen, repeat and record in Lang. Lab.
Thursday: Session 12; 20 minutes – Revision; pair practice.
Thursday: Post-test.
REFERENCES


