# Dynamic Range Compressor Ear Training: Improving critical listening skills through software-assisted practice

Denis Martin

Doctor of Philosophy

Schulich School of Music
Department of Music Research
McGill University
Montréal, Québec

July 23, 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy. 

# Abstract

Traditionally, audio enthusiasts looking to enter the field of sound recording have been trained via a master/apprentice relationship. However, in recent decades a large number of institutions have started to offer training in audio engineering at the post-secondary level. Students are expected to develop their skills during two- to four-year programs that would have normally been acquired through five- to ten-year apprenticeships. As part of their training, students must learn to operate and audibly recognize the effects of a variety of audio signal processors. The following work involved creating, testing, and reporting on the efficacy of a training method and its associated tools for an important and often misunderstood processor: the dynamic range compressor (DRC).

The focused development of the critical listening skills required to recognize these audible effects is called technical ear training (TET). Past research has resulted in training methodologies for a variety of common audio attributes: spectral content, spatial attributes, reverberation, distortion, noise, loudness, MP3 artifacts, and dynamic range. Research into the efficacy of these methodologies is limited but suggests participants improve with increasing training. Overall, continuous development and validation of methodologies for DRC TET have received little previous work.

The DRC TET program created in this study was used to train a group of eight audio engineering technology students over a period of one year. Analysis of the training data revealed that the students improved over the training period at a rate similar to those reported by other researchers. A series of pre-/post-training listening tests were also administered to measure improvement on a set of nearly-related listening tasks including a control group and professional group for reference. The trained students successfully transferred the skills learned to a real-world mixing task, while the control group did not. The trained students also outperformed the professional group.

Overall, it was determined that this DRC TET program was successful. Students improved in their understanding of DRC operation and their ability to detect and quantify audible DRC effects. The software and listening test results presented in this thesis should be useful for educators training audio engineers and a variety of other professionals: researchers training listening test participants, acoustic engineers, electrical engineers, musicians, producers, composers, and more.

# Résumé

Traditionnellement, les passionnés d'audio cherchant à entrer dans le domaine de l'enregistrement sonore ont été formés via une relation maître / apprenti. Cependant, au cours des dernières décennies, un grand nombre d'établissements ont commencé à offrir une formation en ingénierie audio au niveau postsecondaire. Les étudiants doivent développer leurs compétences au cours de programmes de deux à quatre ans qui auraient normalement été acquises dans le cadre d'un apprentissage de cinq à dix ans. Pendant leur formation, les étudiants doivent apprendre à utiliser et à reconnaître de manière audible les effets d'une variété de processeurs de signaux audio. La thèse suivante résume un travail de création, de test et de rapportage sur le succès d'une méthode de formation et des outils associés pour un processeur important, mais souvent mal compris : le compresseur de plage dynamique (DRC).

Le développement ciblé des compétences d'écoute critiques requises pour reconnaître et quantifier ces effets audibles est appelé entraînement auditif technique (TET). Plusieurs chercheurs ont créé des méthodologies TET adressant une variété d'attributs audio courant : contenu spectral, attributs spatiaux, réverbération, distorsion, bruit, volume, artefacts MP3 et plage dynamique. Les recherches sur le succès de ces méthodologies sont limitées, mais des expériences impliquant des technologies TET spatiales et spectrales suggèrent que les étudiants améliorent leurs capacités d'écoute critiques en suivant une formation croissante. Dans l'ensemble, peu de travaux antérieurs ont été réalisés avec des compresseurs de plage dynamique.

Le programme TET lié au DRC créé dans le cadre de cette étude a été utilisé pour former un groupe de huit étudiants en technologies de l'ingénierie audio sur une période d'un an. L'analyse des données de formation a révélé que les étudiants s'étaient améliorés au cours de la période de formation à un rythme similaire à ceux rapportés par d'autres chercheurs. Une série de tests d'écoute avant et après la formation ont également été administrés pour mesurer l'amélioration vis-à-vis un ensemble de tâches d'écoute similaires, notamment par un groupe contrôle d'étudiants et un groupe d'ingénieurs professionnels. Les étudiants formés ont transféré avec succès les compétences acquises vers une tâche de mixage similaire, alors que le groupe témoin ne l'a pas fait. Les étudiants formés ont également obtenu des résultats supérieurs à ceux du groupe professionnel lors de ces tests d'écoute.

Globalement, il a été déterminé que ce programme TET lié au DRC était un succès. Les étudiants ont amélioré leur compréhension du fonctionnement du DRC ainsi que leur capacité à

détecter et quantifier les effets audibles du DRC. Le logiciel et les résultats des tests d'écoute présentés dans cette thèse seront utiles aux enseignants qui forment des ingénieurs du son et divers autres professionnels : chercheurs formant des participants aux tests d'écoute, ingénieurs en acoustique, ingénieurs électriciens, musiciens, producteurs, compositeurs, etc.

# Acknowledgments

I would like to express my sincere appreciation to my advisors Prof. George Massenburg and Prof. Richard King. This work would not be possible without their advice, guidance, expertise, motivation, and collaboration. I also extend this thanks to Prof. Wieslaw Woszczyk, Prof. Martha de Francisco, the instructors of the Sound Recording Area, and Ieronim Catanescu for their continued guidance through my long Qualifying Year, Masters, and PhD studies at McGill University.

I would also like to thank my PhD colleagues Dave Benson, Jack Kelly, Diego Quiroz, Will Howie, Bryan Martin, and Matt Boerum who all lent several helping hands at various stages in my research. The monthly "PhD Colloquium" served as a valuable opportunity to bounce ideas off of an experienced group in a relaxed atmosphere.

A big round of thanks goes to the 24 listening test participants who all volunteered between 6-20 hours of their valuable and limited time to this project. Their attention and motivation are greatly appreciated. Included in this group are also the MMus in Sound Recording classes from 2013-2019, Étienne Boivin and the employees of Studio Lamajeure for participating in and providing feedback on various early implementations of the training program. Special thanks to Jason Corey and Dave Benson for the source of the compressor DSP code used in this software.

Finally, I'm very proud to thank my loving wife, parents, siblings, and extended family for providing me with this opportunity and for their everlasting support and encouragement.

This research was funded in part by the Social Sciences and Humanities Research Council of Canada (SSHRC), the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), and the Schulich School of Music.

# Contributions of Authors

This thesis project is primarily the work of the first author, Denis Martin. Co-supervisors Prof. George Massenburg and Prof. Richard King provided guidance and advice throughout. The supervisors collaborated on the editing and formatting of the document as well as several specific instances outlined below.

**Chapters 3 & 4**

Through the "Advanced Technical Ear Training" course at McGill University co-taught by Prof. Massenburg and Denis Martin, Prof. Massenburg collaborated on the development of the dynamics technical ear training methodology, preparation of stimuli, DSP programming of the dynamic range compressor, and collection of user training data.

**Chapters 5-9**

Both Prof. Massenburg and Prof. King collaborated on the development of the three listening test methodologies (in-context, near-transfer, and far-transfer), Think Aloud experiment methodology, survey questions, preparation of stimuli, and DSP programming of the dynamic range compressor.

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

The field of audio engineering technology (AET) spans a wide range of industries and job titles. Audio engineers are employed in industries encompassing music, audio and video broadcast, film, video games, education, acoustics, research, and audio products. Audio engineers hold titles such as mixing engineer, assistant engineer, live sound engineer, boom operator, sound effects editing engineer, music supervisor, audio designer, audio programmer, course instructor, archivist, consultant, audio/video technician, and research engineer. The focus of the following work is on AET education as it applies to music production. However, the tools and findings can be applied to most of the industries and jobs listed above.

## 1.1 Education in Music Production

### 1.1.1 From Apprenticeship to Post-Secondary

Sound recording technology first took the stage in the late 1800s. At this time, audio engineers were engineers in the true sense of the word: individuals whose presence was required to operate highly technical and uncommon technology. This relationship held true until the 1950s and 1960s when the operation of recording technology became less of a technical process. Audio engineers and producers were then liberated to concentrate on the artistic side of recording [65]. Beginning audio engineers were typically employed as studio assistants in the role of apprentice to a master engineer. Education followed the typical master-apprentice model. However, a shift to offering AET education at post-secondary institutions was already beginning.

According to Roy Pritts, post-secondary education in music production was well rooted by 1960. It began with the Tonmeister program in Detmold, Germany [79]. This model was brought to the United Kingdom in 1966 then four-year degree programs began in the United States in the early 1970s. The Audio Engineering Society (AES) established the "Educational Committee" in 1974 and began listing post-secondary offerings on the "Directory of Educational Programs" in 1979. At that time, 101 programs were listed world-wide. This grew to 211 programs in 1996 and has exploded to over 550 programs today (2019) [7]. This shift was also described by Laurel Cash-Jones; "For many years, audio education was considered an oxymoron, and the only way to enter the recording industry was to grab a broom and start over there…today…all across the country two and four-year university programs and trade schools are helping young people get started on the right track to a career in audio." [8]

### 1.1.2 Learning Outcomes

With thousands of students receiving a post-secondary education in AET each year, a standardized set of learning outcomes would be highly beneficial. It is expected most schools have opted to construct their own list of competencies. However, two notable efforts have been made to construct and publish standardized sets of required skills.

Tough used the Delphi method to examine the opinion of 52 successful future oriented experts from a broad range of audio engineering fields [91]. The result was a list of 160 ranked core competencies that graduates of four-year programs in AET should possess. Among other competencies in categories of "general audio", "digital audio", "music business/business", "music", "electronics", and "communications", the following competencies deserve special attention in the context of the following work.

- #6 "Demonstrate a *basic* knowledge of effects including equalization, reverberation, delays, gates, limiters."
- #8 Demonstrate the ability to pay attention to detail.
- #22 Engineer recording projects as an individual.
- #25 "Demonstrate an *intermediate* knowledge of effects including equalization, reverberation, delays, gates, limiters."
- #35 Demonstrate basic knowledge and basic skill manipulating entry-level professional recording equipment.
- #36 Apply and interpret a technical vocabulary as related to audio.
- #44 Assess the quality of recordings using basic critical listening skills.
- #51 Demonstrate an *intermediate* working knowledge of professional mixing techniques.
- #55 Demonstrate an *intermediate* knowledge of the language of recording production.
- #66 Recognize and appraise differences in audio examples using critical listening skills.
- #136 Design and execute a blind (or double-blind) critical listening test.
- #153 Identify music productions from different recording eras.

What is important to take away from this is students need to learn to use a variety of audio processors and they need to develop a critical ear for sound.

The second resource is the *Game Audio Curriculum Guideline (v. 1.0)*, developed by the Education Working Group of the Interactive Audio Special Interest Group [27]. This document lists 102 core competencies required of audio engineers entering the video game industry. The following are selected from the list of competencies.

- C1. Create a demo reel. Create a portfolio of technical projects together with written accompaniments that explain the thought processes, research methods, design philosophy and approach taken.
- F1. Demonstrate appropriate use of automation, plug-ins and digital signal processing for mixing and production.
- F4. Discuss and implement mixing strategies.
- F5. Master assets appropriately for final delivery.
- G2. Identify and implement room and reverberation effects.
- G4. Demonstrate an understanding of human perception principles and their impact on interactive audio systems.
- J11. Identify common real-time digital signal processing uses and understand their resource needs.
- J12. Mix and master sound design projects both for linear digital video and interactive applications.
- L1. Understand the aesthetics and practices of interactive music.
- L3. Understand and be able to critique contemporary game scores.

Once again, audio students need to understand how to use a variety of traditional audio processors. It can also be argued that audio students need to be able to hear the effects of these audio processors to achieve many of these outcomes (portfolio, final mix, a particular sound aesthetic).

### 1.1.3 Critical Listening and Technical Ear Training

Critical listening, in the context of music production, is the practice of identifying production parameters and diagnosing audio problems by ear. Expert listeners with highly developed critical listening skills can identify finer details in the various attributes of sound, make more consistent judgements about what they hear, and communicate these details and judgements in a clear and concise way [2][6][13][20][78][82].

Technical ear training (TET) is a type of perceptual learning that focuses on the development of critical listening skills. TET is currently being used at a variety of post-secondary institutions, industrial facilities, and research centres. For AET students, TET often comes in the form of instructor led classes and individual practice with a software application. TET programs usually involve specific practice modules covering a variety of audio attributes: spectral content, spatial attributes, dynamic range, loudness, distortion, noise, perceptual coding artifacts, etc. A more detailed review of the applications, methodologies, and targeted audio attributes will be presented in Chapter 2.4.

These TET programs are designed to help trainees master three main skills: 1) the ability to detect infinitesimal differences in the sound of various audio processors, where these differences in sound are caused by a change in processor parameters. Trainees work on their ability to detect subtle changes that have been made to a processor's parameters. 2) The ability to associate changes in the sound of various audio processors with specific controllable processor parameters. Trainees compare the sound of different processor settings and work on their ability to detect which parameters are different and by how much. Participants are trained to use a universal lexicon that can be unambiguously understood by other audio professionals 3) The development of a long-term memory for the sound of various audio processor parameter settings. For example, one of the main goals of equalizer TET is to turn trainees into what Brixen calls a "human spectrum analyzer" [6]; trainees should be able to quantify spectrum modifications by ear without seeing or manipulating any of the audio equipment involved. Finally, some possible beneficial byproducts of the training are an increase in auditory attention, the ability to identify and resist listening fatigue, and an increased understanding of audio processor design and function.

The importance of critical listening is outlined in the learning outcomes above and supported by many others including Reba: "I am confident that many audio educators can agree that of the most invaluable skills that students should develop, critical listening skills are somewhere at the top of that list…" Indelicato et al. "… it is also important for an engineer to recognize good quality sound. One needs to have 'ears', as we say in the industry, to critically evaluate sound fields." and Massenburg "Let's start by restating how important listening skills and tools are to our work in professional audio. Without them, we're just guessing." [26] [61][84]

### 1.1.4 Dynamic Range Compressors

Dynamic range controllers are a family of audio processors commonly used in music production. A survey of several introductory AET textbooks confirms this importance [13][25][58][87]. Each book dedicates several pages or an entire chapter to describing the design and use of various dynamic range controllers. This importance is also outlined in the learning outcomes above. These devices are mentioned more generally as part of the "digital signal processing", "effects", "recording equipment", or "plugins" family of processors and specifically referenced as "limiters".

Dynamic range controllers are used to affect the dynamic range of an audio signal. Individual processors fall into the category of devices reducing a signal's dynamic range (compressor or limiter) or devices increasing a signal's dynamic range (expander or gate). In music production, full dynamic range performances are recorded in recording studios with a low noise floor and plenty of headroom for signal amplitude peaks. These same performances are delivered to an audience listening with a significantly higher noise floor on devices with limited headroom in their amplifiers. Dynamic range compression is often applied to fit the dynamic performance in the limited delivery medium. Dynamic range expanders are less frequently used. The following study is focused on the sound, design, and use of the more commonly used dynamic range compressor (DRC).

#### 1.1.4.1 Design and Function

DRCs come in a variety of types and make use of a variety of circuit designs, electric components, and parameters to control their behaviour. All DRCs make use of two signal paths; an audio path, and a control signal path (Figure 1). The audio path is simply a gain device. The amount of positive or negative gain at any given time is controlled by the control signal. The control signal is generated by the gain computer. The behaviour of the gain computer is modified by the device parameters: threshold, ratio, knee, attack time, and release time. In a feedback design, the control signal is derived from the output signal. In a feed-forward design, the control signal is derived from the input signal. Feedback designs were more common in the early days of audio technology. However, feed-forward methods are less prone to artifacts and are used for their predictability.

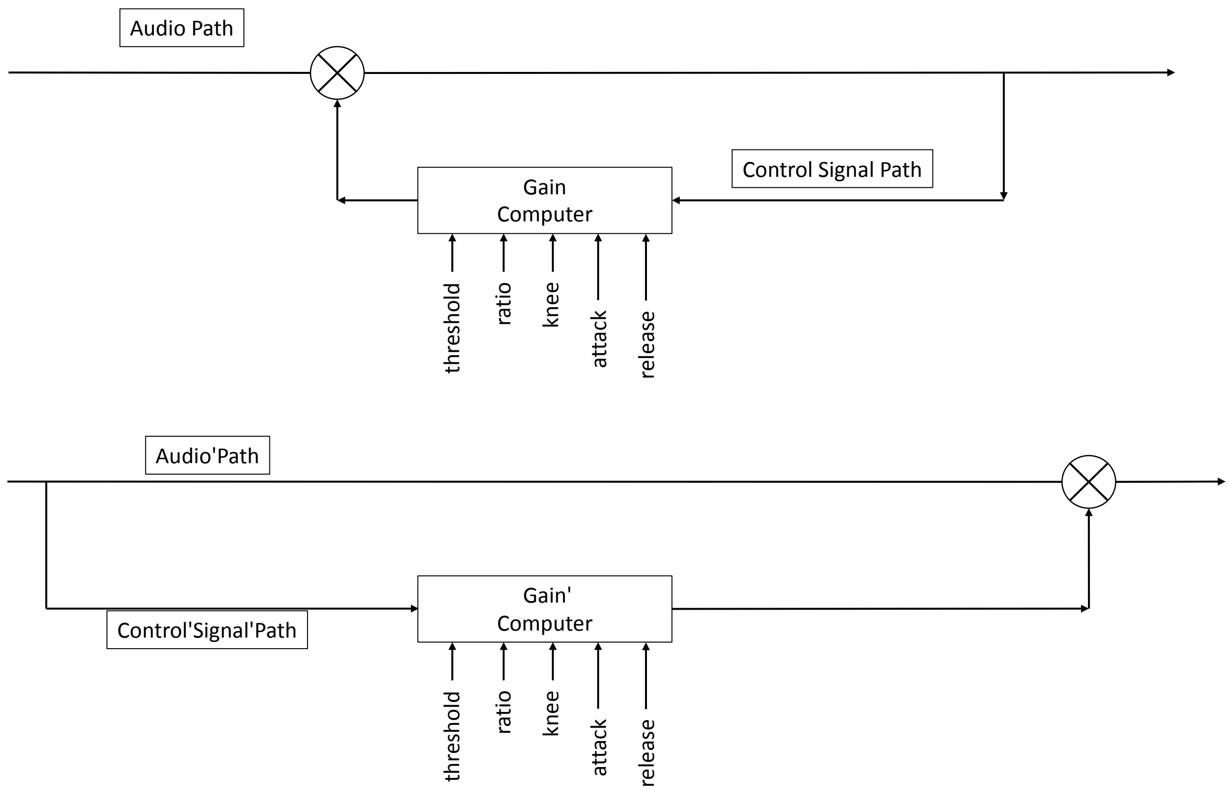


Figure 1: Block diagrams of the signal flow of two typical dynamic range compressor designs: top – feedback, bottom – feed-forward.

The first step in the gain computer is essentially a detector/analyzer. Typical detector/analyzers measure the signal's peak or average amplitude moment to moment. The threshold is defined as a static amplitude. When the signal goes above this amplitude, it will be reduced in level by the gain device. Conversely, when the signal remains below this amplitude, no gain will be applied. For example, with a threshold setting of -20 decibels (dB), when the signal is above this level it will be reduced in amplitude. When the signal's amplitude is below this level no negative gain will be applied.

When the signal passes above the threshold the amount of negative gain applied is calculated from the signal's amplitude and the ratio parameter. The ratio is defined as the slope of the input/output curve above the threshold. The slope ratio below the threshold is 1:1 in a compressor. Above the threshold, the slope ratio can be adjusted anywhere from 1:1 (no reduction) to ∞:1 (maximum reduction). For example, with a threshold setting of -20 dB and a ratio setting of 4:1, a signal that passes the threshold at an amplitude of -16 dB will be reduced to -19 dB. The signal was 4 dB over the threshold and was reduced to be 1 dB over the threshold (divided by 4). The threshold and ratio of a compressor are typically visualized as an input/output graph. Various ratios are visualized on this type of graph in Figure 2**.** The knee parameter can be used to smooth the transition from the 1:1 ratio portion below the threshold to the ratio above the threshold. An example of this is also visualized in Figure 2.

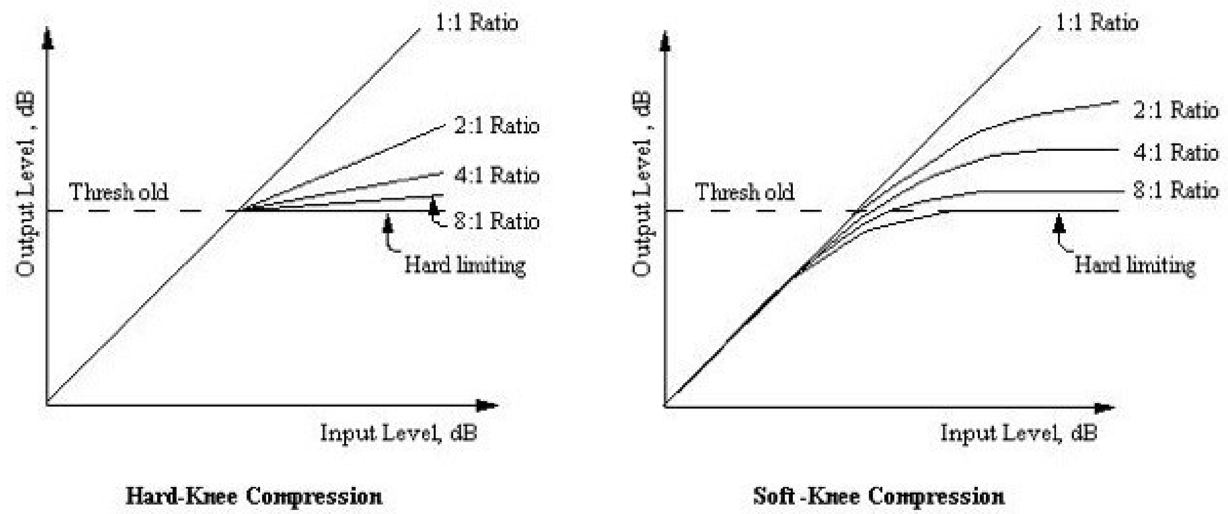


Figure 2: An input/output diagram of a compressor at various ratios and various knee settings.

If the compressor were to apply negative gain immediately as the signal passed above the threshold and return to unity gain immediately after the signal passed below the threshold, the compressor would be acting as a distortion device. The goal of most DRCs is to reduce the dynamic range of the signal without the presence of additional artifacts. The attack and release times are used to rectify this. The attack time slows the rate of gain change whenever the compressor's gain value is decreasing. The release time slows this rate whenever the compressor's gain value is increasing. Overall, the attack and release times serve to slow the compressor's response to a rapidly changing input signal level, smoothing the compressor's change in gain and reducing obvious distortion. Further explanation of the design and application of DRCs can be found in [1][13][25] and [85].

#### 1.1.4.2 Audio Applications

An audio engineer will use a DRC for a variety of reasons in music mixing. One of the most common reasons is to reduce the dynamic range of a signal. This signal could be a single element within a music mix, like a vocal track, or it could be a mix of signals, like the combination of several guitar tracks into a single shared compressor. The compressor can be set-up to reduce the signal's entire dynamic range by a certain factor or to only reduce the loudest portion of a signal's dynamic range. The latter example is typically called peak limiting. The goal is to reduce the level of the signal's loudest transients and to leave the remainder of the dynamic range untouched.

An engineer may want to reduce the dynamic range of a recording so that it can be enjoyed in a variety of listening situations. For example, an orchestral recording can easily occupy upwards of 80 dB of dynamic range [25]. To enjoy that performance the listener must have a noise floor that is sufficiently low that the *pianissimo* passages can be heard and an amplifier with enough headroom that the *fortissimo* passages do not cause distortion. In practice, consumers are listening in scenarios with a peak signal to noise ratio between 12-60 dB [25]. This is not enough to represent the full performance. Dynamic range reduction might be applied to better suit the listener's environment.

An engineer may also want to reduce the dynamic range of a single element to fit it within the context of a full mix. For example, the drums, bass, and guitar might only span a dynamic range of 20 dB in a rock recording. However, the recorded vocal might span a dynamic range of 40 dB. Portions of the vocal performance will be too quiet, and portions will be too loud. A DRC could be employed to bring the vocal into a similar dynamic envelope as the rest of the ensemble.

A DRC is also capable of modifying a signal's timbre by adjusting the gain at a very fast rate. Timbre is considered a multi-faceted attribute. One contributor to the perception of timbre is the attack-decay-sustain-release (ADSR) envelope. When a compressor's attack and release times are set in the millisecond or microsecond range it will modify this envelope. It typically reduces the attack and increases the sustain. Audio engineers use compressors to purposefully modify the timbre of a signal in this way. In this case, the actual dynamic range reduction may be considered a desired or undesired side-effect.

Another artifact that can occur when using a compressor is “pumping”. This occurs when compression is applied to a signal that contains both steady state and transient components. When transients (signal peaks) occur, the compressor is typically applying negative gain. However, the lower-level steady state components in the background will also be reduced. When the transient has passed the compressor will return to unity gain and the steady state components in the background will increase in level. A continued stream of transients will cause background steady state elements to periodically increase and decrease in level. This produces an unrealistic “pumping” effect. Audio engineers sometimes use this unrealistic artifact for artistic uses. It is described as adding “motion”, “energy”, or “movement” to the music.

#### 1.1.4.3 Learning to Operate Dynamic Range Compressors

Through several years of experience in audio instruction, DRCs have been identified as one of the more difficult audio processors for students to learn. The same level of difficulty does not occur with the volume fader, pan pot, equalizer, delay, and reverberator. These devices are primarily linear and time-invariant; their sound is not dependent on the input signal and their sound does not change over time. A DRC is the opposite: it is non-linear and time-variant. A DRCs dependence on the input signal, constantly changing response, and high number of inter-related controls makes it very challenging for students. Best practices are also hard to teach since generalizations are problematic to make. Finally, the visual feedback that is available on most devices can be misleading and uninformative.

If students can understand the complex operation of a DRC they are met with another challenge: learning to hear the resulting effect on the signal. Extreme uses are obvious but subtle ones are not. Students must learn to listen over a period of time since a DRC’s sound is constantly changing. It is hypothesized that this puts a large strain on the student’s auditory memory when making comparisons. In addition, the parameters on a compressor have a mix of logarithmic and linear effects on the resulting sound. It is hypothesized that changes in the ratio, attack, and release parameters are perceived logarithmically while the threshold and knee parameters are perceived linearly.

Overall, compressors deserve a special place in the roster of available audio effects. They are important, complex, non-linear, and time-variant devices that require a solid understanding to use and a strong auditory memory to hear.

## 1.2 Scope of Work and Global Hypothesis

This research project involved the design of a TET methodology, software application, and stimuli set to address DRCs. The software application was used to train a group of eight AET students. A series of listening tests were also administered to measure improvement in the students' critical listening skills. It was hypothesized that the students would improve in their understanding of compressor operation, ability to listen to the sound of compressors, understanding of compressor best practices, confidence, and speed. The results could provide instructors with proven methodologies and tools to teach their TET classes and have a positive impact on all AET education areas.

# 2 Literature Review

## 2.1 Auditory Memory

It was theorized that auditory memory can be broken up into two categories: long-term and short-term [15]. Long-term memory is primarily used for identifying and classifying sounds, for example, the relative pitch of a note. The listener must have had a range of previous experiences exemplifying a full range of pitches so they would know how to weigh this new incoming pitch against the average. In audio engineering, it is important to develop a long-term memory for the full variety of possible processor sounds. An engineer can then classify and categorize what they are hearing and identify processes by ear. Long term auditory memory has a high capacity but is not very detailed. In TET, Letowski suggests dividing the continuous spectrum of possible intensities into more manageable discrete "chunks" [46]. The purpose is to aid in developing and maintaining this long-term memory.

It was also theorized that short-term auditory memory can be broken up into two sub-categories: a very short-lived but high-capacity "echoic memory" and a more enduring but limited-capacity "short term" memory [15]. A similar phenomenon is reported in visual perception. When comparing audio effects back to back, it is likely that linear and time-invariant effects (ex: equalizers) can be evaluated using the high-capacity echoic memory immediately at the moment the engineer "switches" between the two stimuli. However, this is not always possible with non-linear and time-invariant effects (ex: DRCs). Listeners can evaluate some of the audible changes quickly, like modifications of timbre, distortion, or other frequently occurring artifacts. For some other audible changes, like modifications in long-term dynamic range, the listener must evaluate the sound over at least a short period of time. Therefore, the more limited capacity "short term" memory is likely being used in this case. This difference could be the reason why AET students report difficulty in learning to hear some effects of DRCs. It is expected that a listener would become more discriminating when evaluating sounds if their short-term memory could be improved.

## 2.2 Isomorphic Mapping

In TET, the term isomorphic mapping comes from Corey and Benson [13]. Most audio processors are controlled by a set of parameters that have no obvious relation to the perception of the resulting sound. For example, if someone wanted to make the timbre of a signal “brighter” with an equalizer they might employ a *6 dB peaking filter* with a *frequency of 10 kHz* and a *Q factor (bandwidth) of 1*. Even for someone who is familiar with the tone controls on their stereo amplifier and the concept of brightness in timbre would have no idea what to expect by inputting these parameters. Isomorphic mapping is the association of these technical parameters to the perceived sound. Students can practice this skill through systematic repetition [80][81][82][83]. Therefore, it is expected audio students would increase their speed and confidence in using various audio processors by improving their isomorphic mapping skills.

## 2.3 Auditory Sensitivity

A review of Watson’s *Time Course of Auditory Perceptual Learning* reveals that improvements in auditory sensitivity were achieved through systematic training [92]. Specifically, improvements were made in auditory thresholds and just-noticeable-differences (JNDs) of a variety of audio attributes. The learning curve took a logarithmic shape. The time needed to achieve asymptotic performance varied depending on the listening task, complexity of the stimuli, number of stimuli, and familiarity. Improvement was demonstrated in the detection of a sine tone in noise and discrimination of changes in loudness, pitch, and duration. Asymptotic performance was achieved faster (a few hours) in tasks that involved simple detection and took longer (hundreds of hours) in more complex tasks like discrimination and identification. It is expected that audio engineers would become more discriminating and consistent if they could improve their auditory sensitivity; they could detect and identify small changes in audio processors that might otherwise go unnoticed.

## 2.4 Technical Ear Training

TET for audio students is a relatively new field. The earliest documentation of a program comes from Letowski and Miskiewicz at the Chopin Academy of Music in Warsaw, Poland [46][63]. The authors reported that the TET program began in 1977. Instructor-directed training methodologies involved presenting repeated listening experiences to students in a controlled classroom environment. The training focused on many different audio attributes including spectral content, dynamic range, distortion, noise, and pitch. As mentioned above, the goal of TET is to teach students how to hear the effects of various audio processors. More specifically, it can be divided up into three primary aspects: 1) auditory sensitivity, 2) isomorphic mapping, and 3) long and short-term auditory memory.

At the Chopin Academy of Music instructors began by presenting audio processes in an exaggerated form so that their effects were clearly audible and could be easily memorized. For example, what will be referred to as an "absolute ID" task. Students heard a stimulus, then they heard the same stimulus with the audio processes applied. Afterward, they were asked to determine the processor parameters by ear. A second task was also used that will be referred to as "matching". In this task, students had access to their own identical audio processor and attempted to match the sound of their processor to the provided example. These tasks will be described more thoroughly in Chapter 3.1. The audio processes became more subtle as the training progressed and the students' sensitivity increased (*aspect #1*). The students also formed long-term memory references for the varying intensity levels of the processes with repeated exposure to the various manipulations (*aspect #3*). A point was made to describe sound in technical terms during the training. For example, which equipment parameter settings gave rise to the sound. By using this technical language, the students were better able to predict the sound of their equipment, mapping their perception to equipment parameter values (*aspect #2*). This methodology required a classroom environment, a stimulus set, high quality playback equipment, and an instructor.

Several CD and tape-based training programs were developed by Everest, Moulton, and Brixen in order to alleviate the need for an active instructor and classroom listening environment [6][19][20][21][66]. This brought similar listening experiences to a wider audience. One disadvantage of these programs was they were not interactive. User responses were recorded on a sheet of paper and checked against an answer key at the end of the training session.

### 2.4.1 TET Software Applications

Beginning in 1990, Quesnel implemented an equalizer TET program in software for use at McGill University [80]. Implementation in software provided several advantages: training without an active instructor, unlimited number of randomized training questions, instantaneous feedback for the user, advanced response tracking, and adaptation to each user's skill level. Iwamiya et al. also described how a TET application can be implemented using personal digital assistant (PDA) terminals as response devices in a classroom setting [31]. Estep & Meng described a supplemental training task called "mnemonic imagery". After completing a "matching" or "absolute ID" trial the student immediately re-imagines completing the task in as high detail as possible [18]. Moving forward to today, most TET applications make use of repetitive "matching", "absolute ID", "ranking ordering", or "method of adjustment" tasks implemented in software applications.

TET with equalizers has received the most attention overall. However, several other researchers and educators have applied these concepts to other audio processors [6][13][19][20][62][66][77][82]. Training programs have been created to target feedback frequencies, gain devices, reverberation, distortion, audio edit points, and MP3 artifacts [13][18][31][37][52][73][74]. Neher et al. created a TET program to target spatial audio attributes "ensemble width" and "ensemble depth" [67]. One of the main difficulties they encountered was these audio attributes are not directly controlled by any common audio processor. They document the development of an algorithm to unidimensionally scale these attributes and validate their success. So far, Corey & Benson are the only authors to report on a software application and accompanying documentation that targets DRCs [13].

Equalizer TET has received enough attention for it to become effective and commonplace in AET programs. The same scale of development and acceptance is not apparent with other audio processors and attributes. In the context of training students, it is recommended that TET programs should be focused on audio processors, as opposed to audio attributes. These are the tools an engineer has at their disposal to make manipulations to the sound. Audio attributes may be a more appropriate focus when it comes to different applications, like training listening test participants.

Agreement in essential key features becomes clear when reviewing the current TET offerings [14][21][46][55][62][63][66][67][76][80]. Implementation in software provides the advantages summarized above. A diversified stimulus set retains user interest and provides a wide range of experiences. The training system should have gradually increasing difficulty levels. Continuous audio processor parameters should be discretized into "chunks" for memorization. Training in software should be accompanied by instructor directed listening sessions. Letowski discusses the importance of students learning from each other through class discussion and instructor demonstration [46].

### 2.4.2 Proof of Concept and the Learning Curve

Several researchers have attempted to quantify the learning that takes place during TET. Some key questions to answer are: do participants improve from the beginning to the end of the training program? What is the rate of learning? What is the shape of the learning curve? Can participants transfer these skills to everyday audio engineering tasks?

The first researcher to examine this was Quesnel in 1990 [80]. He used an equalizer TET program to train nine students over six months. Students completed 27.4 hours of training on average and improved from an average score of 80 % to 99 % on a beginner training level. The students' response time also improved from an average of 0.87 minutes to 0.44 minutes. On a more difficult task, Quesnel found that the students' score improved from an average of 54 % to 66 %. Quesnel measured this learning using a pre-/post-testing methodology. Students were given the same test before and after the training. Unfortunately, it is not reported whether these improvements are statistically significant.

Bech also reported learning when participants repeatedly took the same loudspeaker rating listening test [1]. 12 university students' performance was measured using the "F-statistic" on each listening test. The total number of hours spent training was 2.5 and it took place over five days. Bech noted that the learning curve was positive. Each participant experienced diminishing returns, approaching an asymptotic performance level. Bech did not provide a statistical model for this learning curve. Instead, he reported that 65 % of participants reached an asymptotic performance level after four repetitions of the listening test.

In 1994, Quesnel & Woszczyk reported the results from another set of students using their equalizer TET program [83]. Seven students were trained over eight months. Students each completed an average of 47 hours of training. Again, a pre-/post-test methodology showed that the students improved from 70 % to 97 % success and a trial time reduction from 8.1 s to 3.1 s. Quesnel reported that each student's time spent training was significantly correlated with the degree of improvement in their test score, but not with their improvement in response time. It is not known exactly which factors contributed to the students' improvement since neither of Quesnel's studies used a control group.

Also in 1994, Olive developed an equalizer TET program and used it to train seven employees at Harman International [76]. Employees completed five training sessions over a period of several days. The average training time is estimated to be 3.75 hours. Participants significantly improved in that time. They achieved an average score of 67 % on the first session and 80 % on the final session. Olive did not comment on the shape of the learning curve. Interpretation of the presented data revealed a linear trend.

Jumping to 2001, Quesnel completed his PhD Thesis, which was later published as a book [82]. This was a final study on his equalizer TET program. A group of five students trained for an unreported number of hours over two semesters. Quesnel compared the students' performance after training to a group of professional audio engineers. The goal was to determine if the skills learned during training were comparable to the skills an audio engineer learns over a career of experience. Quesnel found that the students (average 96 %) significantly out-performed the professionals (average 78 %) in a listening test resembling the training program. The same relationship was found when comparing the groups along Quesnel's "performance index" which takes the students' score and divides it by the trial time. Quesnel found no significant difference between the two groups in their response time. Quesnel noted that the students are at an advantage by being already familiar with the listening test through their lengthy training. Part of the measured differences might be due to familiarization with the listening task and interface and not necessarily a difference in general critical listening skill.

In 2003, Olive reported a comparison between a trained panel of listening test participants and a variety of other listener groups [78]. Olive showed that the trained group could transfer the skills learned in the TET program to a loudspeaker rating task. The trained panel performed "3 to 27 times better" than groups of audio retailers, audio marketers, audio salespeople, and naïve students when measured using the "F-statistic". Unfortunately, it is not known if this difference is statistically significant. However, the study makes use of a huge sample including 268 participants.

Neher et al. conducted a small pilot study by training only three naïve students on two spatial audio attributes: "source distance" and "source width" [71]. An average of 4.3 hours of training was completed over 2.8 weeks. They used a pre-/post-testing methodology and included a control group of two students. The experimenters eliminated the factor of learning due to listening test familiarization by examining the differences between the two groups. The results of the study are mostly unreliable due to the very small sample size. However, they showed that the learning experienced by the control group was on the same scale as the learning experienced by the trained group.

Kassier et al. continued the work of Neher el al. in a series of studies from 2005 to 2007 [33][34][35][36]. The authors were interested in quantifying the transfer of learned skills to a nearly related listening task. Students were trained on two spatial audio attributes: "ensemble width" and "ensemble depth". Training involved in-class demonstrations and individualized training on a software application. 48 students were divided into three groups: a trained group that went through the TET program, a practice group that simply practiced the pre-/post-test repeatedly, and a control group that did not practice or train. The practiced and trained groups trained for three hours, in 30-minute sessions, over one week.

The trained group's performance was measured during the training. Students improved from a "mean absolute ranking difference" score of 1.0 to 0.4 and a response time of 124 s to 95 s on the ensemble depth attribute. They improved from a score of 1.2 to 0.75 and a response time of 127 s to 110 s on the ensemble width attribute. These improvements were statistically significant and the learning curve was linear. The authors mentioned that a logarithmic curve would likely be found with a longer training period.

On a pre-/post-test, improvement was measured in all groups for both performance metrics (rank order and response time). The degree of improvement was statistically higher for the trained and practiced groups over the control group, but the degree of improvement between the trained and practiced groups was similar. This result confirms that some learning occurred due to familiarization with the pre-/post-tests. However, a higher level of improvement occurred when students were trained or practiced.

Another pre-/post-test was conducted designed to measure performance on an unfamiliar but nearly related listening task. The exact same results were reported. Skills learned during training or practicing were transferred to the nearly related listening tasks. There was not a difference between the trained and practiced groups.

Evidence of effective TET programs continues with Liu et al. who reported improvement between pre-/post-test scores in a training program targeting equalizers, loudness, instrument discrimination, and pure tone discrimination [52][53]. Specifically, 57 students improved from a score of 70 % to 95 % with 3.75 hours of training on a pure tone identification task.

Kawahara et al. showed that TET programs were also successful when training adult employees [37]. An unreported number of junior employees at Yamaha Corporation trained around nine hours over four weeks. The learning curve was logarithmic (approaching an asymptotic level).

Participants improved from a score of 80.7 % to 89.7 % on a loudness task and 76.2 % to 86.2 % on an equalizer task.

Nishimura also demonstrated that TET experience could help students understand theoretical concepts [74]. In this case, TET experience with perceptual codecs improved student written exam scores on the same topic by 5-10 %.

Going back to the beginning, Rosciszewska & Miskiewicz from the Chopin Academy of Music reported testing results gathered from 136 students [86]. Students were measured using a pre/post standardized test over 15 years of teaching. Students typically received four years of TET courses throughout their degrees. The pre-/post-test was based on an equalizer TET task and 88 % of students improved. The variance of scores were much less post training which suggested a logarithmic learning curve. The mean improvement in scores appears to be from 58 % to 75 %. On two other tasks, 74 % and 90 % of students improved respectively. The results of this study are very interesting. It involves a massive number of students training for an extended period in a high-level TET program. Unfortunately, the report lacks many details and is generally void of statistical hypothesis testing.

Returning to the main questions outlined at the beginning of this chapter, it is easy to conclude that, yes, 10 separate studies showed that participants improved in their listening ability when participating in a TET program. It is more difficult to come to a conclusion when examining the rate of learning. An average of all studies using percentage as a measurement of performance leads to a learning rate of 2.0/100 points per one hour spent training. However, there is evidence that the learning curve is not linear. Therefore, it cannot be reported as a single rate. Studies using short training periods (around three hours) reported an average learning rate of 3.7 points per hour and studies long training periods (25+ hours) reported a rate of 0.34 points per hour. This suggests a logarithmic learning curve, approaching an asymptotic level of ability with diminishing returns. The curves can probably be approximated as linear for low numbers of hours spent training. However, the logarithmic component must be considered as the training hours increase. It is interesting to note that some studies also reported a decrease in response time during training. Kassier et al. demonstrated that skills learned during TET were transferred to a very closely related listening task, but the question remains whether the skills can be transferred to a more distantly related task. This question is often rephrased as “will TET actually make a student a better audio engineer?”.

A pattern of issues was identified in some of the studies summarized above. Control groups should be included in all pre-/post-testing methodologies. Measurement devices (for example a test score) should be properly calibrated. In some studies, the measurement device was "capped out" by the end of training; all participants were achieving 100 % which is an end point on the measurement device. Also, all hypothesises should be statistically tested and a learning curve should be fit to any training data. Finally, a larger number of participants should be used to increase sample sizes when logistics and class sizes permit.

### 2.4.3 Factors in TET Program Design

Researchers have created TET programs and researchers have validated TET programs. More recently, Researchers have also considered how to optimize TET programs. In 2006, Kassier et al. wrote a summary article outlining factors to be considered in training for optimal transfer of skills [36]. The question was, how can we design TET programs so that the skills learned can be applied to other related circumstances? Their summary of previous research showed that:

- The training task should be similar to the transfer task.
- There should be variety in the training task.
- Intelligence and motivation are large factors.
- Training on an easier task may be more beneficial than a hard task.
- Training time is important, considering a warm-up period and eventual fatigue.
- The training environment and task should be realistic and ecologically valid.

They also found that when testing the transfer of learned skills, the transfer task should be geared toward application of learned skills and not simple recall.

McKinnon-Bassett and Martens published a study comparing two approaches to equalizer TET [62]. They found when students trained by successively "guessing" the settings of the equalizer, as opposed to continuously "sweeping" for the settings, they better transferred their learned skills to a nearly related listening task. This implies that the method of dividing up continuous processor parameters in discretized "chunks" for memorization is effective. This puts an emphasis on long-term memorization instead of relying on short-term comparisons.

Kim et al. (also reported in Kaniwa et al.) found that "individually optimized training sequences" were superior to randomized sequences in the context of equalizer TET [32][38]. The finding assumed that students have more difficulty with some combinations of processor parameters than others. It also assumed that students differ in the parameter combinations that they find difficult. The individually optimized training sequence prioritized training on processor settings that the individual had difficulty with. The software application favoured questions that each individual student typically got wrong. As measured, the difference between the two methods was not entirely conclusive. However, there was some evidence that for one stimulus the students using the optimized sequence significantly improved 15-20/100 points more than the students that used simple randomization.

Kim and Olive also found that working on a reference quality set of headphones was significantly superior to working on a low-level consumer set [40]. Students using the reference headphones improved 10/100 points over 15 weeks and students using their own consumer set only improved 2/100 points over this time. However, the performance of the control group (consumer headphones) was lower in the pre-test than the treatment group (reference headphones). Assuming a logarithmic learning curve, the difference measured could have been due to the difference in the curve's slope as opposed to the difference in headphones.

Kim and Imamura found that a five-channel surround sound speaker system was superior to a binaural version over headphones in the context of two-dimensional spatial TET [39]. This is not surprising, given the current limitations of non-individualized binaural audio. For those without access to a five-channel surround sound system, a lesser degree of improvement was still measured when training the rating of "auditory source width" over headphones.

Finally, Marui and Kamekawa produced a series studies on the varying difficulty between stimuli in an equalizer TET program [59][60]. It was documented in many of the studies above that stimuli choice can have a large effect on the difficulty of the training task. However, it is not known what signal features contribute to the difficulty. The authors hypothesized that it was related to the spectral flatness of the signal. In other words, that an equal amount of energy in all octave bands would make the task easier. The authors did not find a difference in difficulty when examining the stimuli across three different measurements of spectral flatness.

Overall, there are countless factors that influence the efficacy of a TET program. Current research has only "scratched the surface" of what will need to be examined in the future.

### 2.4.4 Gap in Knowledge

The largest gap identified in previous research was the continued development of DRC TET programs. Faders and pan pots are likely the most important signal processors in audio production. However, their operation is simple and their effects are easily identified. The next most important processor is likely the equalizer. TET in this realm has received significant amounts of research both in development and validation. DRCs were identified as the next most important processor family to address.

There were two main areas that were identified for improvement in DRC TET programs. The first, was previous programs trained audio processors in the "mastering" context while ignoring the "mixing" context. To clarify, processors were applied to the whole signal rather than individual elements within the full mix. That setup reassembles a typical workflow in mastering. However, with many more mixing engineers in the world than mastering engineers, an effort to present the processor in a mixing context was made. DRCs should also be applied to a single track within a mix. For example, a DRC could be placed on a lead vocal within the context of a backing track. The second area for improvement was related to the generation of randomized parameters. Processor parameters to be matched are randomly generated in most TET applications. This works well when the audio process is linear and time-invariant. However, it is common to randomly generate a nonsense set of parameters when the processes in non-linear and time-variant. For example, with a DRC, a random threshold that is too high such that the compressor does not ever engage. Solutions to this problem are discussed further in Chapter 3. In addition, this DRC TET program was optimized for transfer of skills according to the recommendations of Kassier et al.

Finally, the efficacy of DRC TET programs had never been tested. This research employed a variety of testing methods (Chapter 2.5) to measure learned skills in-context and on transfer tasks.

## 2.5 Measurements of Learning

It was hypothesized that by taking the DRC TET program students would improve in their:

1. Understanding of compressor operation.
2. Ability to hear the effect of compressors.
3. Understanding of compressor best practices.

4. Confidence and speed in compressor use.

Student learning is difficult to measure. Therefore, researchers in the field of education typically rely on the application of several assessment techniques whose results should be interpreted holistically [5][23][41][93]. Ideally, the methods should span from direct measures to indirect measures. For example, grades and standardized tests as direct measures. These can be combined with surveys and course evaluations as indirect measures. Direct measures are thought to be more objective and indirect measures are thought to be more detailed and insightful. The techniques used can be designed to collect either quantitative or qualitative data. The following study made use of four techniques spanning all of these categories. Some techniques had been used in TET assessment research before, some were a new application.

A. Training data.
B. Pre-/post-testing.
C. Retrospective think-aloud.
D. Exit Survey.

Learning outcomes 1, 2, and 4 were measured by all assessments and learning outcome 3 was primarily measured by assessments C and D.

### 2.5.1 Training Data

Each time a student used the DRC TET software, data was collected on the difficulty level, stimuli, success rate, and other variables. An analysis was conducted to examine changes in the student's success rate over time. Since questions in the training program had an objectively correct answer, this technique constituted a direct measure of a quantitative variable. This technique was the equivalent of examining student practice problems in a mathematics course to examine how their success rate changed with practice. Analysis of training data has been the most commonly used assessment in TET research so far [2][26][34][35][36][37][38][52][53][76][77][78] and is also used in the broader field of educational research [54].

### 2.5.2 Pre-/Post-testing

Students also took a series of three pre-/post-tests before and after the training. The first test presented questions that were identical to those found in the training environment. By contrast, the second test included unfamiliar stimuli, measuring the transfer of skill to a nearly related task. Since the questions in these two tests had objectively correct answers, these techniques constituted a direct measure of quantitative variables. The equivalent technique used in the broader field of educational research is standardized testing [54]. Pre-/post-tests have been very popular in TET assessment [34][35][36][39][62][71][86]. A student control group and a group of professional engineers also took these tests. The control group was used to isolate some of the factors that may have been contributing to learning. Similar to Quesnel's research, the trained students were compared against professional engineers to evaluate whether the listening skills developed through TET were comparable to those developed by professionals over years of work [82].

The third pre-/post-test was designed to measure changes in skill in applying DRCs. Since it was impossible to directly measure this ability, participant consistency in applying DRCs was measured instead. This technique was an indirect quantitative measurement. Consistency in adjusting mixing parameters (usually blind, to prevent bias) has been used as a measurement of listener "performance" in audio engineering research several times before. Levine et al. examined the effect of various simulated (binaural) and natural acoustic control room conditions on the consistency of a lead vocal level setting within a mix [50]. King et al. examined the effect of various acoustic control room conditions and playback systems on the consistency of a lead instrument or vocal level setting within a mix and high frequency level preference [42][43][44][45]. Participant consistency, as measured by the $F$-statistic, was used to evaluate listening test participant performance by Bech and Olive [2][78]. The International Telecommunication Union Radiocommunication Sector also suggests that statistics representing the variance in the response data from participants can be used as an "evaluation of listener expertise" [29].

### 2.5.3 Retrospective Think-Aloud

Retrospective think aloud experiments are designed to study human behavior [51]. They are commonly applied in usability applications to gather data on user intents and reasoning during a task [17]. They are a sub-category of the more broadly defined think aloud protocol where qualitative user explanations are collected after the task instead of during. The main goal of these experiments is to examine how and why participants react and perform in certain situations.

The two student groups described in the previous chapter participated in a retrospective think-aloud experiment at the end of the training period. During this experiment, students were asked to adjust the settings of a compressor in a real-world mixing simulation. After their adjustment, they were asked to explain their thought process and methods of adjustment. This explanation was subsequently graded. This experiment involves indirect evaluation of both quantitative and qualitative variables. So far, think-aloud protocols have not been used in TET assessment but have a history of use in the broader field of educational research [9][15].

### 2.5.4 Exit Survey

Finally, the trained students also answered a series of open-ended questions in an exit survey. This experiment involved indirect assessment of qualitative data. An equivalent technique in the broader field of education would be a student course evaluation which is used at many university institutions as an evaluation of teaching effectiveness [24]. The exit survey in this study was designed to collect feedback on the quality of the training program, perception of improvement, and suggestions for the future. No previous research was found where surveys or student course evaluations were used to assess the effectiveness of a TET program or course. It's highly possible that this data exists, but unfortunately reports were not found.

# 3 Training: Software Application and Program Material

## 3.1 Training Methodology

The DRC TET software (named "*dyntet*") used in this study was built on previously successful methodologies. It involved drill-and-practice exercises where tasks were practiced over and over within a specifically constructed training environment. The tasks involved listening to a DRC in both mixing and mastering contexts. To explain, DRC was applied to a single musical element (ex: vocal, guitar, snare drum) within the context of a complete musical work (ex: rock, pop song), or directly to the master buss. Trainees tried to determine an unknown set of DRC parameters strictly by ear. Trainees were forced to analyze what they are hearing within the context of the compressor parameters that gave rise to the resulting sound. The training began at an easy level where the differences between compressor settings were exaggerated. The differences become more subtle as the training progressed and the trainees' discrimination became more refined.

### 3.1.1 Segmentation and Long-term Memory

During the training, the continuously adjustable compressor parameters (threshold, ratio, attack, release, and make-up gain) were divided into a series of discrete steps. As outlined in Chapter 2, it is important to break down continuous parameters into categorical (and sub-dividable) "chunks", so that they can be memorized. This brought up the question of how the various compressor parameters should be discretized. What are the useful end points of each parameter? For instance, what is the fastest attack time we should be interested in, and what is the highest perceptually relevant ratio? Should the parameter scales be treated as linear, exponential, or logarithmic? The chunks were chosen in an attempt to segment the parameter range into a series of perceptually equal steps. By surveying the control ranges and front plate markings of many different commercially available compressors, some information was gained about what equipment designers thought of these questions. However, there was a high amount of variance. For the current implementation of this training program, the end-points and discrete steps were drawn from these sources and professional experience with DRCs. This is an area that could be explored further. Statistical analysis of the training data resulting from the use of this software provided some insight.

### 3.1.2 The Matching Task

Of the two training tasks implemented in the *dyntet* software, the first was a matching task. In this task, the trainee first listened to the actively compressed target musical element with the compression in place but with hidden parameters (labeled *question*). Then they compared it to an uncompressed version (labeled *answer*). The trainee had control over a second compressor that was applied to the unprocessed version (*answer*). The trainee applied the compressor parameters they believed would exactly match the *question* to the *answer*. The trainee could switch freely between the *question* and *answer* and a third uncompressed stream (labeled *bypass*) while matching the sound. They submitted their answer when they were satisfied and received feedback on their work. Trainees had a chance to re-listen to the audio streams before moving on to the next trial. Figure 3 shows a block diagram of the signal flow.

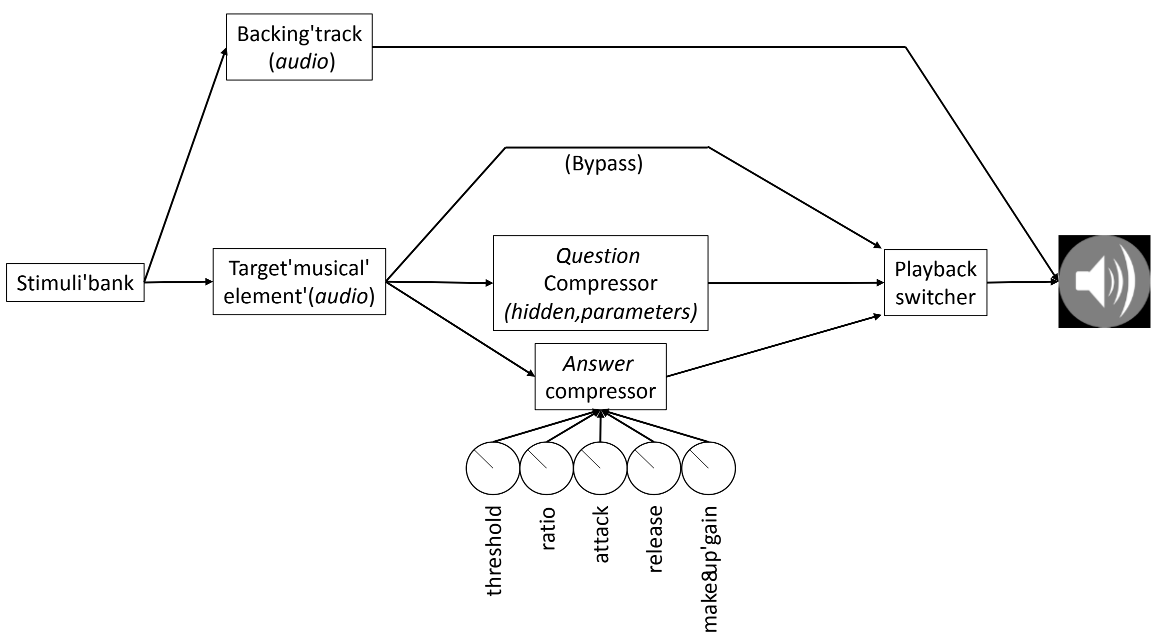


Figure 3: Signal flow for the matching task in the *dyntet* software.

### 3.1.3 The Absolute ID Task

The matching exercise was designed to develop the trainees' isomorphic mapping and discrimination skills. The absolute ID task was designed to develop their long-term memory. In the matching task, trainees could switch between the *bypass*, *question*, and *answer* audio streams freely. The underlying task was to adjust the *answer* compressor parameters until it sounded the same as the *question*. In the absolute ID task, the *answer* audio stream was not available to the trainee for listening. The trainee listened to the *bypass* version of the audio and then the *question* with a set of hidden compressor parameters applied. After listening, they guessed what those parameters were. The trainee was not able to hear what their *answer* sounded like and compare it to the *question*. They had to use their long-term auditory memory instead of making comparisons with their short-term memory. With repeated practice, the trainee was forced to memorize the generalized sound of the various discrete parameter steps. Figure 4 shows a block diagram of the signal flow for the absolute ID task.

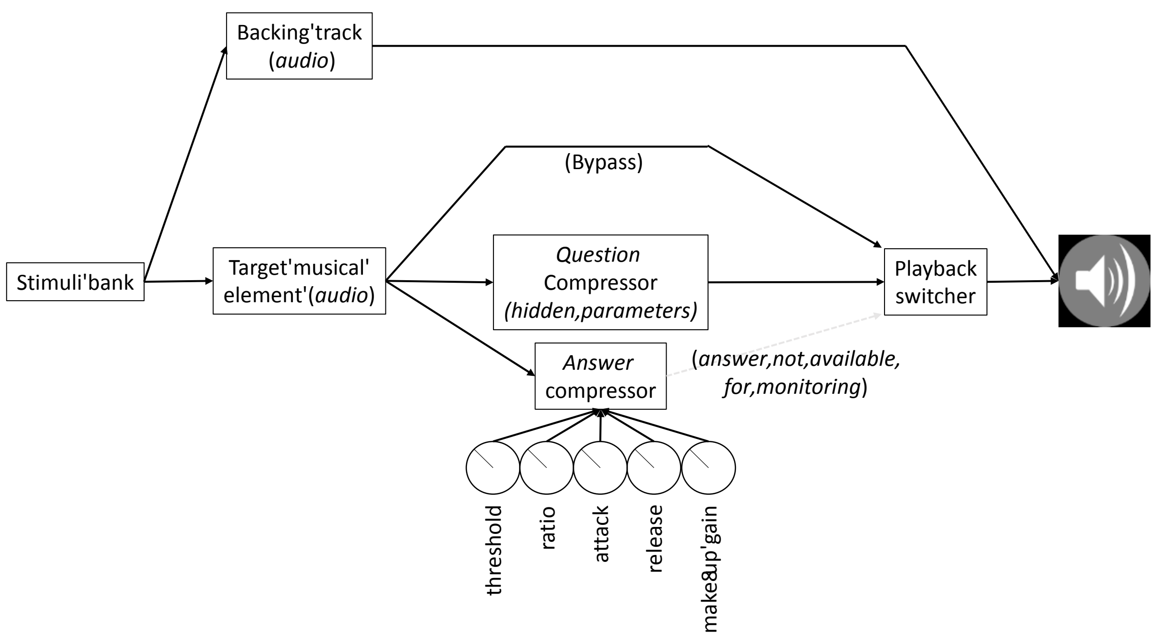


Figure 4: Signal flow for the absolute ID task in the *dyntet* software.

### 3.1.4 Trainee Prerequisites, Training Administration, and Training Length

The *dyntet* training program required that trainees understand the design, function, and operation of compressors. This previous knowledge could be introductory since practice with the application should yield a much deeper understanding. The program did not assume any previous TET experience. The *dyntet* program was designed to take place over 12 weeks and included short (~15 minute) weekly class listening sessions with the instructor. During class time, progress of each student was reviewed, problem areas received additional focus and instruction, and the next difficulty level was introduced. To address problem areas, students were asked to note stimuli they had difficulty with. During class time, the instructor reviewed those stimuli, pointing out specific aspects of the sound that students should pay attention to. To introduce the next difficulty level, the new task was explained, and several trials were completed as a class. During this, the instructor once again directed the student's attention to various important aspects of the sound.

While class time was important, trainees were required to practice individually for at least one hour per week. It was recommended that trainees practice in sessions of 30 minutes or less to reduce the effect of listening fatigue. It was also recommended that trainees practice in a conventional studio control room with full range loudspeakers. As discussed in Chapter 2, for maximum transfer of learned skills to real-world tasks, the training environment should mirror the working environment closely. Data was collected to examine the effects of training in non-ideal listening environments. Data was also collected to examine the effects of listening fatigue.

### 3.1.5 Program Structure

The training program structure was divided into four categories: *Ratio, Timing, Threshold,* and *Comprehensive.* These categories were each designed to target specific compressor parameters. The categories were divided into sub-levels of increasing difficulty.

#### 3.1.5.1 Ratio (Levels 1.x)

The ratio category was the first category presented to trainees. It was designed as an introduction to hearing DRC and involved manipulation of the ratio control. The threshold, attack, and release parameters were all preset and not changeable by the user for both the *question* and *answer*. To achieve an exact match between the *question* and *answer* audio streams the user only needed to adjust the ratio and make-up gain parameters. While the ratio was the parameter under focus, make-up gain was available for adjustment out of necessity. When manipulating the ratio to try and find a match, a higher ratio inevitably results in a quieter overall output that needs to be increased for direct comparison. The goal of the ratio level was to examine changes in dynamic range, due to a change in ratio, while holding the average volume of the stimuli constant. Therefore, the make-up gain parameter was provided to the user for adjustment of the average volume of the stimuli. To eliminate it from the training task, automatic make-up gain would need to be implemented, however, this does not reflect how the majority of DRCs operate. In addition, it's important for students to understand the importance of constant level matching when adjusting and listening to DRCs. The preset level of the make-up gain in the question, that students had to match, was selected to maintain a similar average volume to the stimuli when the DRC was bypassed.

In general, the threshold was preset so that the entire dynamic performance of the musical element fell above the threshold. With this setting, the ratio control influenced the entire dynamic range of the performance. It also made changes in the ratio obvious and easy to hear. The attack and release controls were preset "to taste" by professional mix engineers with a desire to impart as little artifact as possible on the ADSR envelope of the musical element. This is standard practice when the desired result is transparent compression. Table 1 shows a summary of the difficulty levels. Difficulty was manipulated by introducing additional steps in the ratio control and by introducing more difficult stimuli. For this level, it was hypothesized that stimuli with a larger initial dynamic range would be easier than stimuli with a smaller initial dynamic range. For example, a closely recorded vocal on the easy side of the spectrum and a distorted guitar amplifier on the difficult side. Both the matching and absolute ID tasks were included and DRC was applied in both mixing and mastering contexts.

Table 1: All difficulty levels available in the *dyntet* software.

| Category | Level | Task | Adjustable Parameters | Parameter Steps |
|---|---|---|---|---|
| **Ratio** | Level 1.0 | Matching | Ratio | Ratio (1:1, 8:1) (Vocal stimuli only) |
| | Level 1.1 | Matching | Ratio | Ratio (1:1, 8:1) |
| | Level 1.2 | Matching | Ratio | Ratio (1:1, 2:1, 8:1) |
| | Level 1.3 | Matching (buss comp) | Ratio | Ratio (1:1, 2:1, 4:1) |
| | Level 1.4 | Absolute ID | Ratio | Ratio (1:1, 2:1, 8:1) |
| **Timing** | Level 2.0 | Matching | Attack, release | Attack (.5 ms, 20 ms)<br>Release (50 ms, 1 s)<br>(Percussion stimuli only) |
| | Level 2.1 | Matching | Attack, release | Attack (.5 ms, 20 ms)<br>Release (50 ms, 1 s) |

| | | | | |
|---|---|---|---|---|
| | Level 2.2 | Matching (buss comp) | Attack, release | Attack (.5 ms, 20 ms)<br>Release (50 ms, 1 s) |
| | Level 2.3 | Absolute ID | Attack, release | Attack (5 ms, 20 ms)<br>Release (50 ms, 1 s) |
| **Threshold** | Level 3.1 | Matching | Threshold, ratio | Threshold (-15 dBFS, -30 dBFS)<br>Ratio (2:1, 20:1) |
| | Level 3.2 | Absolute ID | Threshold, ratio | Threshold (-15 dBFS, -30 dBFS)<br>Ratio (2:1, 20:1) |
| **Comprehensive** | Level 4.1 | Matching | Ratio, attack, release | Ratio (1:1, 2:1, 4:1, 8:1, 20:1)<br>Attack (.5 ms, 20 ms)<br>Release (50 ms, 1 s) |

#### 3.1.5.2 Timing (Levels 2.x)

The timing category was designed to teach trainees the effect of changing the attack and release parameters. In this category, the threshold and ratio were both preset and not changeable by the user. The user only controlled the attack, release, and make-up gain parameters. While changes in the attack and release parameters do not alter the overall output of the DRC as much as changes in the ratio, the make-up gain control was still provided to facilitate level matching during comparisons. The make-up gain parameter played this same role in all implemented training levels and was always available for adjustment. The threshold was preset the same as in the ratio category, so that the entire dynamic performance of the musical element fell above the threshold. The ratio was always set to 4:1. Again, DRC was present in both mixing and mastering contexts and both the matching and absolute ID tasks were used. Difficulty was controlled by linking and unlinking the attack and release controls together and by introducing more difficult stimuli. For this level, it was hypothesized that transient stimuli would be easier than steady-state stimuli. For example, percussion stimuli were thought to be easier than bass or soft synth keyboards. A detailed summary is presented in Table 1.

#### 3.1.5.3 Threshold (Levels 3.x)

The threshold category was designed to teach trainees the effect of changing the threshold parameter in combination with the ratio parameter. Engineers have the choice of using a high threshold with a high ratio or a low threshold with a low ratio to achieve a given amount of gain reduction on a signal. In fact, a variety of combinations in between these extremes could produce the same amount of average gain reduction. Again, the attack and release parameters were preset so that the compressor artifact was minimized. The user adjusted the threshold and ratio while using the make-up gain to facilitate overall level matching. DRC was presented in the mixing context only and both the matching and absolute ID tasks were used. Multiple difficultly levels were not used. A detailed summary is presented in Table 1.

#### 3.1.5.4 Comprehensive (Levels 4.x)

The comprehensive category was designed to bring all the parameters together in a more complex task. There was certainly a problem-solving component to this task given that several different combinations of parameters could sound quite similar. The category was used to demonstrate to the trainees the high level of interaction between the parameters. The category only had one level. It used the matching task in a mixing context. Details are presented in Table 1.

## 3.2 Software Application

### 3.2.1 Interface Design

The methodology described above was implemented in a software application that ran in a web browser using HTML, CSS, and JavaScript. The interface included controls for level selection, answer submission, next trial, the three audio streams, sound mute, and the six compressor parameters. There was also a gain reduction meter for the *answer* compressor that could be turned on or off. Finally, there was a “Cheat (solo)” function that allowed the user to examine the question and answer with the backing track muted, thereby presenting the single element on its own. This was only available after the answer was submitted and was included to allow the trainee to closely examine incorrect answers.

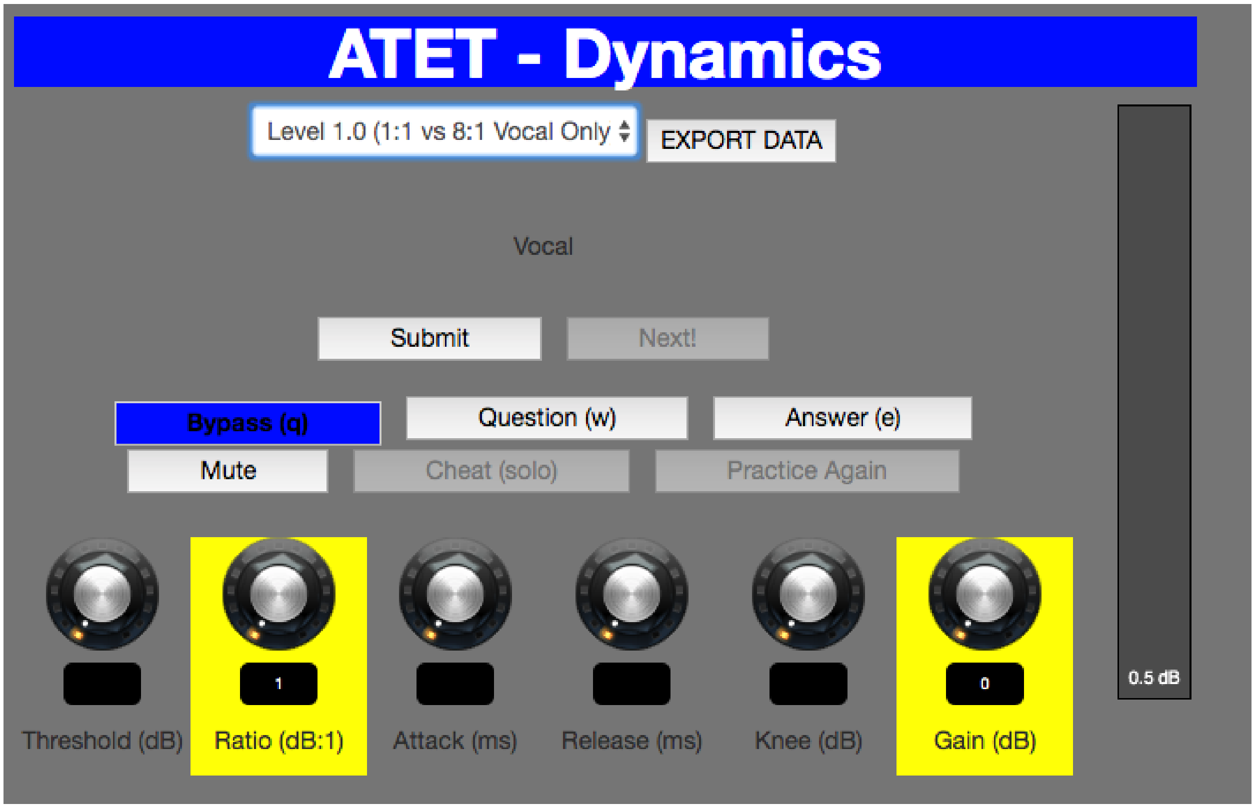


Figure 5: *dyntet* software user interface.

### 3.2.2 Audio Processing and Compressor Design

Real-time audio processing was possible in the browser using the Web Audio API [94]. The API had several pre-programed interfaces to handle audio storage, playback, routing, and signal processing. In the *dyntet* application, the various sound files were downloaded to the user's computer and stored for playback. The three audio streams played back simultaneously and were switched using quickly ramping gain devices for seamless transitions. The stock Web Audio API compressor implemented in Google Chrome featured pre-emphasis filters in the side chain and auto make-up gain that could not be turned off. Therefore, it was not useful for generalized training. The compressor used in *dyntet* was a modified version of the compressor programmed by Corey and Benson for use in the training software accompanying *Audio Production and Critical Listening: Technical Ear Training* [14]. The design was a generic digital compressor using a feedforward side-chain, log domain time constants, and a log domain peak level detector placed after the time constant calculation. A block diagram of the signal flow is displayed in Figure 6.

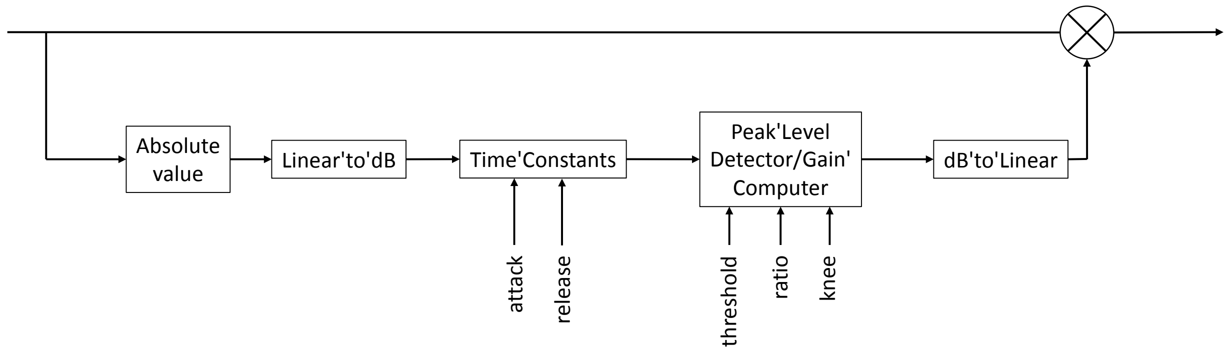


Figure 6: A block diagram of the signal flow of the compressor used in *dyntet*.

### 3.2.3 Software and Hardware Specification

The software was only debugged for Google Chrome but will likely work on all browsers in the future. The user was required to download around 200 MB of audio files upon loading the page. Since the digital signal processing requirement was low, there have not been any reported performance issues when used with any relatively recent computer model.

## 3.3 Program Material/Stimuli

Audio material was included in the *dyntet* application and spanned many genres (44 different musical excerpts). Each excerpt consisted of a musical element to be compressed and a corresponding backing track. In addition, some un-mastered mixes were included for the mastering context exercises. All audio came from professionally released mix sessions. The musical excerpts were 5-20 s selections that were automatically looped in the software. The wide range of loop lengths were required so that custom musically acceptable loops could be formed for each piece. Relatively short excerpts were created so that trainees could easily evaluate the same musical phrase over and over. As mentioned above, listening for compressor parameters involves listening to the evolution of sound over time: being able to compare the same short phrase back to back with different settings was critical. The trainee's short-term auditory memory would be exhausted as they waited for the same phrase to come around again in longer excerpts. It was hypothesized that loops around 10 s would be ideal. An effort was made to create loops as close to that length as possible, but sometimes musical phrases required a longer excerpt. It's possible that the loop length had an impact on the difficulty of the task.

Most stimuli were available at 96 kHz 24-bit resolution in an offline version of the software. The files were down-sampled to 44.1 kHz 24-bit to reduce download time in the online version. A summary of the stimuli can be seen in Table 2.

Table 2: List of stimuli included in the *dyntet* software.

| Name | Genre | Target Musical Element | Number of Excerpts | Stimuli Length per Excerpt |
|---|---|---|---|---|
| "Adultery" | Indie Rock | Acoustic Guitar | 2 | 24 s, 14 s |
| "Candy" | Indie Rock | Acoustic Guitar | 2 | 8 s, 7 s |
| "Story Teller Man" | Pop/Folk | Acoustic Guitar | 2 | 9 s, 9 s |
| "I Ja Fin" | Jazz/R&B | Drum Kit | 1 | 20 s |
| "She's Like" | Pop/Folk | Drum Kit | 2 | 14 s, 14 s |
| "CC Rider" | Blues | Electric Bass | 2 | 21 s, 10 s |
| "No God" | Indie Rock | Electric Bass | 2 | 14 s, 14 s |
| "Young Lover" | Pop/Folk | Electric Bass | 2 | 17 s, 18 s |
| "CC Rider" | Blues | Electric Guitar | 2 | 13 s, 5 s |
| "Future Punch" | Jazz | Electric Guitar | 2 | 11 s, 7 s |
| "Like Yr Type" | Electronic | Keyboard | 1 | 11 s |
| "Future Punch" | Jazz | Kick Drum | 2 | 11 s |
| "Bury Me" | Pop | Master Buss | 2 | 14 s, 14 s |
| "Cloud Of Smoke" | Acoustic | Master Buss | 1 | 26 s |
| "I Will Be Water" | Electronic | Master Buss | 1 | 9 s |
| "In My Blood" | 80s Pop | Master Buss | 1 | 13 s |
| "Psycho Activity" | Jazz | Master Buss | 1 | 10 s |
| "A Vision Of You" | Jazz | Piano | 2 | 10 s, 10 s |
| "Fever Nostalgic" | Electronic | Snare Drum | 2 | 10 s, 20 s |
| "Revolution" | Jazz/R&B | Snare Drum | 1 | 14 s |
| "A Vision Of You" | Jazz | Vocal | 3 | 12 s, 9 s, 8 s |
| "CC Rider" | Blues | Vocal | 3 | 10 s, 5 s, 7 s |
| "In The Meantime" | R&B, Funk | Vocal | 3 | 7 s, 7 s, 5 s |
| "Tell Me" | R&B | Vocal | 2 | 8 s, 5 s |

The audio material is likely the most critical asset when developing a TET application for DRCs. Access to the professionally released mix sessions was needed and required all signal processing intact and adjustable. It was also important that the target element have no previous compression applied. With this, it was possible to mix down the target individual elements without compression but with all the other effects engaged (ex: equalization). A corresponding backing track (the full mix minus the individual element) was also prepared. As discussed in Chapter 2, this wide range of stimuli is important for the generalization and transfer of learned skills. These requirements make assembling a large library very difficult. In this study, these multi-track sessions were sourced from the authors' professional work as well as particularly strong class projects completed by past students at McGill University. Criteria for selecting stimuli were high-quality, subjectively requiring compression, varied instrument types, and the other aspects mentioned previously.

## 3.4 Grading and User Evaluation

In *dyntet*, a simple preliminary grading system was implemented. If the trainee got all the required DRC parameters correct, they received full points. If the trainee got any of the required parameters wrong, they received zero points. Make-up gain was excluded from the grading since it was only included as a level matching tool. The goal of the training was not to examine the students' ability to level match stimuli, although this might be a by-product of the training. The application displayed an ongoing success rate to the user.

## 3.5 Comparison to Previous Training Programs

The *dyntet* training program had many features in common with Corey & Benson's DRC TET application [14]. When comparing, a few advantages of the *dyntet* software emerged; 1) Many different types of high-quality program material were included with the software. This offered a wide range of stimuli for convenience and promoted transfer of skills. 2) Training was available in a mixing context in addition to a mastering context. This offered a wider range of experiences, and again, promoted the transfer of skills. 3) The compressor settings that trainees had to match/absolute ID were selected from an ecologically valid set of presets (custom creations from professional mixing engineers). The presets were generated within the constraints of the "chunks" and endpoints previously defined for each parameter. Corey & Benson made use of randomly generated settings. Their approach likely came from successful work in equalizer TET where randomly choosing a center frequency to boost or cut will often result in an audible effect.

Unfortunately, in DRC TET randomly generated settings can result in excessively subtle or obvious effects. Because of this, Corey & Benson were limited to what DRC parameters could be included in the training and instructors were unable to completely control the intensity of the effects.

# 4 Testing Training Effect I: Training Data

## 4.1 Introduction

A TET program designed to teach audio students to memorize the audible effects of DRCs was developed and described in Chapter 3. This TET program was used to train eight students and student training data was collected during all training sessions. The following analysis was an investigation into student improvements on the various training tasks. An analysis is shown of the trial by trial response data generated by all students while using the *dyntet* training application. This evaluation of improvement, or lack thereof, was considered "in-context" because the data comes from the training environment itself. If improvement was measured it could be due to any number of the following factors; 1) familiarization with the testing application, 2) familiarization with the stimuli, 3) familiarization with the compressor algorithm, 4) increased listening skill level due their ongoing education, 5) increased listening skill level due to the training application. Factors 1-3 would be of interest to some researchers training listening test participants. However, they are less interesting to an educator training audio engineers. The *dyntet* training program was designed for audio students and was considered a success if it improved Factor 5: increased listening skill level due to the training application. Those skills would likely be transferrable to every day audio engineering tasks. Examination of the in-context training data was the first step in evaluating the influence of these factors on student improvement.

The main research questions for this study were: do student skill levels improve with increasing amounts of training? What is the average rate of improvement? What is the shape of the learning curve? Several other research questions were of interest: does student performance depend on the time of day of training? Does student performance depend on the use of loudspeakers or headphones? Is the listening room acoustic a factor? Do students differ in their overall skill levels? Do students differ in their improvement rates and curves? Are the difficulty levels equally spaced? How long does it take each student to complete a level? Are there too many gradations in difficulty, or too few? Are all the stimuli within any one level equally difficult?

## 4.2 Methodology

The training program was administered as part of the Advanced Technical Ear Training (ATET) course offered by the Graduate Program in Sound Recording at McGill University, Montreal, Canada. This seminar follows the Technical Ear Training course that students take in first year of their studies. The preliminary TET course focuses on equalizer training. ATET focuses on several other audio attributes: dynamic range (compression), loudness, spatial attributes, special audio effects, etc. A detailed description of the ATET course is available in Martin & Massenburg [55].

The DRC TET described in the previous chapter took place over a period of 14 weeks. The training was completed as take-home assignments. The students had the flexibility to train whenever and wherever they wanted: in the studio, at home, on loudspeakers, on headphones, etc. Students were asked to train a minimum of one hour per week. The students' work was reviewed every two weeks. Awards were given out during the in-class reviews as an additional incentive and to "gamify" the training. Awards were offered for the highest number of correct answers, most time training, and to the student who improved the most. Table 3 shows the schedule of training. Students were instructed to start with the first level within each category and only progress to the next level when they could consistently achieve 80 % correct. It was expected that students would complete a level every 1.5 – 2 weeks (90-120 minutes of training).

Table 3: DRC training schedule for the 2017-2018 "Advanced Technical Ear Training" course.

| Time Period | Level Number | Level Category | Adjustable Parameters |
|---|---|---|---|
| **Weeks 1-6** | Level 1.x | Ratio | Ratio |
| **Weeks 7-10** | Level 2.x | Timing | Attack, Release |
| **Weeks 11-12** | Level 3.x | Threshold | Threshold, ratio |
| **Weeks 13-14** | Level 4.1 | Comprehensive | Threshold, ratio, attack, release |

The following data was collected by the software for every trial during training: trial start date and time, trial end date and time, stimuli used, training level, *question* compressor settings (unknown settings to match), *answer* compressor settings (student's attempt at matching the *question* compressor), participant identification, audio monitor type (speakers or headphones), listening room type (control room, treated room, un-treated room), and student comments. The latter three data points were self-reported by the participants. The listening room types had the following descriptions: control rooms were purpose built acoustically treated listening rooms with a low noise floor and high-quality reference monitors, un-treated rooms had no acoustic treatment, treated rooms were somewhere in between control rooms and un-treated rooms. For example, a dedicated room at the student's residence containing a modest application of acoustic treatment.

## 4.3 Results

The eight students involved in the training program had a mean age of 25.13 (SD = 2.42) years, a mean of 14.63 (SD = 5.13) years of musical experience and a mean of 4.31 (SD = 2.05) years of recording experience. Almost all the students in the Masters in sound recording program held a Bachelor of Music degree as part of the admission requirements. This explained the high number of years of music experience. Some of those students had recording experience from their previous degrees and others began studying the topic upon arrival at McGill University. The least experienced students had a minimum of two years of intensive sound recording training when beginning the training program. Overall, the students can be described as highly motivated and skilled audio students.

The students trained every second week between Oct. $6^{th}$, 2017 and Feb. $22^{nd}$, 2018, with a month-long holiday break during Dec. 2017. The cumulative number of hours spent training and cumulative number of trials completed for each student is shown in Table 4. The mean number of hours was close to the assigned number of hours, however, there was a great deal of variance in the number of trials completed within those hours.

Table 4: *Cumulative time* (hours) and *cumulative trials* for each student over the entire training period.

| Student | Cumulative Time (hours) | Cumulative Trials |
|---|---|---|
| **1** | 11.77 | 660 |
| **2** | 12.93 | 615 |
| **3** | 12.77 | 895 |
| **4** | 11.13 | 708 |
| **5** | 10.28 | 607 |
| **6** | 6.93 | 369 |
| **7** | 12.33 | 1686 |
| **8** | 14.27 | 1183 |
| **Mean (SD)** | 11.55 (2.22) | 840.38 (416.27) |

The distribution of those hours over each level is visualized in Figure 7. Within the 1.x series of levels (ratio category), the students progressed very quickly through Level 1.0, quickly through Level 1.1, and spent the bulk of the time on Level 1.2. Some students spent a fair amount of time on Level 1.3. Only a few students made it to Level 1.4. Within the 2.x series of levels (timing category), students progressed very quickly through Level 2.0 and spent the bulk of the time on Levels 2.1 and 2.2. Only one student made it to Level 2.3. Overall, students did not spend a lot of time on the remaining levels. The data from Level 2.3 was removed from the following analysis since only one student spent any time on it. The data from Level 4.1 was removed because no students achieved a correct response. The data from Level 1.0 was also removed because students scored almost all responses as correct. This level was too easy.

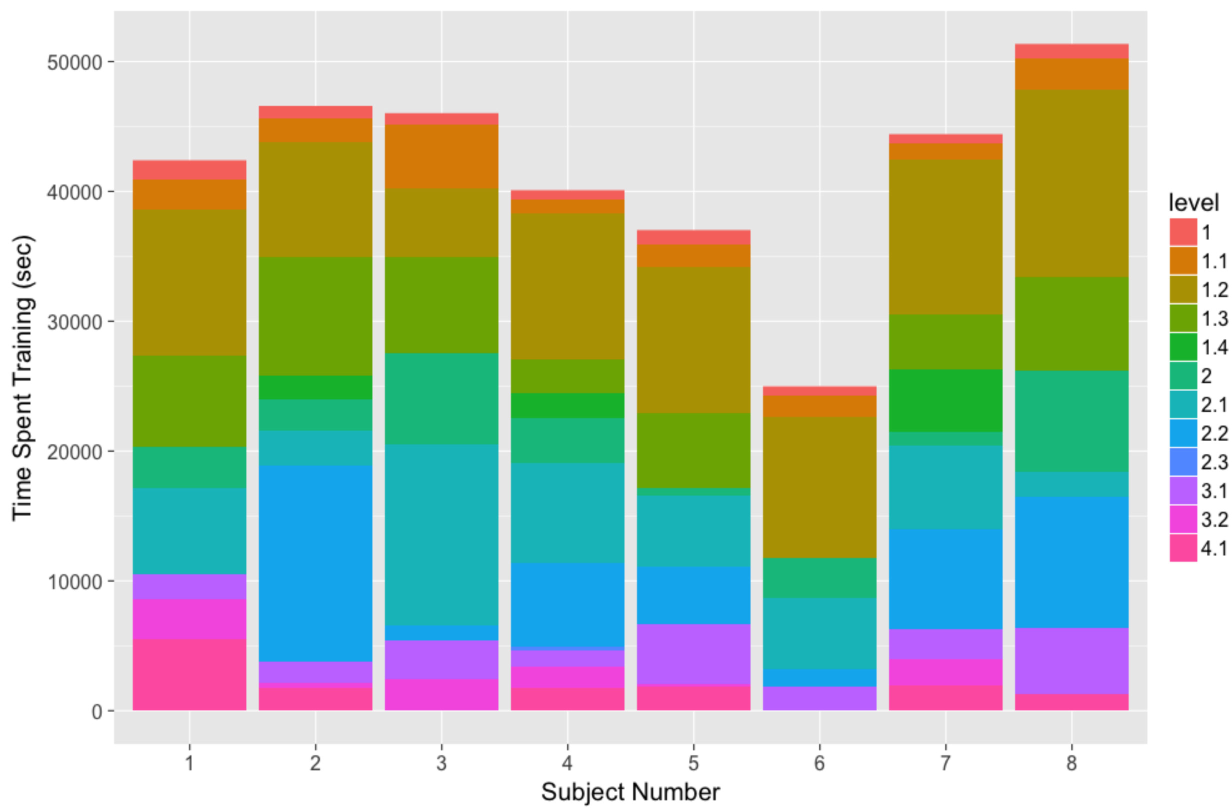


Figure 7: The number of seconds spent training on each level for each student.

### 4.3.1 Defining Skill Level and Amount of Training Completed

A couple smaller questions were tackled before addressing the main research questions. How should student skill level be defined within the context of the TET program? The obvious statistic was the percentage of correct responses. However, some previous studies had also shown an improvement in student response time. To examine this, the relationship between the number of correct responses and the time taken to complete a trial was plotted. This relationship is shown for each student in Figure 8.

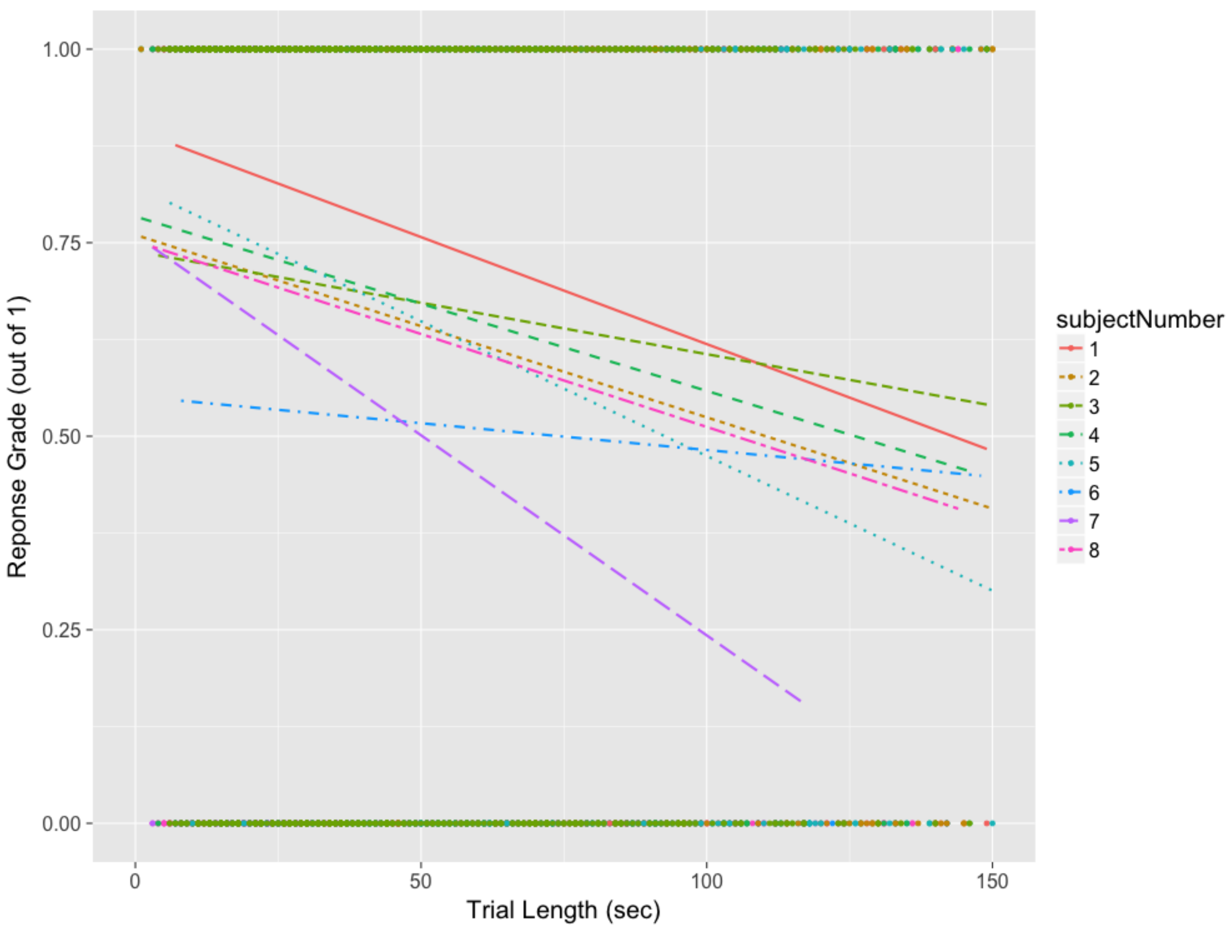


Figure 8: A scatter plot with linear regression lines fitted for each student. The plot shows the relationship between achieving a correct response, and the amount of time taken to complete the trial.

A mixed-effects logistic regression model was built on the data using *trial time* to predict *correct response (True or False)*. S*ubject number* was input as a random effect (intercept) to compensate for the overall differences in ability between participants. The results are presented in Table 5 and show that for each increase in one second of trial time the odds of a correct response were 1.01 times lower. It was hypothesized that the percentage of correct responses and student response times would be correlated and that they both would represent different parts of the overall skill level. The results were consistent with the hypothesis. The model had a very low $R^2$ value, so there was a great deal of unexplained variance between the variables. Therefore, while *correct response* and *trial time* were related, they were likely measuring different characteristics of overall skill level. These aspects were considered separately in the following analysis.

Table 5: Results of the mixed logistic regression model. *Trial time* as predictor of *correct responses, subject number* as a random effect. Model $x^2(1) = 103.91$, $p < .001$. $R^2 = .022$ (Hosmer-Lemeshow).

| | B (Standard Error) | Odds ratio (Confidence Interval) | *p* |
|---|---|---|---|
| Intercept | 1.05 (0.11) | | <.001 |
| Trial Length | -0.01 (0.000010) | 0.99 (0.99 - .099) | <.001 |

A second question that had to be answered was how should the amount of training that has been completed be measured? It made sense to examine the cumulative amount of time the student had spent training and the cumulative number of trials the student had completed. A repeated measures correlation coefficient was computed for the two variables. *Cumulative time* and *cumulative trials* were highly correlated ($r(6404) = .93$, $p < .001$). *Cumulative time* and *cumulative trials* were very similar and both metrics were considered in the main analysis below.

### 4.3.2 Main Analysis: Participant Improvement

Both performance metrics (*correct response* and *trial time*) were examined in relation to both measurements of training (*cumulative trials* and *cumulative time*) to determine whether students improved with increasing training.

#### 4.3.2.1 Number of Correct Responses

The evolution of each student's grade (1 for correct, 0 for incorrect) over the student's cumulative number of minutes spent training is visualized in Figure 9. Examination of this plot revealed that 43 out of 62 linear regression lines for each level, within each student, indicated an increase in the students' grades with increasing time spent training. A similar plot with similar results was constructed with the number of *cumulative trials* as the predictor. In this plot, 44 out of 62 linear regression lines indicated an improvement in the number of correct responses. A linear model was chosen to represent the learning curves since the dataset was limited in size, the training period was not particularly long, and it was the simplest model available to explain the data. As concluded in the literature review, the learning curve for low numbers of hours spent training can be approximated as linear. With a much higher number of training hours, a logarithmic curve would have likely been a better fit to the data.

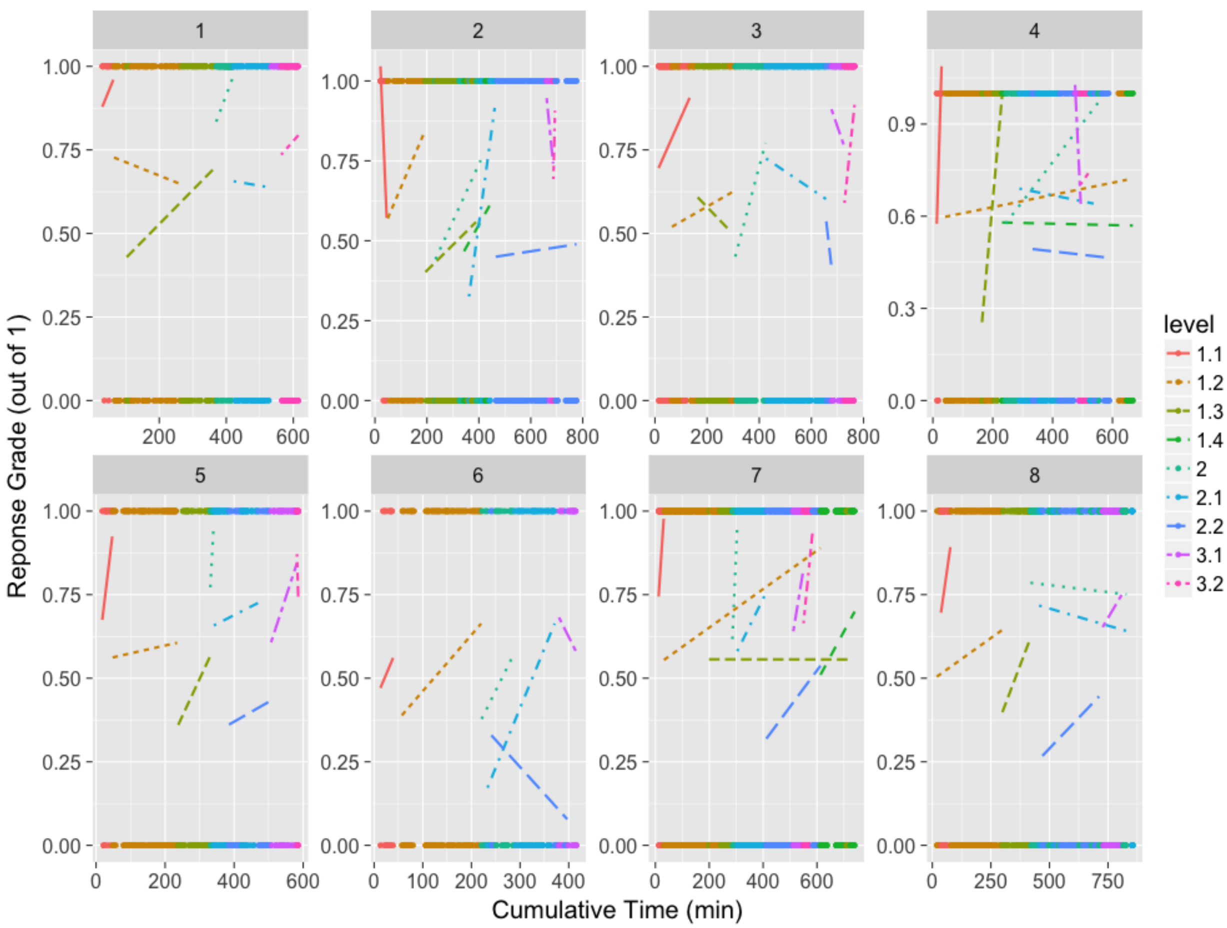


Figure 9: A scatter plot with linear regression lines fitted for each student and each level. The plot shows the change in student grades with increasing number of minutes spent training. The student numbers are indicated at the top of each graph.

Two mixed-effects logistic regression models were fit to the data: 1) using *cumulative time* to predict *correct response,* and 2) using *cumulative trials* to predict *correct response* (Table 6).

The *subject number* and *level* variables were input as random effects (intercepts) to compensate for any differences in skill level between students and any differences in difficulty between levels.

Table 6: Results of the two mixed logistic regression models. 1) *Cumulative time* and 2) *cumulative trials* as predictor of *correct responses. Subject number* and *level* as random effects.
Model *cumulative time* $x^2(1) = 16.60$, $p < .001$. $R^2 = .014$.
Model *cumulative trials* $x^2(1) = 8.39$, $p < .01$. $R^2 = .0071$.

| | B (Standard Error) | Odds ratio (Confidence Interval) | *p* |
|---|---|---|---|
| Intercept | 0.18 (0.25) | | |
| Cumulative Time (Min) | 0.0011 (0.00027) | 1.0011 (1.00056 – 1.0016) | <.001 |
| | | | |
| Intercept | 0.41 (0.23) | | |
| Cumulative Trials | 0.00042 (0.00015) | 1.00042 (1.00014-1.00072) | <.01 |

Both *cumulative time* and *cumulative trials* were statistically significant predictors of *correct response*. The Chi-Square and $R^2$ values for each model revealed that using *cumulative time* to predict *correct response* was a better fit than using *cumulative trials*. It was found previously that *cumulative time* and *cumulative trials* were highly correlated, as expected, and therefore can't both be used together in a predictive model. Therefore, only the first model using *cumulative time* was examined further, since it better represents the amount of training required to improve a student's success rate. Averaged across all students and all training levels, for each minute a student spent training the odds ratio of getting a correct response increased by 1.0011. When a student completed one minute of training the odds of them getting a correct response was 1.0011 times greater than before that one minute of training. For example, if a student began with a success rate of 50 % and completed 11.55 hours (693 minutes) of training (the average in this study) it was expected the odds of them getting a correct response would be 2.14 (CI = 1.47 – 3.03) times greater. That equals a success rate of 68 % (CI = 60 % – 75 %) after training. Or, an improvement of 1.6 (CI = 0.9 – 2.2) points per hour. It was hypothesized that students would improve along both performance metrics: the percentage of correct responses and response time. These findings supported the hypothesis that students' success rates would improve with increasing training. A linear curve was used in this analysis for the reasons described above.

### 4.3.2.2 Time Taken per Trial

The evolution of each student's response time over the student's cumulative number of minutes spent training is visualized in Figure 10. The trial time was calculated as the difference between the moment the student clicked "Next Trial" and the moment they submitted their response. Only trials with correct responses were considered. It would have been very easy for a student to enter an incorrect response very quickly, but that would not have represented better performance. Therefore, the following was an examination of the change in the students' time to arrive at a correct response. Examination of this plot revealed that only 34 out of 62 regression lines for each level, within each student, indicated a decrease in the students' response time. A similar plot with similar results was constructed with the number of cumulative trials as predictor. In this plot, only 29 out of 62 regression lines indicated an improvement (decrease) in response time.

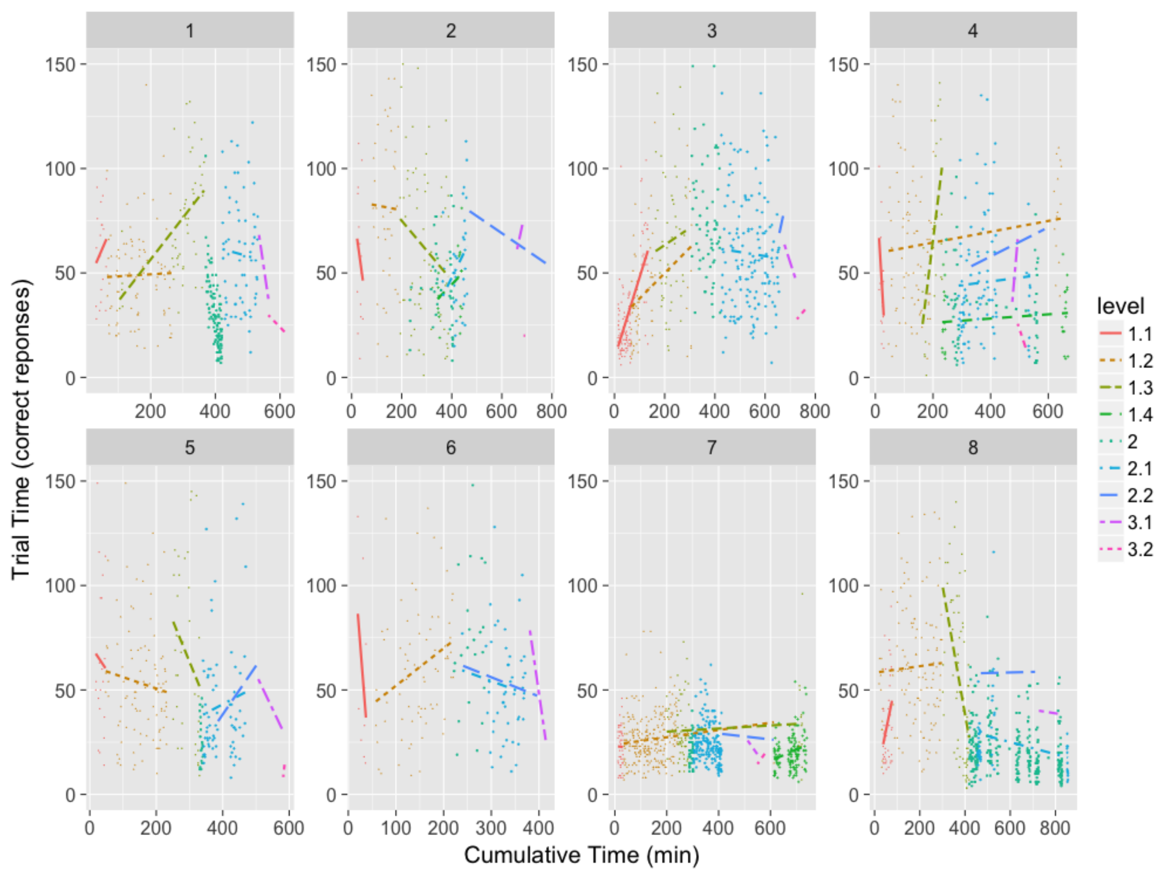


Figure 10: A scatter plot with regression lines fitted for each student and each level. The plot shows the change in time taken to arrive at a correct response with increasing number of minutes spent training. The student numbers are indicated at the top of each graph.

Two linear mixed-effects regression models were fit to the data: 1) using *cumulative time* to predict *trial time,* and 2) using *cumulative trials* to predict *trial time* (Table 7). The *subject number* and *level* variables were input as random effects (intercepts). A linear model was chosen again for the reasons described in Chapter 4.3.2.1.

Table 7: Results of the mixed regression models. 1) *Cumulative time* and 2) *cumulative trials* as predictors of *trial time, subject number* and *level* as random effects.
Model *cumulative time* $x^2(1) = 2.8495$, $p = .091$. $R^2 = .0028$.
Model *cumulative trials* $x^2(1) = 0.22$, $p = 0.64$. $R^2 = .00026$.

| | B (Standard Error) | T | *p* |
|---|---|---|---|
| Intercept | 54.15 (6.91) | 7.832 | |
| Cumulative Time (Min) | -0.011 (0.0067) | -1.68 | 0.091 |
| | | | |
| Intercept | 48.51 (6.98) | | |
| Cumulative Trials | 0.0020 (0.0038) | 0.525 | 0.064 |

Neither *cumulative time* nor *cumulative trials* were significant predictors of *trial time*. It appears that the training did not influence the time required to arrive at a correct response. This finding was opposite the hypothesis that students' response times would improve with increasing training. Since the trial time was measured from the moment the student started the trail, to the moment they submitted their answer, it's possible the measurement does not reflect the amount of time the student was actually working. For example, a student may have begun a new trial but then taken a break before continuing. Where possible, these obvious outliers were removed from the dataset.

### 4.3.3 Secondary Analysis: Additional Factors

For the secondary analysis, student skill level was only examined as a function of *correct response*. The amount of training completed was only examined as a function of *cumulative time* spent training.

#### 4.3.3.1 Effect of Time of Day

Figure 11 shows the relationship between the time of day of training and the student success rate. There was no previous hypothesis as to the shape of the regression line, so an exploratory locally weighted regression line was fit to the data. Several different polynomial mixed-effects logistic regression models were fit to the data to describe the trend, but a statistically significant fit was not achieved.

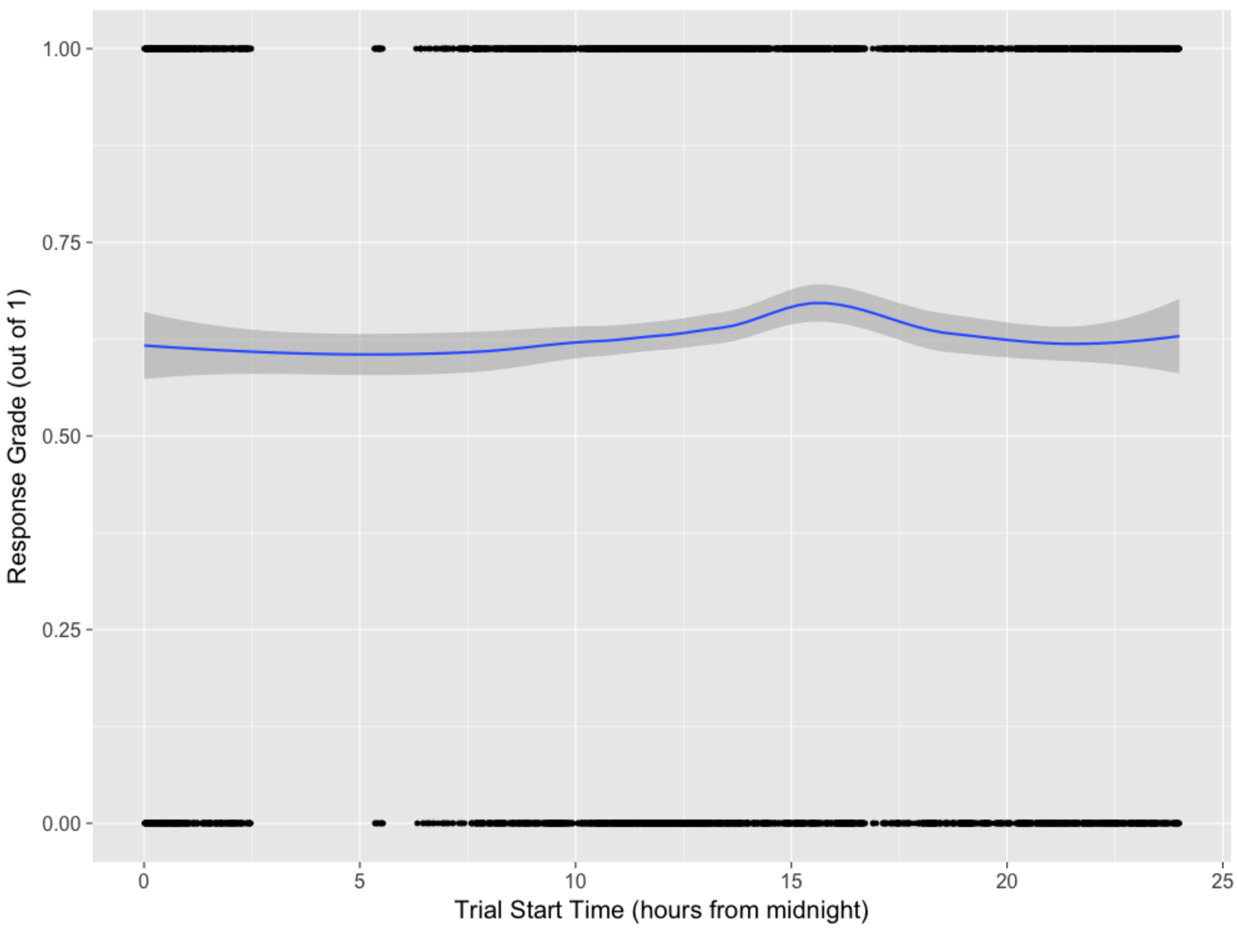


Figure 11: A scatter plot with a locally weighted regression model was fit to the data ("LOESS" model in R, ggplot2 [9]) The plot shows the change in the overall student success rate as a function of time of day. The shaded area represents the standard error.

#### 4.3.3.2 Effect of Listening Environment

Data was collected to describe each student's listening environment for each trial completed. This data was compiled to be one of four categories: *headphones*, speakers in a *control room*, speakers in a *treated room*, or speakers in an *untreated room*. The average success rate for each student as a function of listening environment is visualized in Figure 12. The first observation was that not all students trained in all possible locations. The second observation was that there appeared to be an interaction effect between the *subject number* and *monitors*. Students appeared to differ in which environment led to the best performance. Unfortunately, there was not enough data to construct a stable model to examine this interaction. In addition, a mixed-effects logistic regression model showed that *correct response* could not be significantly predicted by *monitors* (*subject number* and *level* as random effects). It was expected that students would perform better on loudspeakers than headphones. When using loudspeakers, it was expected that performance would be better in acoustically treated control rooms compared to untreated living spaces. This was based on professional experience and the results of King et al. and Leonard et al. [43][49]. Overall, the relationship between *monitors* and success rate was undetermined in this study.

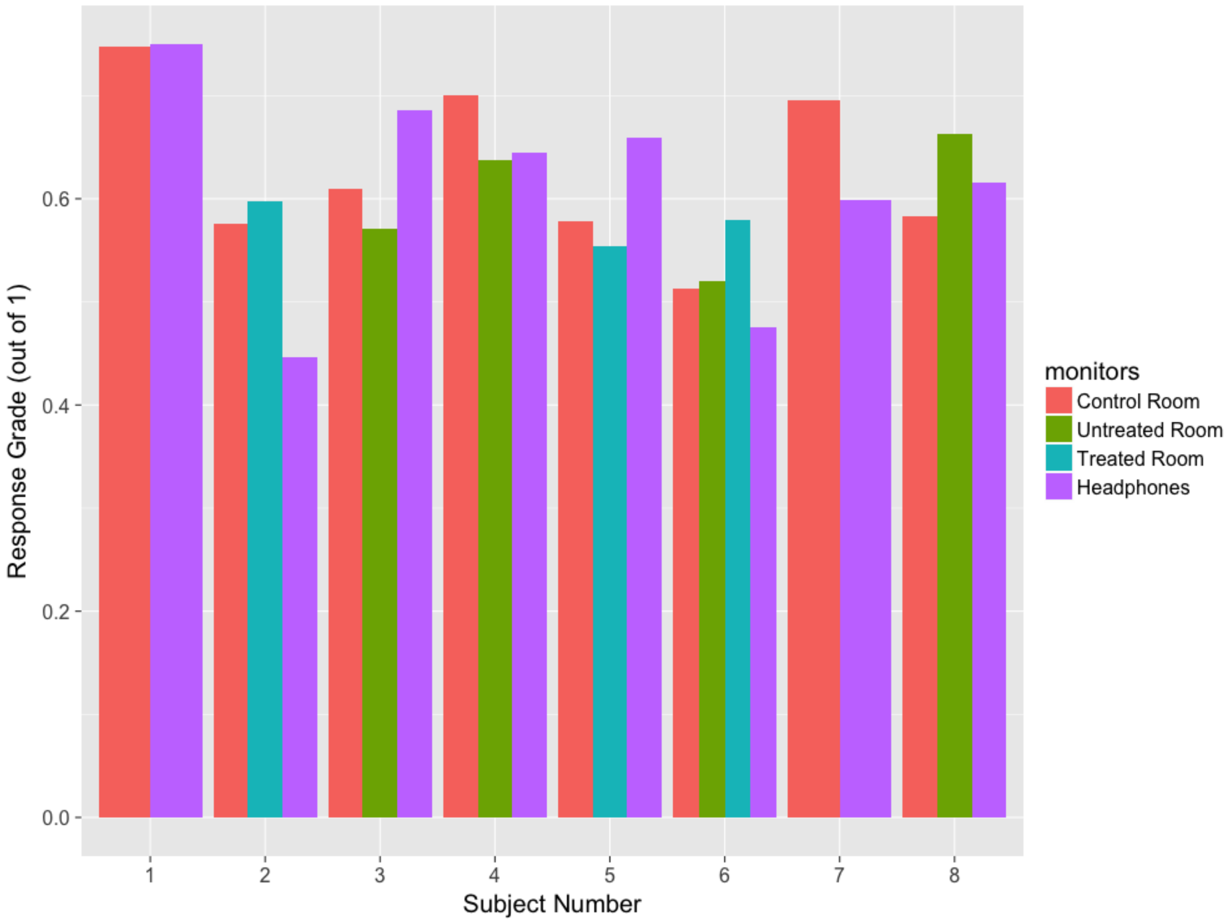


Figure 12: A bar plot displaying the probability of a correct response as a function of listening environment, for each student.

#### 4.3.3.3 Examination of the Training Structure

Figure 13 shows that the difficulty and progression through each series of levels was not ideal. The difficulty levels should have been equally spaced. The students should have taken roughly the same amount of time to complete each level. The gradations in level should have been set such that students were moving quickly enough through the levels to feel confident in their own progression but not so fine that students could complete a subsequent level without any additional training. Only Level 2.0 seemed to follow that objective. The rest of the levels seemed to start too difficult and take much longer to arrive at the target success rate (1.2, 1.3, 1.4, 2.1, 2.2) or the target success rate was achieved almost immediately (Levels 3.1 and 3.2).

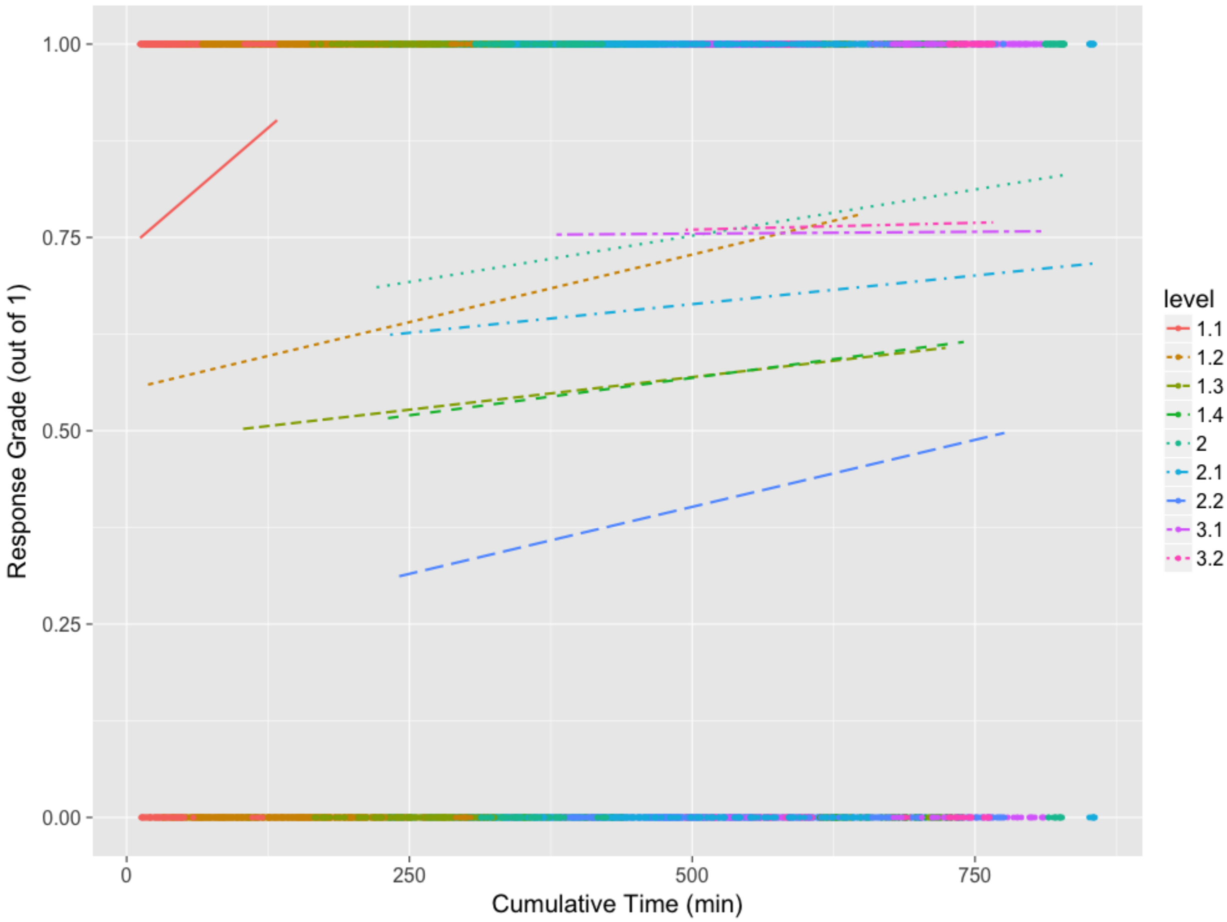


Figure 13: A scatter plot displaying the probability of a correct response for each level as a function of cumulative time spent training (improvement). Data has been averaged across all students.

Figure 14 shows that there was massive variation in the difficulty of each stimulus within each level. While this result was expected, the goal of the training software and preset settings was to have equal difficulty across all stimuli within a given level.

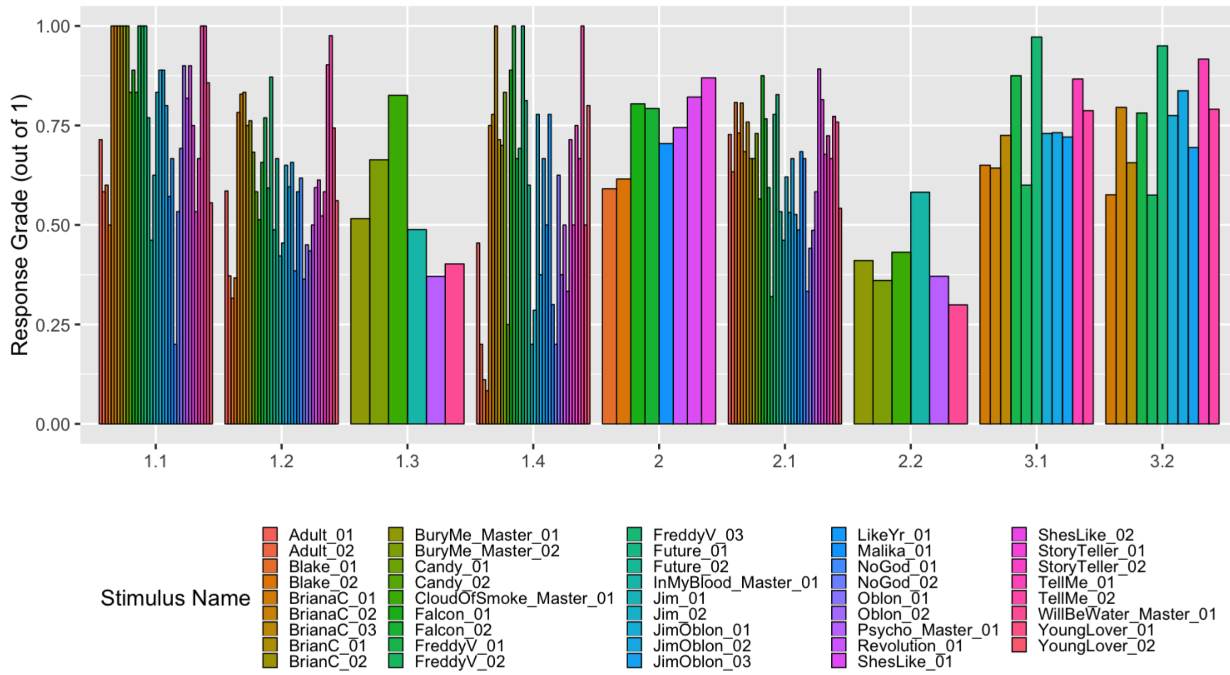


Figure 14: A bar plot displaying the probability of a correct response as a function of the stimulus. Data for each level is treated separately.

### 4.3.3.4 Differences Between Participants

Figure 15 shows that there were differences in the overall skill level between students. Teaching experience led to the hypothesis that students would differ in their skill levels, roughly following a standard distribution of scores. It was also expected that students would differ in their improvement rates and curves, however, the nature of those differences were difficult to estimate. This result supported the hypothesis. A mixed-effects logistic regression model was built with *subject number* predicting *correct response* (random effect *level*). This model was a significantly better fit to the data than the baseline model excluding the *subject number* predictor ($x^2 = 56.41$, $p < .001$). Multiple comparisons of means were conducted using Tukey Contrasts and showed that this difference in performance existed between several pairs of participants. Participant 1 performed significantly better than Participants 3, 5, 6, 7, and 8 and Participant 6 performed significantly worse than all other participants. It is interesting to note that Participant 1 completed an average amount of training and that Participant 6 completed the least amount of training. There was likely a difference in the improvement rate between students. Unfortunately, there was not enough data to build a stable model to examine this relationship.

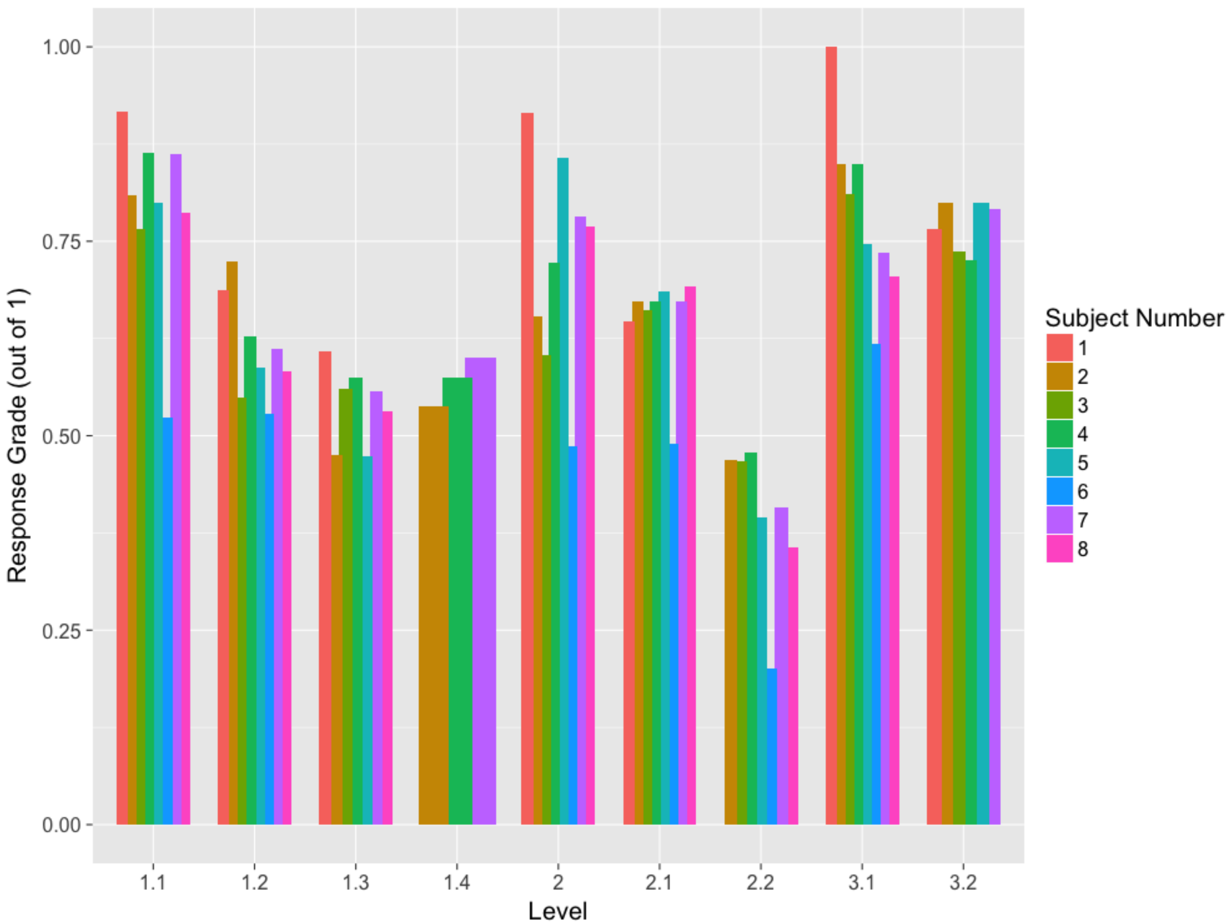


Figure 15: A bar plot displaying the probability of a correct response as a function of subject number. Data for each level is treated separately.

## 4.4 Discussion

Based on the analysis above, the training was considered to be successful "in-context". All students demonstrated improvement on many training levels. An overall improvement was achieved when student data was averaged. No direct comparisons of the rate of improvement could be made to other DRC TET programs, but comparisons could be made to equalizer TET programs. In Chapter 2, the average short-term and long-term learning rates were calculated as 3.7 points per hour and 0.34 points per hour respectively. In the *dyntet* study, the rate was 1.6 (CI = 0.9 – 2.2) points per hour. Overall, the improvement rate achieved in the *dyntet* study seemed similar to the rates seen in equalizer TET.

It was unfortunate that there was not enough data for each student within each level to examine the precise shape of the learning curve. A linear model was chosen. The implication of this choice is that the regression model is not be useful for predicting user performance levels for training times outside the window of available data. For example, the model is not useful for predicting the performance level of a hypothetical student who has completed more than 14 hours of training.

It was surprising that the time required to arrive at a correct response did not decrease with training. In this study, students were encouraged to take their time when answering questions, not to go as quickly as possible. The correlation between response time and percentage of correct responses show that the difficulty of the task was, at the very least, a factor in the response time. Adjustments to the training program should be made to encourage students to work on that part of their overall listening ability since response time is considered a valuable skill for an audio engineer. Another possible explanation for this result is to reemphasize that compressors are non-linear and time-variant devices. To hear their effects, students must listen to the excerpt for a minimum amount of time. This is not the case in equalizer TET, where the difference is heard immediately. All previous reports of improvements in response time have been in TET programs targeting linear and time-invariant processors. Finally, response time may be a problematic performance variable to begin with, since it does not measure it directly. For example, students may lose focus for a duration of time, take a break, or decide to pay particular attention to a given trial. All of these actions would result in a longer response time but are not necessarily directly related to task performance.

In this dataset, *cumulative time* was a better predictor of student improvement than *cumulative trials*. It would be interesting to see if this relationship holds true across other TET scenarios. This suggests that there is no advantage for students to quickly move through trials. Instead, it may be better to take time and carefully focus on each question.

When looking at the secondary research questions, students appeared to have peak performance at 16h00. This might seem surprising at first since the average Canadian work day ends around 17h00. However, reviewing the students' class and studio schedules revealed their classes rarely began before 10h00 and they were often involved in recording sessions in the evening and overnight. Therefore, 16h00 may have felt more like mid-day to these "night owls".

When looking at the effect of listening environment, previous research would suggest there might be a difference in mixing consistency between listening on loudspeakers versus headphones [43][49]. A similar difference was not seen in this data set. Trends suggest there may be a participant dependent effect. The effects of playback device and control room acoustics on audio activities is still not totally understood. The continued collection of training data will hopefully provide insight into these potentially cost saving alternatives.

Examination of the overall training structure revealed a few areas where it should be optimized. Some levels were either too easy or too difficult. Intermediate levels are required in those cases. The goal was to equalize the relative differences in difficulty between each successive level. Making this change will allow students to expect regular well-timed positive feedback as they complete each level in the target range of 90-120 minutes of training. There was also a massive difference in the difficulty of each stimuli within each level. This was not desired. Students should not have felt penalized when a particularly difficult stimulus was randomly chosen for them. This imbalance should be equalized by adjusting the bank of compressor presets. It is hypothesized that with all these changes the rate of improvement could be increased.

Finally, two things should be considered when interpreting these results. First, the measured student improvement could be due to any combination of the five factors discussed in Chapter 4.1. Second, while many data points were collected and statistical significance was achieved for the main analysis, the sample subject pool only consisted of eight students and might not be representative of the entire audio student population.

# 5 Testing Training Effect II: In-Context Listening Test

## 5.1 Introduction

The second method used to measure improvement in this study was to run a series of controlled listening tests for several listener groups. With the first test (hereafter named pre-test) given before training began and the second test (post-test) given after the training, differences in performance between the two tests might be attributed to the training program. Three separate listening tests were administered pre- and post-training to measure three separate applications of learned skills: in-context, near-transfer, and far-transfer. This chapter discusses the in-context listening test. The test is considered in-context because it is Level 1.2 from the *dyntet* program. The only difference between this data and the data analyzed for Level 1.2 in the previous chapter is this data comes from a more tightly controlled listening test. The near-transfer listening test is the topic of Chapter 5 and the far-transfer listening test is the topic of Chapter 6.

The in-context listening test was administered to three different groups of listeners. The first group was the trained students discussed in the previous chapter. If improvement was measured between the pre-test and post-test it would be difficult to know exactly which factors were contributors. To review, it could be due to any number of the following factors: 1) familiarization with the testing application, 2) familiarization with the stimuli, 3) familiarization with the compressor algorithm design, 4) increased listening skill level due to their ongoing education, 5) increased listening skill level due to the training application. A control group was created to eliminate other factors since Factor 5 is of most interest in this study. This control group was made up of eight students from the first year of the same graduate program and did not receive the *dyntet* training. While eliminating Factors 1-3 will be the subject of later chapters, this control group was created to eliminate Factor 4: increased listening skill level due to their ongoing education. It was assumed that improvement due to their ongoing education would be equal for both listener groups. Unfortunately, they were not receiving precisely the same instruction or starting with the same overall skill level since the students were in different years of study. In addition, the control group was actively participating in an equalization TET program throughout the year which could have contributed to developing their overall listening abilities. While this introduced confounding variables, an alternative was not available. A third listener group was included as an additional benchmark. This group was made up of working professionals who graduated from the program.

Quesnel showed that audio students could develop listening skills through TET that professionals learned over years of work [82]. In this study, professionals were chosen who had significant equalizer TET experience but who had no DRC TET experience. This way, the professional group would not be at a disadvantage because of an unfamiliar testing interface.

The first research question in this listening was will the untrained control group improve from pre-test to post-test? The main research question was, will the trained students improve more than the untrained students? In addition, how will the professional graduate group compare? Will evidence of improvement be visible in the participant comments and difficulty ratings? Will the stimuli affect participant performance? Will the time of day effect performance? Will the listening test block order effect performance?

## 5.2 Methodology

### 5.2.1 Participants

Participants were voluntarily recruited for the listening test via email notifications sent via McGill University internal e-mail groups. All students (16 total) in the MMus in Sound Recording program volunteered to participate. In addition, eight professional audio engineers who are also graduates of the program volunteered. All eight trained students and all eight professionals participated in all tests. All eight untrained students participated in the pre-tests but only seven were available to participate in the post-tests.

### 5.2.2 Listening Task

The listening test was derived from Level 1.2 of the *dyntet* training software. The task was identical: to adjust and match the settings of a compressor to a set of unknown reference settings by ear. Like Level 1.2, the task was simplified so that the participant only had to match the ratio and make-up gain parameters. All other compressor parameters were pre-matched. Again, the controls were divided into discrete steps (ratio = 1:1, 2:1, or 8:1, make-up gain = integer dB steps). The compressor was applied to a single musical element within the context of a complete mix. For this test, the complete stimuli set from Level 1.2 was reduced to six different stimuli targeting a range of instruments: acoustic guitar, electric guitar, bass, drumset, piano, and snare drum. Each stimulus was repeated three times so that all possible ratios were used on each stimulus, for a total of 18 trials. The presentation order was randomized for each participant. The testing interface was identical to the interface used in the training program except the participants would not get feedback on their answers.

Table 8: List of stimuli used in the in-context listening test.

| Name | Genre | Target Musical Element | Number of Excerpts | Stimuli Length |
|---|---|---|---|---|
| “Adultery” | Indie Rock | Acoustic Guitar | 1 | 24 s |
| “She’s Like” | Pop/Folk | Drum Kit | 1 | 14 s |
| “Young Lover” | Pop/Folk | Electric Bass | 1 | 17 s |
| “Future Punch” | Jazz | Electric Guitar | 1 | 11 s |
| “A Vision Of You” | Jazz | Piano | 1 | 10 s |
| “Fever Nostalgic” | Electronic | Snare Drum | 1 | 10 s |

The three pre-tests (in-context, near-transfer, and far-transfer) were scheduled during a one-month period before the training began. The three post-tests were scheduled during a one-month period after the training terminated. The pool of recent graduates was scheduled to test during the training period. Participants were not scheduled for more than one test per day to minimize listening fatigue. All participants received a randomized order of the three listening tests. The task was explained to participants verbally by the experiment administrator at the beginning of each listening test. No guidance on how to listen to the DRC was given. The participants could complete 1-3 example pre-trials in the presence of the administrator to familiarize themselves with testing interface, listening task, and scale of audible differences between the stimuli as suggested by Bech & Zacharov [3].

### 5.2.3 Surveys

Before the first listening test all participants completed a survey to collect information about their relevant listening experience. The following data points were collected: years of previous musical training, years of previous musical experience, years of previous audio training, years of previous audio experience, hours spent participating in listening tests, years of TET, and age. “Training” was defined as years of experience with specific instruction or lessons.

"Experience" was defined as years participating in the activity. At the end of each listening test participants were asked to volunteer any comments they had.

### 5.2.4 Listening Room and Equipment

The listening tests were completed in the Critical Listening Lab at the Centre of Interdisciplinary Research in Music Media and Technology (CIRMMT). The lab is an ITU-R BS.775-1 compliant room and can be described as a high-quality listening environment [28]. Audio playback followed a path from Google Chrome browser running on a Mac -> RME MADI HDSPe audio card -> RME M-32 DA Converter -> Crookwood C10 7.1 monitor controller -> Classé CA5200 amplifier -> two B&W 802D loudspeakers. Participants were seated in a standard listening position and were free to adjust the listening level. Participants viewed the web-application on an apple monitor and interacted with a mouse and keyboard (interface details in Chapter 3).

Figure 16: A photo of the CIRMMT Critical Listening Lab. (Reproduced with permission from www.cirmmt.org)

## 5.3 Results

### 5.3.1 Participants

The demographic data for each listener group is summarized in Table 9. All three listener groups had statistically similar ($p > .05$) experience across all variables except for audio experience. The trained and untrained students had statistically similar levels of audio experience ($p > .05$) but the graduates had more experience than both groups of students ($p < .01$). This was a positive result since the untrained students were used as a control group for the trained students. The graduates were also a good comparison since they only differed in their additional audio experience, presumably from working in the audio field after graduating. On average, participants took 42.8 minutes to complete the listening test.

Table 9: Listener group demographic data.

| | Untrained Students Mean (Standard Deviation) | Trained Students Mean (Standard Deviation) | Graduates Mean (Standard Deviation) |
|---|---|---|---|
| **Music Training (years)** | 14.7 (6.2) | 13.8 (6.3) | 12.5 (3.1) |
| **Music Experience (years)** | 17.4 (3.9) | 11.9 (3.7) | 16.5 (5.7) |
| **Audio Training (years)** | 3.5 (1.9) | 4.0 (2.0) | 5.8 (3.0) |
| **Audio Experience (years)** | 3.5 (3.0) | 3.2 (1.9) | 10.4 (4.3) |
| **Technical Ear Training (years)** | 2.1 (2.4) | 2.1 (1.6) | 2.3 (1.4) |
| **Age (years)** | 25.5 (3.8) | 25.1 (2.3) | 29.9 (4.2) |

### 5.3.2 Primary Research Questions: Participant Improvement

#### 5.3.2.1 Number of Correct Responses

Figure 17 shows the differences in success rates for each student group for the pre- and post-test. The graduate group performance is included for reference.

A mixed-effects logistic regression model was fit to the data using the *listener group* and *pre-/post-test* variables to predict *correct response*. The *subject*, *stimuli*, *trial time of day*, and *block order number* variables were input as a random effect to account for overall differences in ability between participants, difficulty between stimuli, fatigue according to time of day, and experience gained from the block ordering of listening tests. The interaction effect between *listener* group and *pre-/post-test* was also included to examine any differences in improvement between the two groups. The graduates were excluded from this part of the analysis. This model was not a statistically significant fit to the data when compared to a baseline model ($x^2 = 6.18$, $p = .10$). Both groups of students performed similarly in both tests. In this case, the error bars overlapped for all measurements; the confidence in the measured differences in performance were very low. The error bars spanned around 16 % and the measured differences in performance were around 8 %. A greater difference in performance would have to be measured or a larger data set would need to be collected for there to be a higher degree of confidence. For example, train the participants longer or use more participants.

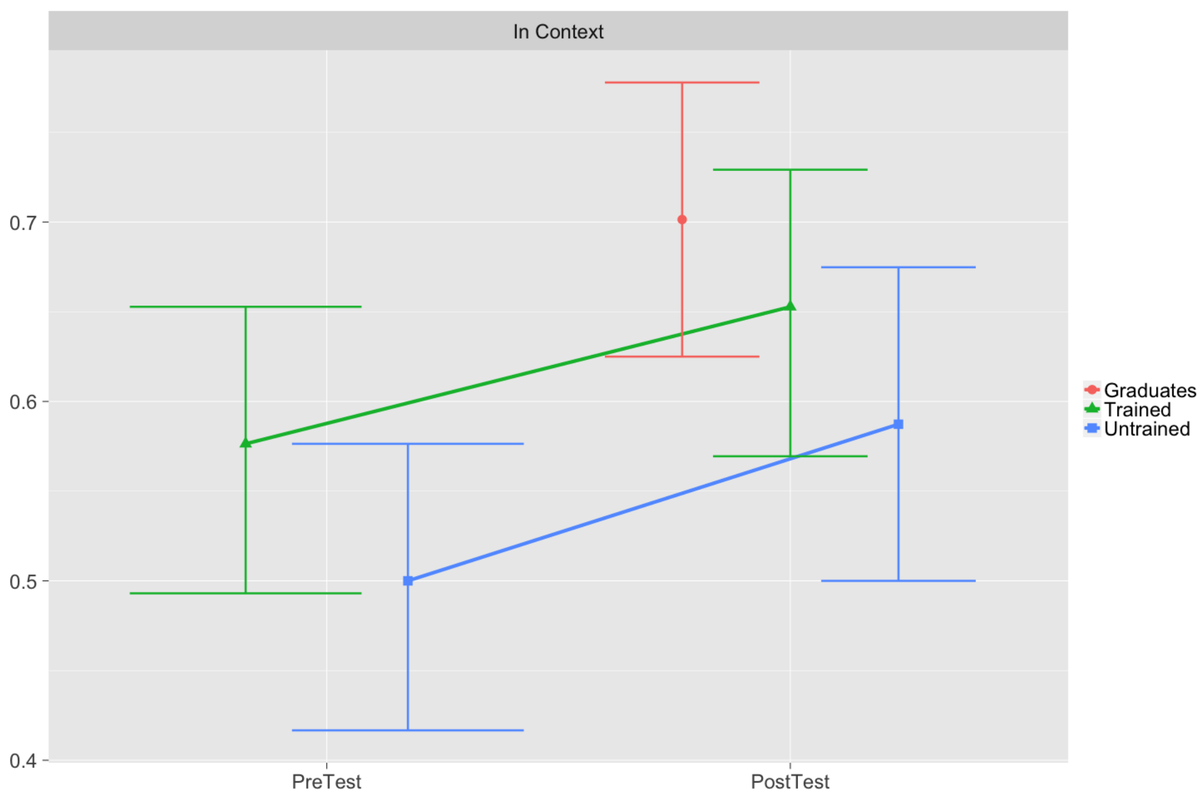


Figure 17: A line plot displaying the probability of a correct response as a function of the pre- and post-test by listener group.

It was hypothesized that the control group would improve. This improvement was expected to be a result of their ongoing education or familiarization with the listening test. It was also hypothesized that the trained group would improve to a greater degree. It was expected that this improvement would be partially due to their own ongoing education and familiarization with the listening test, like the control group, but that the *dyntet* training would lead to an even overall improvement. The results of this analysis did not support either of the hypotheses.

##### 5.3.2.1.1 Comparison to Graduates

The performance of the graduates compared to the students was evaluated using a mixed-effects logistic regression model with *listener group* predicting *correct response*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. This model was a statistically significant fit when compared to a baseline model ($x^2$ = 14.47, $p$ < .01). Multiple comparisons of means using Tukey contrasts revealed that differences only existed between the professionals (70 % correct) and the untrained students during the pre-test (50 % correct). It was expected that the graduates' performance would be higher than the untrained students but lower than the trained students (after training). This result would be similar to that of Quesnel. The results partially confirmed the hypothesis that the graduates would perform better than the students in the pre-test. However, this was only true for the untrained group.

#### 5.3.2.2 Time Taken to Complete a Trial

Figure 18 shows the differences in the trial time taken to arrive at a correct response for each group for the pre- and post-test.

A mixed-effects regression model was fit to the data using the *listener group* and *pre-/post-test* variables to predict *trial time*. The *subject* and *stimuli* variables were input as a random effect. Variables *trial time of day* and *block order number* could not be included as random effects since the high number of variables and low number of data points caused the model to fail to converge. The interaction effect between *listener* group and *pre-/post-test* was also included. The graduates were excluded from this part of the analysis. This model was not a statistically significant fit to the data when compared to a baseline model ($x^2$ = 4.01, $p$ = .26). This, along with the results of the previous chapter, confirmed the new hypothesis that the time taken to arrive at a correct response was not improved in this type of listening task.

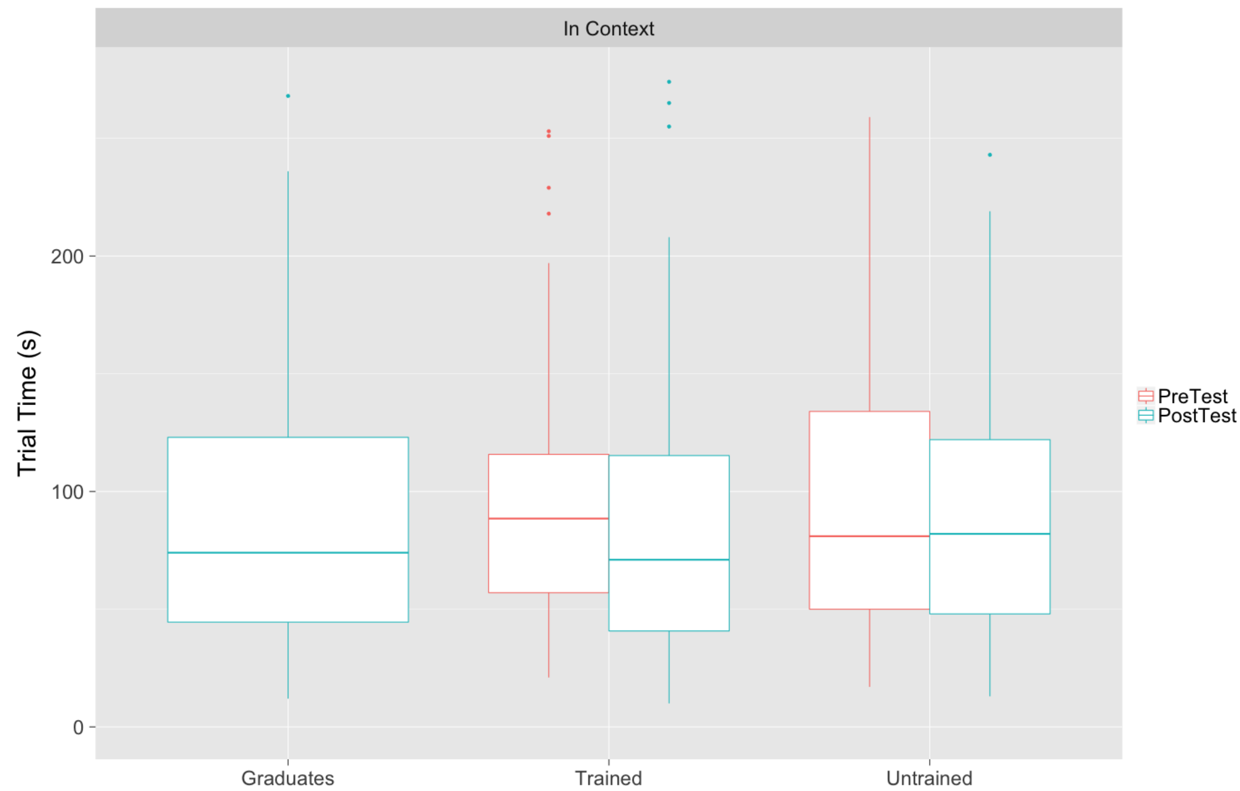


Figure 18: A box plot displaying the trial time taken to arrive at a correct response as a function of the pre- and post-test by listener group.

#### 5.3.2.2.1 Comparison to Graduates

Figure 18 also shows that the graduates seemed to take a similar amount of time per trial as the other listener groups. A mixed-effects regression model was fit to the data using *listener group* to predict *trial time*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. There were no differences between the graduates and student groups ($x^2 = 5.46$, $p = 0.24$), this provided additional support for the hypothesis.

### 5.3.2.3 Participant Comments

There were two participant comments that referred to differences in difficulty between the pre-test and post-test. From the trained group of students: "This time around, I knew what to focus on in order to hear compression…I felt confident going in there compared to the first test where I had no clue what to listen for other than the pumping." From the untrained group of students: "I will say that it felt easier to hear the differences, but I'm not sure." These comments supported the hypothesis that both student groups would perceive improvement from pre-test to post-test. There was no clear evidence of a difference in the degree of improvement between the groups.

### 5.3.3 Secondary Research Questions: Additional Factors

#### 5.3.3.1 Effect of Stimuli on Performance

Figure 19 shows the success rate for each stimulus. The results of the previous chapter suggested that there would be a difference in difficulty between the stimuli. A mixed-effects logistic regression model was built using *stimuli* to predict *correct response* with *subject number* input as a random effect. There was a significant difference in success rates between stimuli ($x^2$ = 12.14, $p < .05$). Multiple comparisons analysis using Tukey contrasts showed that the only significant pairwise difference was between the two extremes: acoustic guitar (48 %) and snare drum (67 %) ($p < .05$). This result supports the hypothesis.

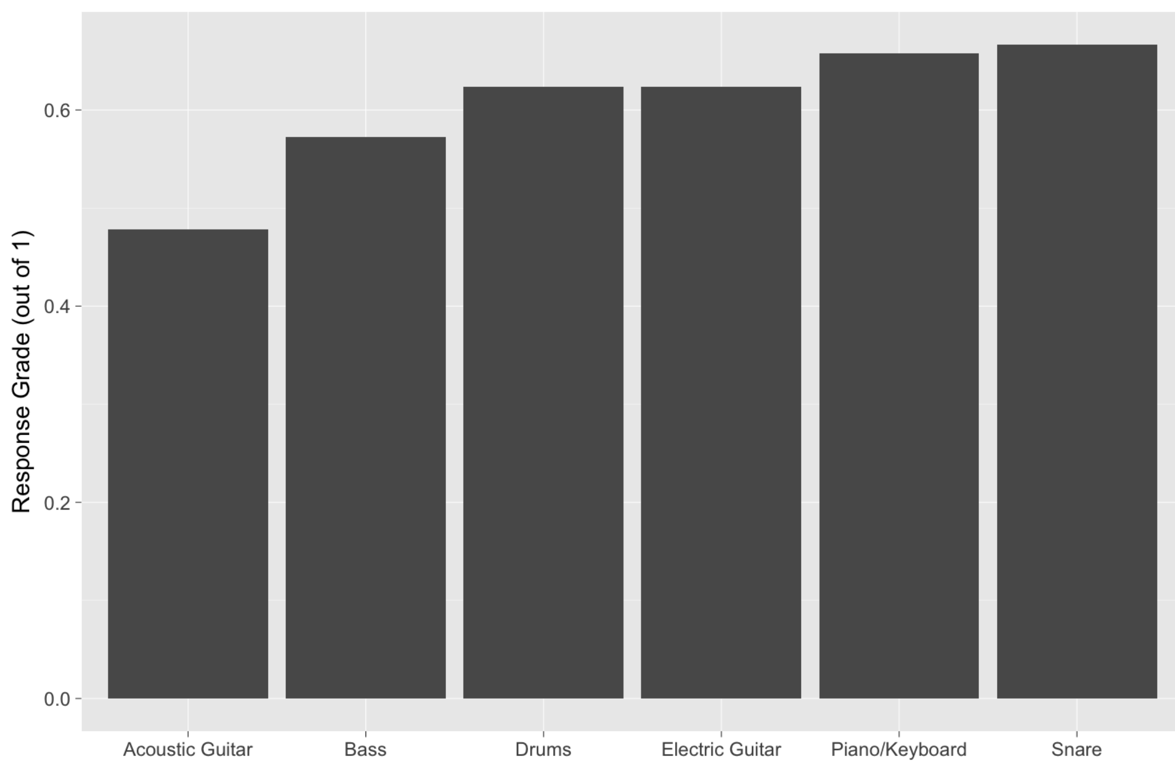


Figure 19: A bar plot displaying the probability of a correct response as a function of the stimulus type.

### 5.3.3.2 Effect of Time of Day on Performance

Figure 20 shows the participant success rate as a function of time of day. A statistically significant difference was not found in the previous chapter. For this listening test, the trends showed decreasing performance throughout the day, with a local maximum in the early afternoon. Several different types of curves were fit to the data with a mixed-effects logistic regression model using t*ime of day* to predict *correct response.* S*ubject number* and *PreTest or PostTest* were input as random effects. A statistically significant fit was not found. The best fit came from a general linear model depicting a downward trend dropping off 1.9 % per hour throughout the day ($x^2$ = 3.38, $p$ = .066). This model is visualized as the red line in Figure 20. Without a statistically significant fit, the hypothesis could not be confirmed.

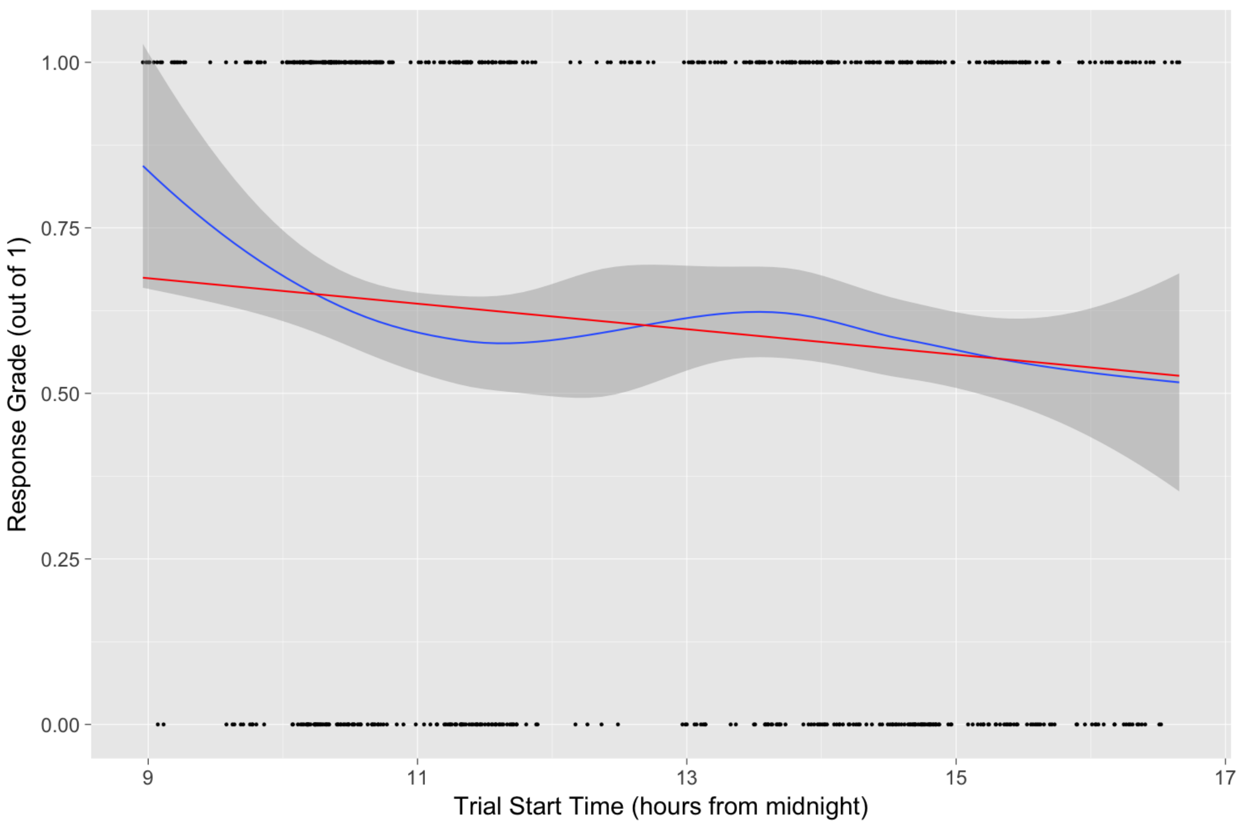


Figure 20: A scatter plot with a locally weighted regression model was fit to the data ("LOESS" model in R, ggplot2 [9]). The red curve is a linear regression model fit to the same data. The plot shows the change in the overall participant success rate as a function of time of day. The shaded area represents the standard error.

#### 5.3.3.3 Effect of Block Order on Performance

The three listening tests (in-context, near-transfer, far-transfer) were administered in a randomized order for each participant. It is possible a small amount of experience gained in one test could have affected the results in a subsequent test. However, it was hypothesized that this block ordering would have an insignificant effect. Indeed, there was no significant effect ($x^2$ = 2.31, $p$ = .31).

#### 5.3.3.4 Participant Comments

The participant comments that did not discuss elements of learning or improvement could be divided into roughly three categories: 1) comments on varying difficulty, 2) comments on listening strategy, 3) comments on the listening test. Comments were converted to short form and can be seen in Table 10. There were no comments about listener fatigue.

Table 10: Participant comments divided up into three categories. The number beside each comment indicates the number of times that comment was recorded.

| Comments on varying difficulty | Comments on listening strategy | Comments on the listening test |
|---|---|---|
| Acoustic guitar was difficult (4) | Listening to changes in bleed from hi-hat in snare drum (2) | Liked the quizzing format (1) |
| 8:1 Ratio was easy (3) | Listening to changes in background reverberation (1) | Interface simple and intuitive (1) |
| Piano was easy (1) | Listening strategy changed depending on stimuli (1) | Test was too long (1) |
| Snare drum was easy (1) | | Awesome to practice with (1) |
| Drums were difficult (1) | | |
| Bass was difficult (1) | | |

Overall, the participants agreed that the acoustic guitar stimulus was hard. This correlates with the results. Listening strategies also appeared to be stimuli specific. In three cases, participants reported hearing changes in background information in the signal. Participants also appeared to appreciate the format and set-up of the test. One participant found the test too long.

## 5.4 Discussion

The results of the analysis showed that there was no significant improvement in the number of correct responses for either student group. It is possible that the hypothesis was wrong and that no learning took place. It is also possible that for the small number of data points collected (576 trials divided between four main conditions) that there was too much uncertainty and error variance in the model to make a reliable measurement. Statistical power is limited when examining a categorical variable (correct or not) compared to examining an interval variable like response time. More data points should have been collected to obtain more certainty. Perhaps, the listening tests could have been administered a second time to each participant since recruiting additional students was not possible. Alternatively, more variables could have been collected to explain more of the error variance. The influence of unmeasured and uncontrolled variables was visible when individual student performance was examined (Figure 21). It was expected that students would improve at different rates, but several students scored far lower on the post-test than the pre-test. Random chance will always be a factor in the variance, and it is possible that some students became far worse at the task over the training period. However, the latter was unlikely given the results of the previous chapter. Instead, it was hypothesized that there were some additional key factors involved. With additional information on the schedules of the students, it was hypothesized that overall fatigue levels played a much larger role in this dataset than expected. There was also precedence for this when looking at the larger field of standardized testing [90]. While evidence was not found in the written participant comments, several students made verbal comments about fatigue levels to the experiment administrator. As part of their studies, these students frequently conducted overnight recording sessions followed by classes the next morning. Some reported getting less than recommended amounts of sleep. Each day of the week sometimes represented a completely different schedule than the previous, so fatigue levels were difficult to predict without precise schedule details. In future studies using pre-/post-testing, a much greater effort will need to be made to control for differing levels of fatigue or other confounding variables. Little variance was seen in participant success rates according to time of day, so additional types of measurements will have to be explored.

To summarize, there was no concrete evidence that the trained group of listeners improved at a higher rate than the untrained listeners. The pre-/post-testing methodology of this study suffered from a small number of categorical data points and a large variation in participant scores due to unexplained variables.

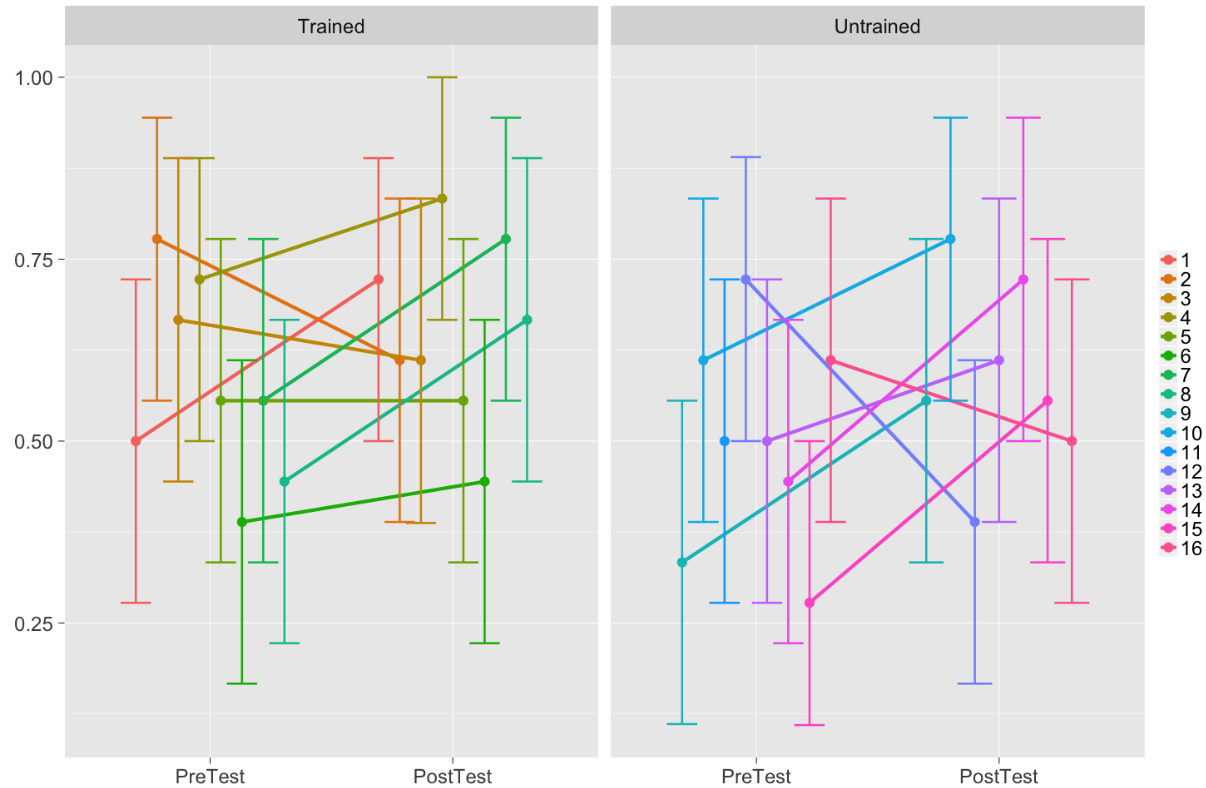


Figure 21: A line graph showing the percentage of correct responses for each student for both the pre-test and post-test.

# 6 Testing Training Effect III: Near-Transfer Listening Test

## 6.1 Introduction

In the previous chapter, the in-context listening test was used to measure improvement within the training environment. In this chapter, a near-transfer listening test was used to measure improvement in a nearly related environment. In the ATET classes and post-training user survey (Chapter 9) the trained group of students reported that they "learned the stimuli". Students reported identifying unique aspects of each sound to which they paid attention with repeated practice. The goal of the *dyntet* training program was to teach skills that could be applied to new situations, so a measurement of the students' ability to transfer their learned skills to a different stimuli set was desired. Therefore, the near-transfer listening test was designed to be exactly the same as the in-context listening test – the only difference was the use of unfamiliar stimuli.

In Chapter 4, the training data was analyzed and the measured improvement could have been due to any of the five stated learning factors (Chapter 4.1 and Chapter 5.1). In the previous chapter, the in-context listening test attempted to eliminate the influence of Factor 4 with a control group (increased listening skill level due to their ongoing education). The same control group was included in this near-transfer test. This near-transfer test also attempted to eliminate the influence of Factor 2 (familiarization with the stimuli) by using an unfamiliar stimulus set.

The main research question in this chapter was will the trained students be able to transfer their learned skills to unfamiliar stimuli? Otherwise, all other research questions were the same as the previous chapter. Will the trained group improve? Will the untrained group improve? How do these groups compare to the graduates? What will be the effects of stimuli, time of day, and block order on the success rate?

## 6.2 Methodology

Except for the stimuli, the methodology for this listening test was identical to the in-context listening test (Chapter 5.2). The same three groups of participants were used, the listening task was Level 1.2 of the *dyntet* training application, and the same number of trials were completed. While the stimuli were different, the instrument types were kept consistent with the in-context listening test: acoustic guitar, electric guitar, bass, drumset, piano, and snare drum. Participants were given verbal directions on the task and interface at the beginning of the listening test and were allowed one or two practice trials. A survey was given at the end of the test to collect participant comments. The test took place in the same Critical Listening Lab at CIRMMT.

Table 11: List of stimuli used in the near-transfer listening test.

| Name | Genre | Target Musical Element | Number of Excerpts | Stimuli Length |
|---|---|---|---|---|
| "Theo" | Indie Rock | Acoustic Guitar | 1 | 9 s |
| "Clarity" | Indie Rock | Drum Kit | 1 | 15 s |
| "Booblast" | R&B/Rap | Electric Bass | 1 | 14 s |
| "Dirty Pictures" | Indie Rock | Electric Guitar | 1 | 10 s |
| "Laisse tomber la cause" | R&B | Piano | 1 | 11 s |
| "Mwen vlé sav" | R&B/Jazz | Snare Drum | 1 | 16 s |

## 6.3 Results

On average, participants took 38.1 minutes to complete the listening test, slightly less than the in-context listening test.

### 6.3.1 Primary Research Questions: Participant Improvement

#### 6.3.1.1 Number of Correct Responses

Figure 22 shows the success rates for each group for the pre- and post-test. A mixed-effects logistic regression model was fit to the data using the *listener group* and *pre-/post-test* variables to predict *correct response*. The *subject*, *stimuli*, *trial time of day*, and *block order number* variables were input as a random effect. The interaction effect between *listener* group and *pre-/post-test* was also included. The graduates were excluded from this part of the analysis. Both groups of students performed statistically the same in both tests ($x^2 = 3.22$, $p = .36$). Similar to the in-context listening test, the standard error bars overlap for all measurements. Before obtaining the results of the previous chapter, it was hypothesized that the trained students would be able to transfer their learned skills to a new set of stimuli, but that the measured improvement would be less than the measured improvement in-context. It was also expected that the untrained group will improve slightly, but less than the trained group. After obtaining the results of the previous chapter, it was expected that this listening test would suffer from the same lack of data points and unexplained variance. Overall, this result did not support the original hypothesis that both student groups would perform better in the post test, and that the trained group would experience a higher level of improvement.

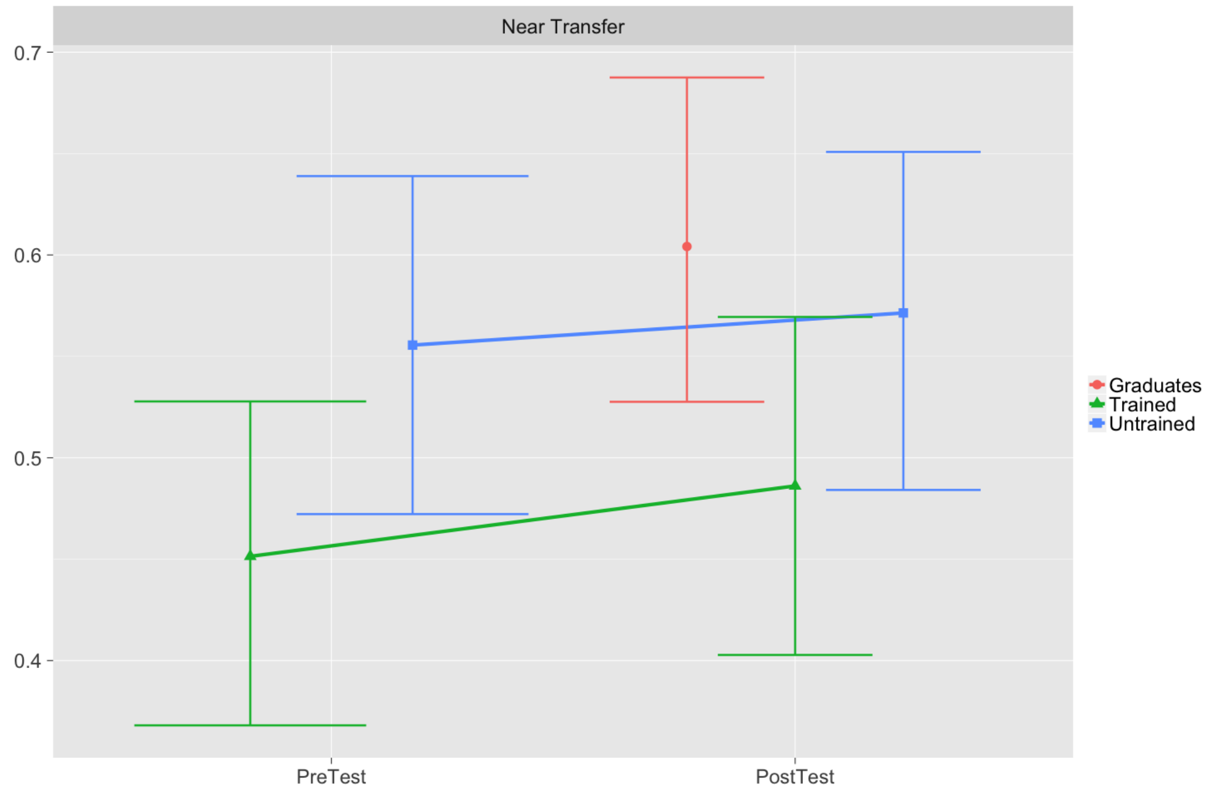


Figure 22: A line plot displaying the probability of a correct response as a function of the pre- and post-test by listener group.

##### 6.3.1.1.1 Comparison to Graduates

To examine the performance of the graduates compared to the students a mixed-effects logistic regression model was fit to the data using *listener group* to predict *correct response*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. Differences did not exist between the professional graduates' performance the students' ($x^2 = 8.95$, $p = .062$). It was hypothesized the graduates' success rate would be higher than the untrained group for both tests, but worse than the trained group in the post-test. The limited accuracy of the model did not allow the examination of that relationship.

### 6.3.1.2 Time Taken to Complete a Trial

Figure 23 shows the trial time to arrive at a correct response for each group for the pre- and post-test. A mixed-effects regression model was fit to the data using the *listener group* and *pre-/post-test* variables to predict *trial time*. The *subject*, *stimuli*, *trial time of day*, and *block order number* variables were input as random effects. The interaction effect between *listener* group and *pre-/post-test* was also included. The graduates were excluded from this part of the analysis. This model was not a statistically significant fit to the data when compared to a baseline model ($x^2$ = 2.24, $p$ = .52). This result supported the new hypothesis that there would be no differences in the time taken to arrive at a correct response in either group for either test.

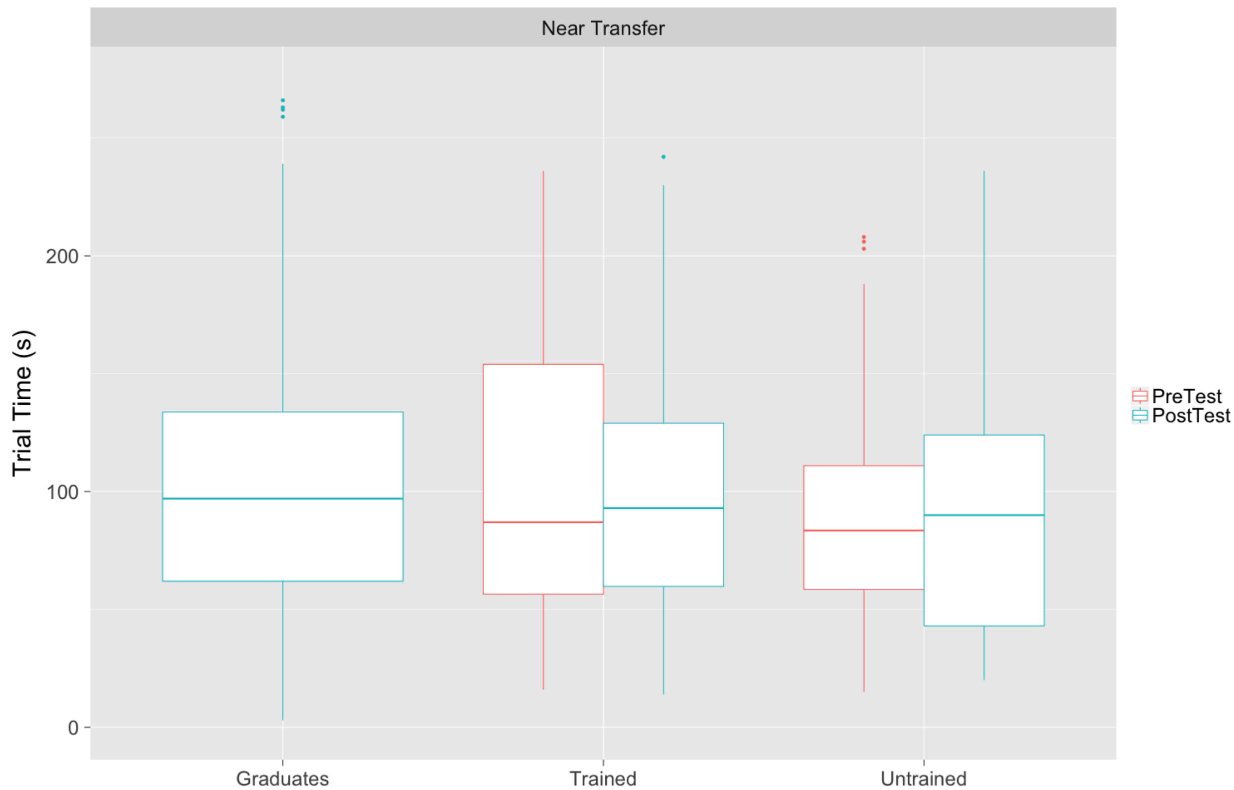


Figure 23: A box plot displaying the trial time taken to arrive at a correct response as a function of the pre- and post-test by listener group.

##### 6.3.1.2.1 Comparison to Graduates

Figure 23 also shows that the graduates seemed to take a similar amount of time per trial as the other listener groups. A mixed-effects regression model was fit to the data using *listener group* to predict *trial time*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. There was no difference between the graduates and student groups ($x^2$ = 5.67, $p$ = 0.23), this supports the hypothesis.

#### 6.3.1.3 Participant Comments

Three students in the trained group and two students in the untrained group indicated improvement between the pre-test and post-test. The trained group said, "…more confident in knowing what I'm trying to listen for than at the beginning of the year", "I could tell the difference more than the last time I took the test", and "after the training, I had an idea of the artifact to listen for instead of being completely left in the dark…". The untrained group said, "easier than last time" and "I think the second time it may have gone better". These comments provided support for the hypothesis that the students would improve from the pre-test to the post-test.

### 6.3.2 Secondary Research Questions: Additional Factors

#### 6.3.2.1 Effect of Stimuli on Performance

Figure 24 shows the differences in the success rate for each stimulus. A mixed-effects logistic regression model was built using *stimuli* to predict *correct response* with *subject number* input as a random effect. There was a significant difference in success rates between the stimuli ($x^2$ = 25.4, $p < .001$). A multiple comparisons analysis using Tukey contrasts showed that significant pairwise differences exist: bass harder than drums ($p < .05$), bass harder than electric guitar ($p < .001$), and snare harder than electric guitar ($p < .01$). This result supported the hypothesis of varying difficulty between stimuli.

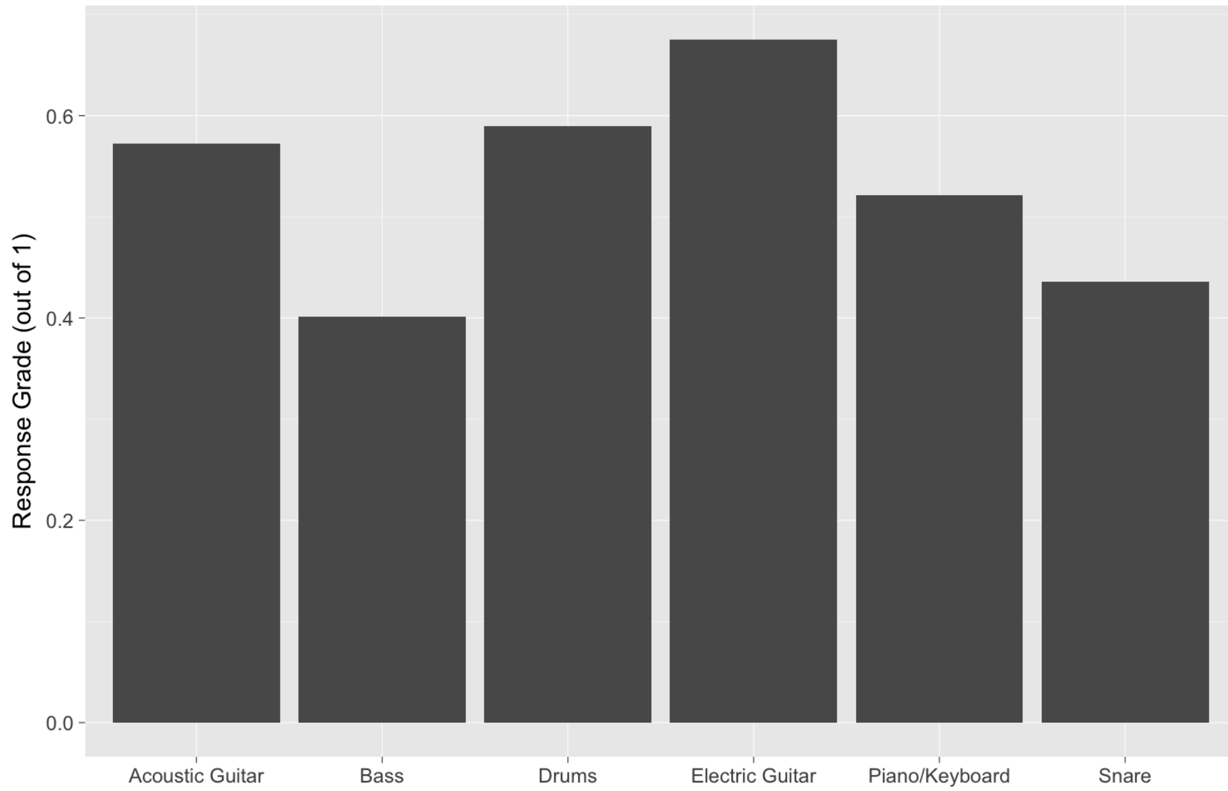


Figure 24: A bar plot displaying the probability of a correct response as a function of the stimulus type.

#### 6.3.2.2 Effect of Time of Day on Performance

Figure 25 shows the participant success rate as a function of the time of day the listening test was completed. The trend was an overall parabolic curve throughout the day with a global minimum and local maximum in the early afternoon. Several different types of curves were fit to the data with a mixed-effects logistic regression model using t*ime of day* to predict *correct response.* S*ubject number, PreTest or PostTest,* and *listening test block order* were input as random effects. A statistically significant fit was found using a second order parabolic curve ($x^2 = 6.99$, $p < .05$); performance decreased until midday at which point it started to improve again. This model is depicted as the red curve in Figure 25. A statistically significant fit was not found when also including the small local maximum at 13h00 in a more complex model. When combining the results from Chapter 5 an overall trend emerged. Performance was high in the morning and decreased throughout the day with a possible small local maximum in the afternoon.

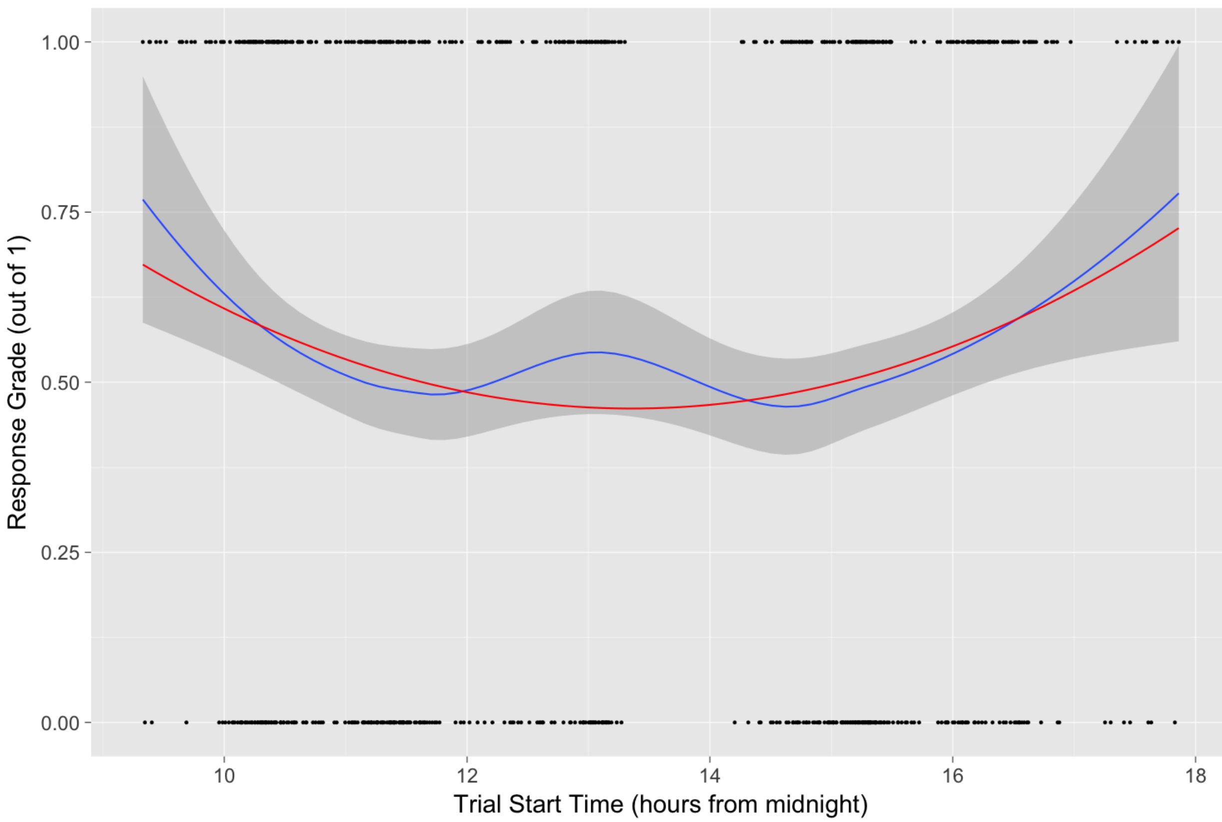


Figure 25: A scatter plot with a locally weighted regression model was fit to the data ("LOESS" model in R, ggplot2 [9]). The red curve is a second order polynomial regression model fit to the same data. The plot shows the change in the overall participant success rate as a function of time of day. The shaded area represents the standard error.

### 6.3.2.3 Effect of Block Order on Performance

The randomized order in which participants took the listening tests (in-context, near-transfer, far-transfer) had no significant effect on participant performance ($x^2 = 4.29$, $p = .12$).

### 6.3.2.4 Participant Comments

The participant comments that did not discuss elements of learning or improvement were divided up into three categories: 1) comments on varying difficulty, 2) comments on listening strategy, 3) comments on the listening test. Comments were converted to short form and can be seen in Table 12. One student indicated a high level of fatigue in their comments for one test.

Table 12: Participant comments divided up into three categories. The number beside each comment indicates the number of times that comment was recorded.

| Comments on varying difficulty | Comments on listening strategy | Comments on the listening test |
|---|---|---|
| Bass difficult. (3) | I listened for timbral changes. | Was difficult. (10) |
| Snare is difficult. | Sustained notes made compression easier to hear. | Test was long. (2) |
| Piano difficult. | | Looping shorter excerpts would be helpful. (2) |
| It gets easier the more times you hear the same stimuli. | | Good training. |
| Acoustic guitar difficult. | | |
| All instruments were equally difficult. | | |
| Bass and guitar easy. | | |

Overall, the impression of the participants that the bass sample was difficult correlated with the results. Listening strategies were slightly different for this test. Participants discussed timbral changes and that sustained notes in the target instrument made the task easier. Most participants indicated that this test was difficult. This agrees with the lower overall scores compared to the in-context listening test. More participants complained about the length of this test, even though it took a similar amount of time to complete. Finally, two participants indicated that shorter musical loops would have made the task easier.

## 6.4 Discussion

The results of this listening test were very similar to the previous in-context listening test. No improvement was measured in the number of correct responses for either student group. This gave no support for the main hypothesis that the trained listeners would experience a larger improvement than the untrained listeners. Although, some support of the hypothesis was found in the participant comments. This study also suffered from the same problems of statistical power. Additional evidence of fatigue playing a role in unexplained variance was visible. Participant number two made the comment “Tired, rough morning” after the pre-test. Figure 26 displays the success rates for both the pre-test and post-test for each participant individually. The largest difference in scores between tests came from participant number two. Indeed, the fatigued pre-test score was much lower. In addition, there was no clear pattern of student increase or decrease in performance when examining both the in-context and near-transfer listening tests together. It was clear that unexplained variables were playing a large role in the results. The analysis of this listening test also showed no differences in time taken to arrive at a correct response. Taken with all the previous results, it's clear that this was not a useful performance metric for this listening task since it did not change with practice.

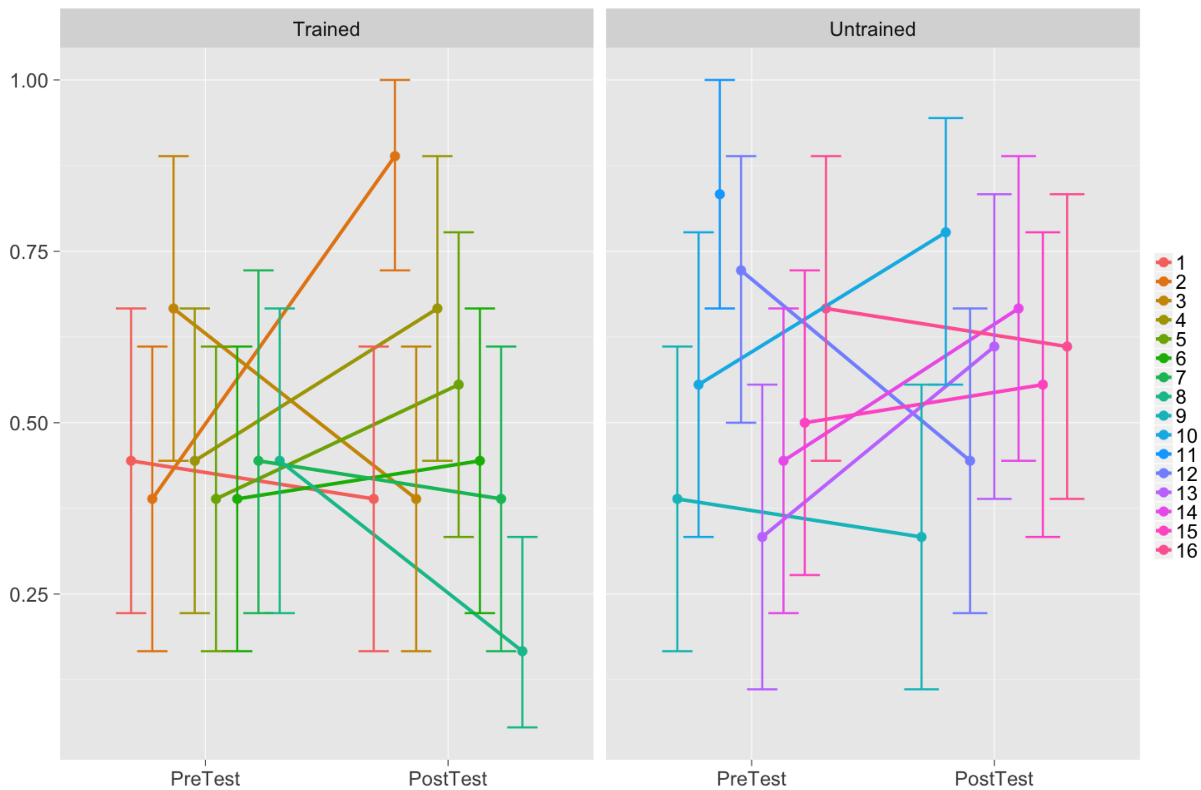


Figure 26: A line graph showing the percentage of correct responses for each student for both the pre-test and post-test.

There was additional support for the hypothesis that difficulty varies between stimuli. No pattern emerged of particularly difficult instrument types. For example, the snare drum was relatively easy in the in-context listening test and relatively hard in the near-transfer listening test. It is possible that instrument types are not inherently easy or difficult, instead, the nature of the stimulus' dynamic range, ADSR envelope, and musical context might play a role.

The time of day variable was proving to be a significant predictor of participant performance. Under the assumption that the time of day was related to participant fatigue levels, this was additional support that listener fatigue was an important variable.

To summarize, there was no concrete evidence that the trained group of listeners improved at a higher rate than the untrained listeners with unfamiliar stimuli. The pre-/post-testing methodology of this study suffered from a small number of categorical data points and large variations in participant scores due to unexplained variables. There was additional evidence that listener fatigue was a factor.

# 7 Testing Training Effect IV: Far-Transfer Listening Test

## 7.1 Introduction

The in-context listening test (Chapter 5) was designed to measure participant improvement within the training context. The near-transfer listening test (Chapter 6) examined participants' ability to transfer learned skills to new stimuli. Finally, this far-transfer listening test was designed to examine the participants' ability to transfer learned skills to a more distantly related task. Many researchers have documented an increase in participant scores due to training (Chapter 2.4.2). However, less research has addressed the application of these learned skills to ecologically valid listening tasks. When presenting this type of research to the audio community, a typical research question is "does TET actually make you a better audio engineer?" This far-transfer listening test resembled a real-life mixing environment: setting up the parameters of a compressor on a vocal to taste. Two unrealistic factors were introduced to reduce bias: the compressor parameters were controlled by endless rotary encoders and there was no visual feedback. The participants had to make their adjustments by ear only. Very few stimuli were used, and participants were asked to apply their preferred settings to the same stimulus multiple times. Since performance was difficult to measure directly in this task, it was measured indirectly by looking at consistency. Participant performance was measured as the ability to repeatedly set-up the compressor with the same parameters on the same stimuli. To be successful, participants would need to be able to hear the effects of the compressor with detail and establish a stable preference. The assumption was that a skilled listener could form a consistent preference immediately and repeatedly set a compressor accurately by ear without visual references. It was hypothesized that this skillset informs on experience, maturity, and a mastery of DRCs. These are all aspects that successful audio engineers should possess, and may come closer to answering the question, "does TET actually make you a better audio engineer?"

Returning to the stated learning factors, the in-context listening test attempted to eliminate the influence of Factor 4 with a control group. The same control group was included in this test. The near-transfer test attempted to eliminate the influence of Factor 2 by using different stimuli. The far-transfer listening test used unfamiliar stimuli as well. Finally, the far-transfer test also attempted to eliminate Factors 1 and 3 by using a different listening test and compressor design (#1 familiarization with the testing application, #3 familiarization with the compressor algorithm design). If improvement in the trained students was greater than improvement in the untrained students, it was most likely due to Factor 5: increased listening skill level due to the training application. This far-transfer listening test represented the culmination of this study. It isolated the learning factor of interest and tested students in an environment that informed on their abilities as an engineer.

There were several research questions in this study. Will both student groups improve from the pre-test to post-test? Will the trained students improve more than the untrained students? How do the student groups compare to the graduate professionals? Do the participants' comments show any signs of self-reported improvement? Will the training cause a change in the students' preferences for compressor settings? For example, do the students use less compression overall after the training? How do the student's preferences compare to the graduates? Is participant performance affected by any secondary factors (stimuli, time of day, block order of listening tests)?

## 7.2 Methodology

The methodology for this listening test was quite different than the two previous tests. However, several aspects remain the same. The exact same listener groups were used: trained students, untrained students, and professional graduates. Participants were given verbal directions on the task and interface at the beginning of the listening session and were allowed one or two practice trials before starting. A survey was given at the end of the test to collect participant comments. The test took place in the same Critical Listening Lab at CIRMMT.

### 7.2.1 Listening Task

The listening task used in this experiment was a method of adjustment preference task. Participants were asked to adjust the parameters of a DRC to taste. The DRC was applied to a vocal and a backing track was provided. Participants were not instructed to be as consistent as possible trial to trail, just to apply their most preferred settings each trial. Precisely, participants were instructed to apply the settings "that brought the audio as close as possible to what they believed was ideal, given the parameters available to them". To simplify the task, participants were only given access to the compressor's ratio and make-up gain controls. All other parameters were custom preset for each stimulus by a panel of three professional audio engineers. The numeric parameter settings of the compressor were hidden from the participant and adjustments were made via endless rotary encoders. Therefore, participants were forced to set the compressor parameters by ear. This eliminated visual bias and prevented participants from simply repeating their previous settings by memory. The compressor settings had end-points (minimums and maximums), but the rotary encoders did not. When the participant hit an end point they could keep spinning the rotary encoder but the sound would not change. This eliminated the possibility of users setting the compressor by feel/touch instead of by ear. This functionality was explained to the participants to avoid confusion.

Each trial began with the ratio set to the minimum (1.1:1) and the make-up gain set to 0 dB. Participants could take as long as they wanted to arrive at their preferred response and set the overall listening level according to their preference. The interface (Figure 27) allowed participants to switch between having the compressor engaged ("Process"), disengaged ("Bypass"), or mute the audio. When participants were happy with their settings, they clicked "Submit". When they were ready for the subsequent trial, they clicked "Next!".

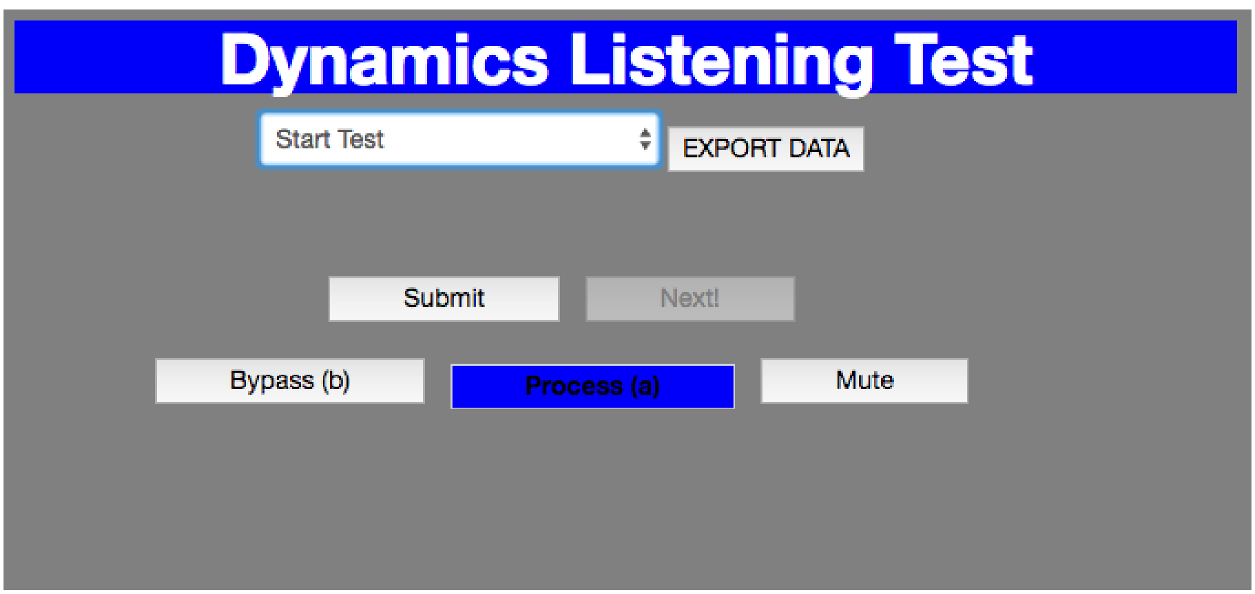


Figure 27: The far-transfer listening test user interface.

### 7.2.2 Compressor Design and Settings

Another goal of the far-transfer listening test was to present participants with a new compressor design that was not used in the training program. The chosen compressor featured a rotation point control instead of the typical threshold. The rotation point represents the single input level where 0 dB of gain is applied to the signal. Decreasing inputs below the rotation point have increasing positive gain applied to them. Increasing inputs above the rotation point have decreasing negative gain applied to them. Several input/output curves are plotted in Figure 28 that demonstrate a rotation point with varying ratios.

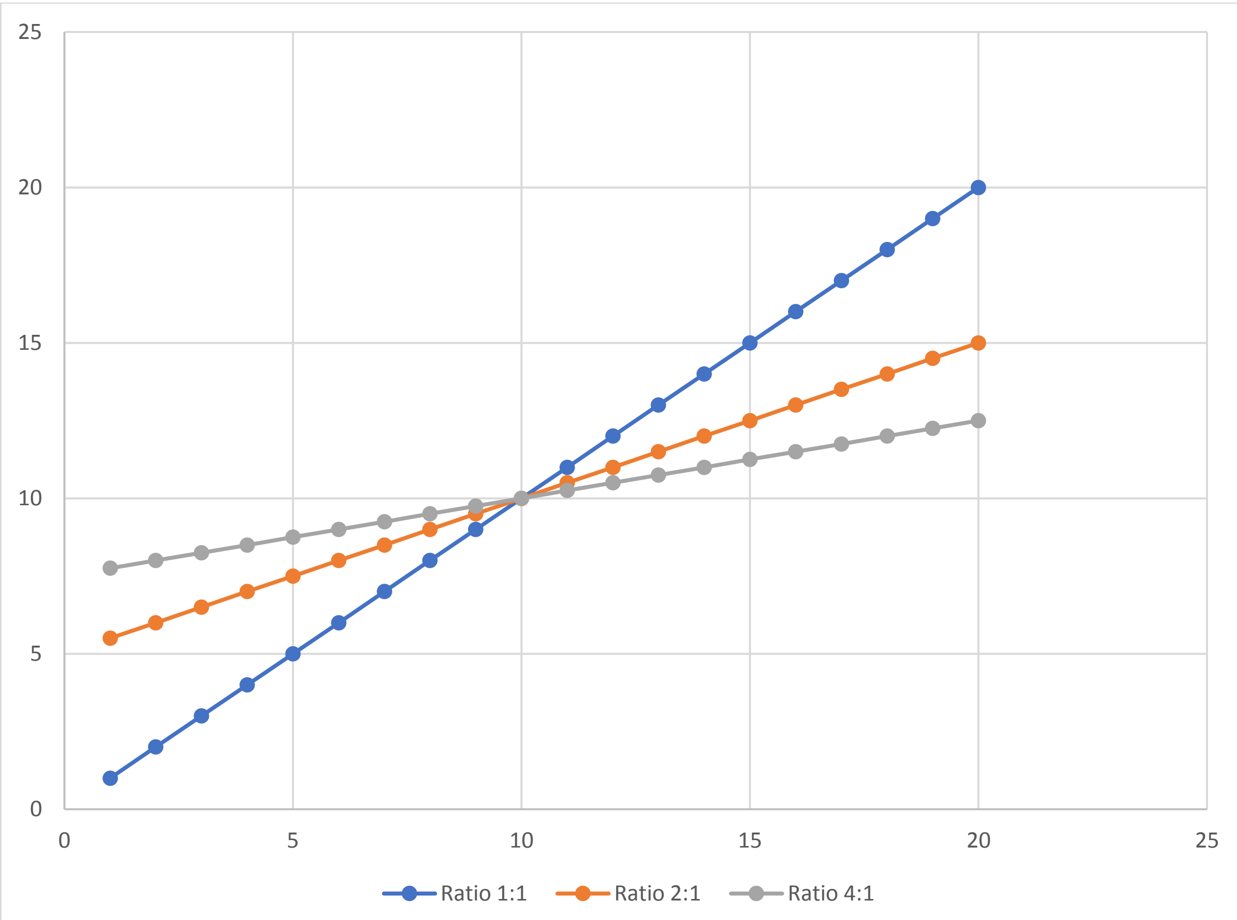


Figure 28: An input/output graph for a rotation point compressor at a variety of ratios. The rotation point in this case is 10 dB.

In this case, the threshold is defined as the point under which the gain is constant, not the point at which the gain is 0 dB. Figure 29 shows several input/output curves demonstrating a threshold at various ratios.

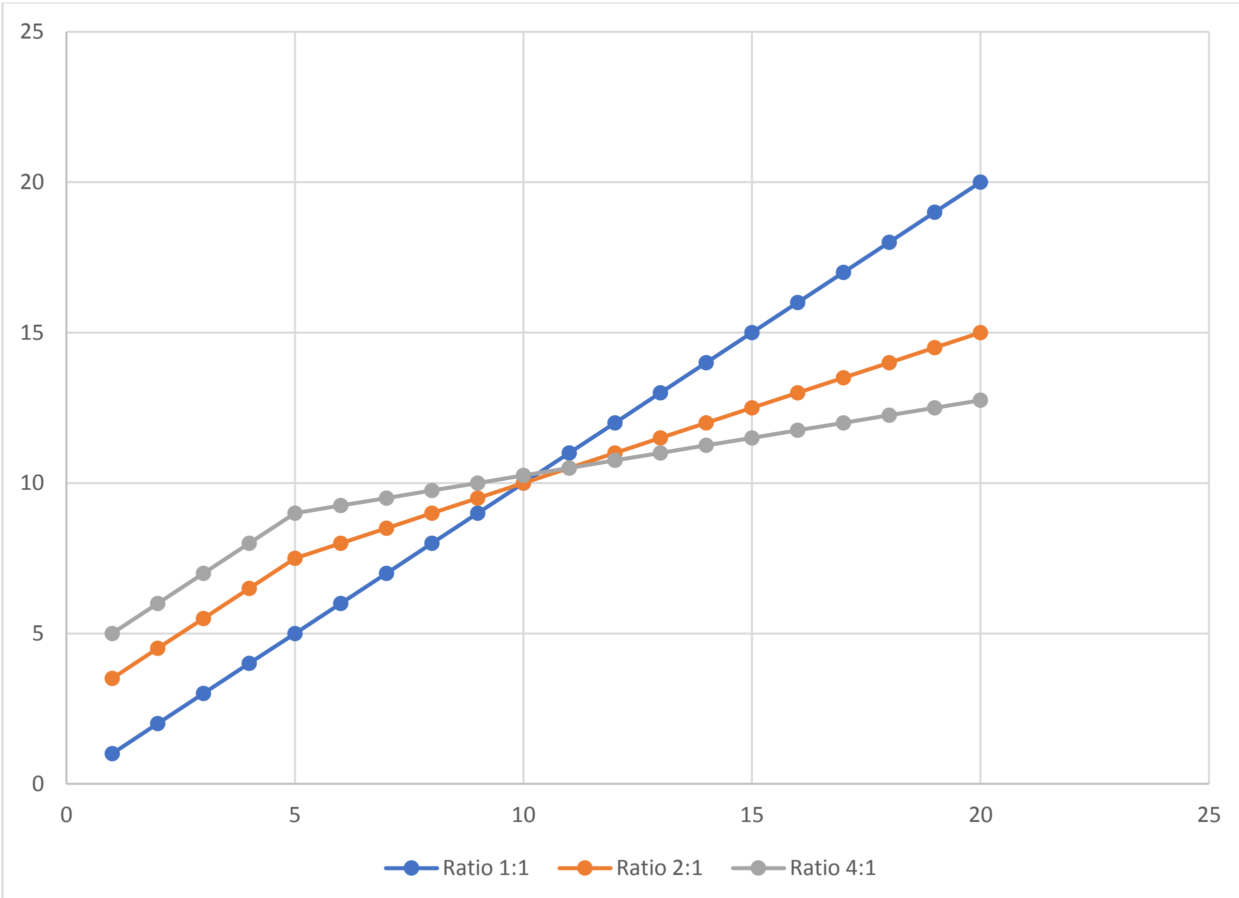


Figure 29: An input/output graph for a rotation point compressor at a variety of ratios. The rotation point in this case is 10 dB, the threshold is 5 dB.

The goal of this design was to facilitate an automatic make-up gain when the ratio of the compressor was changed. This only works under a specific set of circumstances. The following describes these circumstances as they applied to the first stimulus used in this study. The vocal track's dynamic range was subjectively too large, and a DRC was used to reduce it. The vocal's performance existed within a 40 dB range, with the quietest note at -40 dBFS (decibels full scale), and the loudest note at 0 dBFS. The goal was to take that entire range and reduce it equally throughout. The rotation point was set near the middle of the dynamic range, -20 dBFS. The threshold was set such that the entire vocal performance sat within the compressors active range, at -40 dBFS, so that the 1:1 range began just below the vocal's quietest note. With this set-up, the ratio could be adjusted on its own to affect changes to the vocal's dynamic range. The vocal's average level stayed the same and the make-up gain did not need to be continuously adjusted. This set-up is visualized at various ratios in Figure 30.

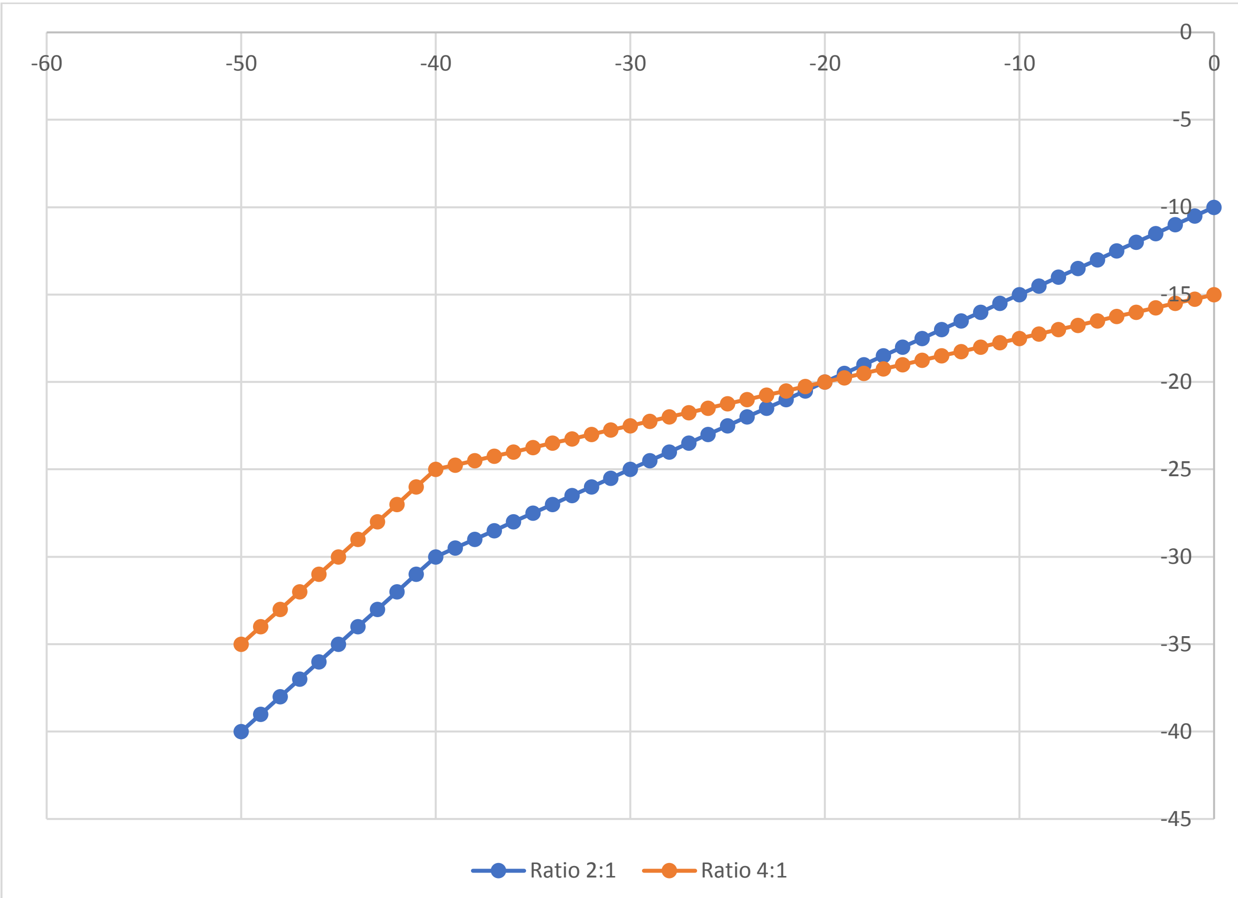


Figure 30: A rotation point compressor at a variety of ratios. Rotation point is -20 dB and threshold is -40 dB.

This type of compressor set-up is not common in the audio engineering field. Instead of explaining the unfamiliar inner workings of the compressor design to the participants in the study, the compressor was simply described as having "automatic make-up gain" so that "you could adjust the ratio without having to constantly adjust the make-up gain as well" and that "as you move the ratio, the average vocal level should stay the same while the dynamic range gets larger and smaller." Participants were also instructed that the make-up gain would then play a different role; "you can think of it as a fader, to sit or place the vocal in the mix at the overall level that you prefer". With all of this in the mind, the simplest explanation for the controls given to the participants was that the ratio controlled the dynamic range and the make-up gain controlled the overall level. The attack and release controls were implemented in the same manner as the compressor in the training program.

When mapping the ratio and make-up gain controls to the rotary controller the parameters were given end-points and were discretized. The make-up gain had a minimum value of -20 dB and maximum value of +20 dB. The range was discretized into 0.5 dB steps. The ratio had a minimum value of 1.1:1 and maximum value of 20:1. The ratio range was logarithmically discretized and mapped to 31 steps. As discussed in Chapter 3, a logarithmic mapping was used because it was hypothesized the reduction in dynamic range would be perceived logarithmically. For example, a perceived change in dynamic range from a ratio of 2:1 to 3:1 would be much larger than the perceived difference between a ratio of 19:1 and 20:1. The goal was to create perceptually equal changes in dynamic range between each ratio step. The exact mapping used can be seen in Table 13.

Table 13: Mapping of the ratio parameter to the steps in the rotary encoder.

| **Steps** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ratio (x:1)** | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.8 | 2 | 2.2 | 2.4 | 2.6 | 2.9 | 3.2 | 3.5 | 3.9 |
| **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** | **31** |
| 4.3 | 4.7 | 5.2 | 5.7 | 6.3 | 6.9 | 7.6 | 8.4 | 9.2 | 10.1 | 12 | 13.2 | 14.5 | 16 | 18 | 20 |

### 7.2.3 Stimuli

Three different stimuli were used for the listening test. Each stimulus was presented six times for a total of 18 trials. All stimuli featured a lead vocal within the context of a backing track. Summary information is available in Table 14. The stimuli spanned three genres, R&B, music theatre, and indie pop. All audio came from professional mix sessions at a sample rate of 96 kHz and bit depth of 24 bit. Musically acceptable loops were created ranging from 13-25 s. As with the stimuli used in the training program, access to the professional mix sessions was needed to create the vocal stem without compression. For more details on stimuli creation, see Chapter 3.3.

Another requirement of the stimuli was for the vocal to have an inappropriately large recorded dynamic range. With this, the participants would be more or less obliged to apply compression. The dynamic range was calculated for the vocals and backing tracks using loudness units according the ITU-R BS.1770-4 standard [30]. The backing tracks had a small loudness range of 2 LU. All vocals had a larger loudness range, ranging from 6 LU to 19 LU, and therefore would likely need DRC to "fit" into the track.

Table 14: List of stimuli used in the far-transfer listening test.

| Name | Genre | Target Musical Element | Vocal/Backing Track Loudness Range (LU) | Stimuli Length |
|---|---|---|---|---|
| "Marj" | Indie Pop | Lead Vocal | 6 LU/2 LU | 13 s |
| "Universe" | R&B | Lead Vocal | 19 LU/2 LU | 25 s |
| "Leave" | Musical Theatre | Lead Vocal | 10 LU/2 LU | 13 s |

## 7.3 Results

It was hypothesized that changes to the ratio setting on a DRC are perceived logarithmically. The ratio setting was discretized and assigned to ratio steps to create perceptually equal "chunks". Figure 31 and Table 15 summarize the ratio step response data and the corresponding true ratio data. Both the true ratio and logarithmic ratio steps demonstrated a right skew, but far less with the ratio steps. In addition, the three central tendency measures (mean, trimmed mean, and median) converged more closely when using the ratio steps. For statistical testing a normal distribution is preferred, therefore, the following analysis made use of the ratio steps and not the corresponding true ratio. There were no changes in the rejection or acceptance of the null hypotheses when using the true ratio instead of the ratio steps and robust statistics.

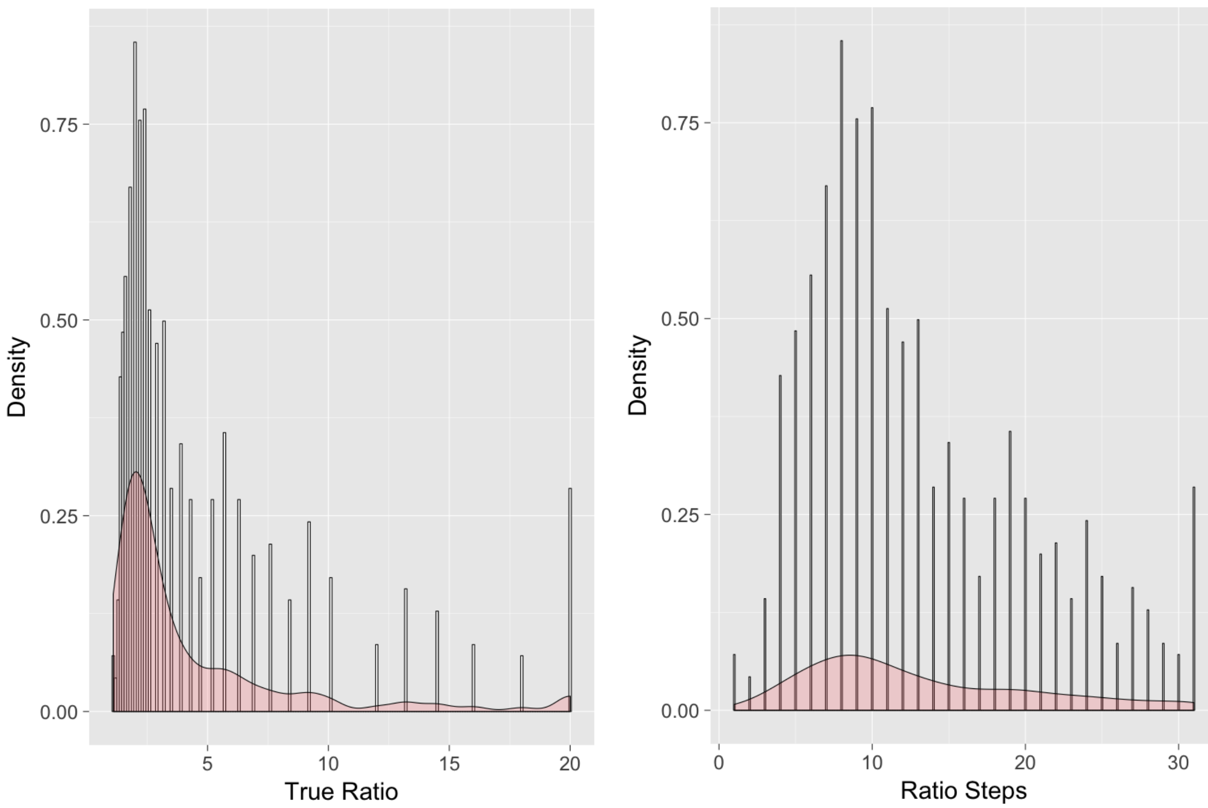


Figure 31: Two histograms showing the distribution of the ratio responses linearly (True Ratio) and logarithmically (Ratio Steps).

Table 15: Summary statistics on the distribution of ratio responses. Comparison between linear representation of ratio (True Ratio) and logarithmic (Ratio Steps).

| | Min | Max | Mean | Trimmed mean | Median | Standard Deviation | Skew |
|---|---|---|---|---|---|---|---|
| **True Ratio** | 1.1 | 20 | 4.4 | 3.5 | 2.6 | 4.2 | 2.2 |
| **Ratio Steps** | 1 | 30 | 13.1 | 12.4 | 11 | 7.3 | 0.8 |

In this experiment, the compressor was designed and pre-set so that the ratio control could be adjusted with automatic make-up gain. To test the success of the configuration the correlation between the ratio and make-up gain was examined. Positive correlation between the two parameters was expected with a traditional compressor design. There should not have been correlation between ratio and make-up gain with this compressor design. This correlation is visualized in Figure 32. Upon first inspection, there appeared to be a small negative correlation. A mixed-effects regression model was built using *ratio* to predict *make-up gain*. *Subject* was input as a random effect. There was no significant correlation between the make-up gain and ratio settings ($x^2$ = .088, $p$ = .77). Automatic make-up gain was successfully implemented. The make-up gain was analyzed as a separate entity below: an indication of the preferred overall vocal level.

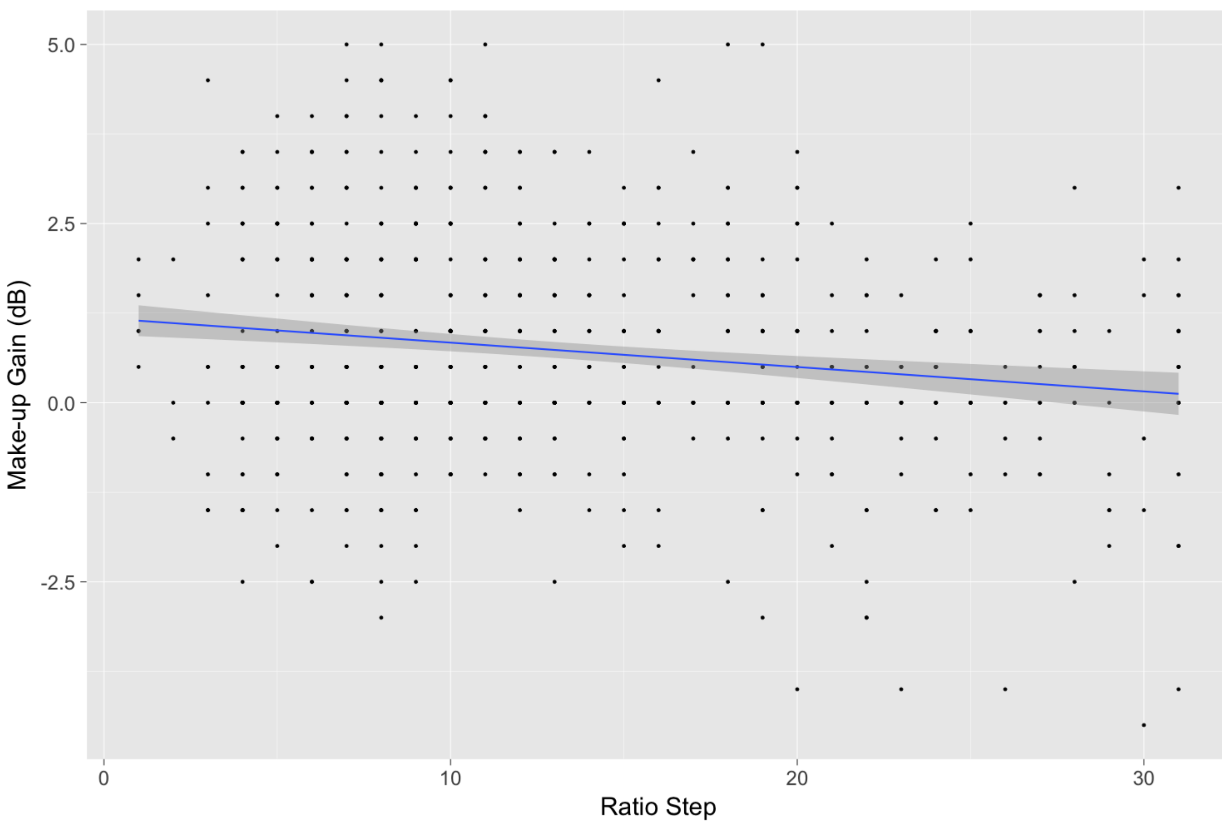


Figure 32: A scatterplot and linear model showing the relationship between the ratio and make-up gain settings for all participants.

On average, participants took 36.8 minutes to complete the listening test.

### 7.3.1 Primary Research Questions: Participant Improvement

#### 7.3.1.1 Response Variance

The response variance was a measurement of participant consistency in applying the same compressor settings to the same stimuli repeatedly. To account for the overall differences in ratio and make-up gain preferences between participants and for each stimulus, both compressor parameters were normalized for mean (but not standard deviation). Normalization took place within each participant, stimuli, and between the pre-/post-test. This way, a participant who used much higher ratios overall could be directly compared with a participant who used lower ratios overall. It also compensated for differences between stimuli, in that one of the stimuli might warrant a higher ratio overall. Finally, it compensated for any difference in overall ratio or make-up gain preference changes from the pre-test to the post-test. The means were normalized so that the standard deviation of the listener group responses across all stimuli could be directly compared.

Figure 33 displays the normalized ratio settings for each listener group for both the pre-test and post-test. The mean/median was centered around a value of zero for each group due to the normalization. What was of interest was the degree of spread in responses. Levene's test for equality of variance between groups was used to assess if the changes in variance were statistically significant. The trained group of students experienced a statistically significant reduction in the variance in their responses from a standard deviation of 4.85 ratio steps in the pre-test to 3.31 ratio steps in the post-test ($F(1, 286) = 16.89$, $p < .001$). The untrained group of students did not experience a change in the standard deviation of their responses ($F(1, 268) = 1.41$, $p = .24$). The trained group of students were more consistent in their responses after the training while the untrained group of students had no change in their consistency. This supported the hypothesis that the trained students would improve at a higher rate than the untrained students. In this case, the hypothesis that the untrained students would still improve a small amount was not supported.

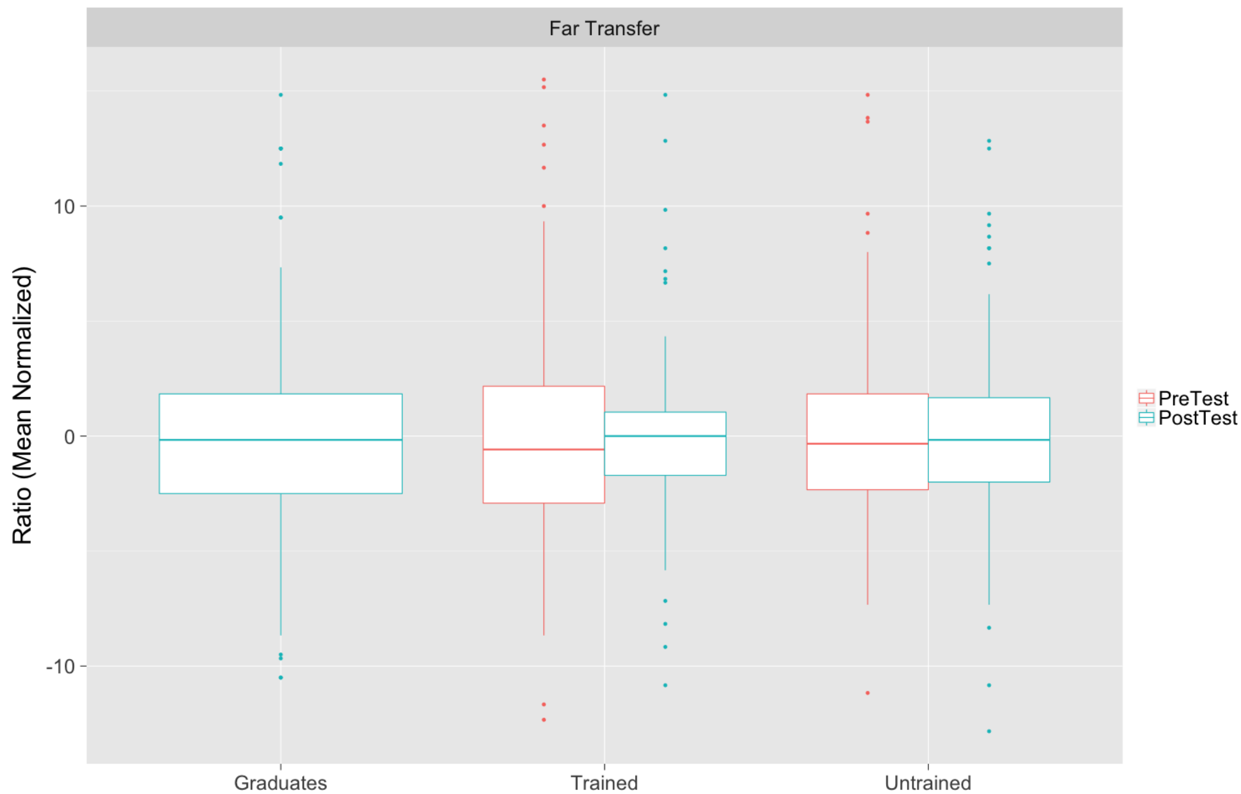


Figure 33: A boxplot showing the variance in ratio settings for each group of participants and for each test.

Figure 34 displays the normalized make-up gain settings for each listener group for both the pre-test and the post-test. The trained group of students demonstrated a statistically significant reduction in the variation in their responses from a standard deviation of 0.78 dB to 0.57 dB ($F(1, 286) = 4.60$, $p < .05$). The untrained group of students also demonstrated a reduction in the variation in their responses from a standard deviation of 0.75 dB to 0.44 dB ($F(1, 286) = 25.52$, $p < .001$). Without a statistical test for the interaction effect, it is not known if the reduction demonstrated by the untrained group was significantly larger than the reduction demonstrated by the trained group. These results supported the hypothesis that participant training in DRC would not necessarily transfer to other more distantly related tasks (setting of static volume levels). Since all students improved from test to test, it was not clear which learning factors contributed to that improvement.

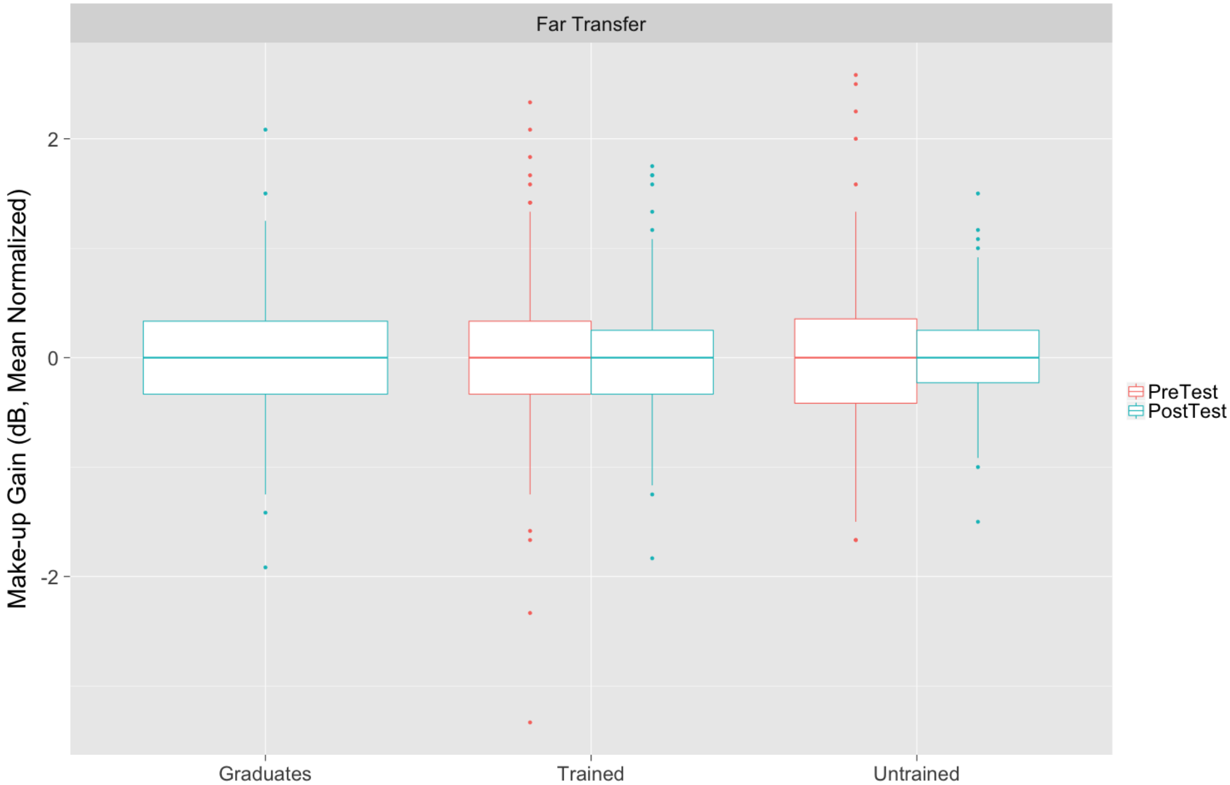


Figure 34: A boxplot showing the variance in make-up gain settings for each group of participants and for each test.

#### 7.3.1.1.1 Comparison to Graduates

Pairwise Levene's tests with the Bonferroni correction for multiple comparisons (new significance level of $p < .0125$) showed that the professional graduates had significantly more variance in their ratio responses (SD = 4.53 ratio steps) than the trained group in the post-test (SD = 3.31 ratio steps) ($F(1, 286) = 10.64$, $p < .01$). There was no significant difference in variation between the graduates and the trained group in the pre-test or between the graduates and the untrained group in either test. This supported Quesnel's findings that with training, students can outperform professionals. The hypothesis that the graduates would have a significantly lower variance than the untrained students was not supported.

Another set of pairwise Levene's tests with the Bonferroni correction on the make-up gain responses shows that the professional graduates had significantly less variation in their make-up gain responses (SD = 0.57 dB) than the untrained students in the pre-test (SD = 0.75 dB) (*F*(1,

286) = 7.58, $p < .01$) but in the post-test the untrained students had significantly less variation in their make-up gain responses (SD = 0.44 dB) than the professional graduates ($F(1, 286) = 7.06$, $p < .01$). The professional graduates were statistically similar to the trained group of students for both tests. These findings were difficult to explain and did not support the hypotheses.

### 7.3.1.2 Time Taken to Complete a Trial

Figure 35 shows the response time for each listener group for both the pre-test and post-test.

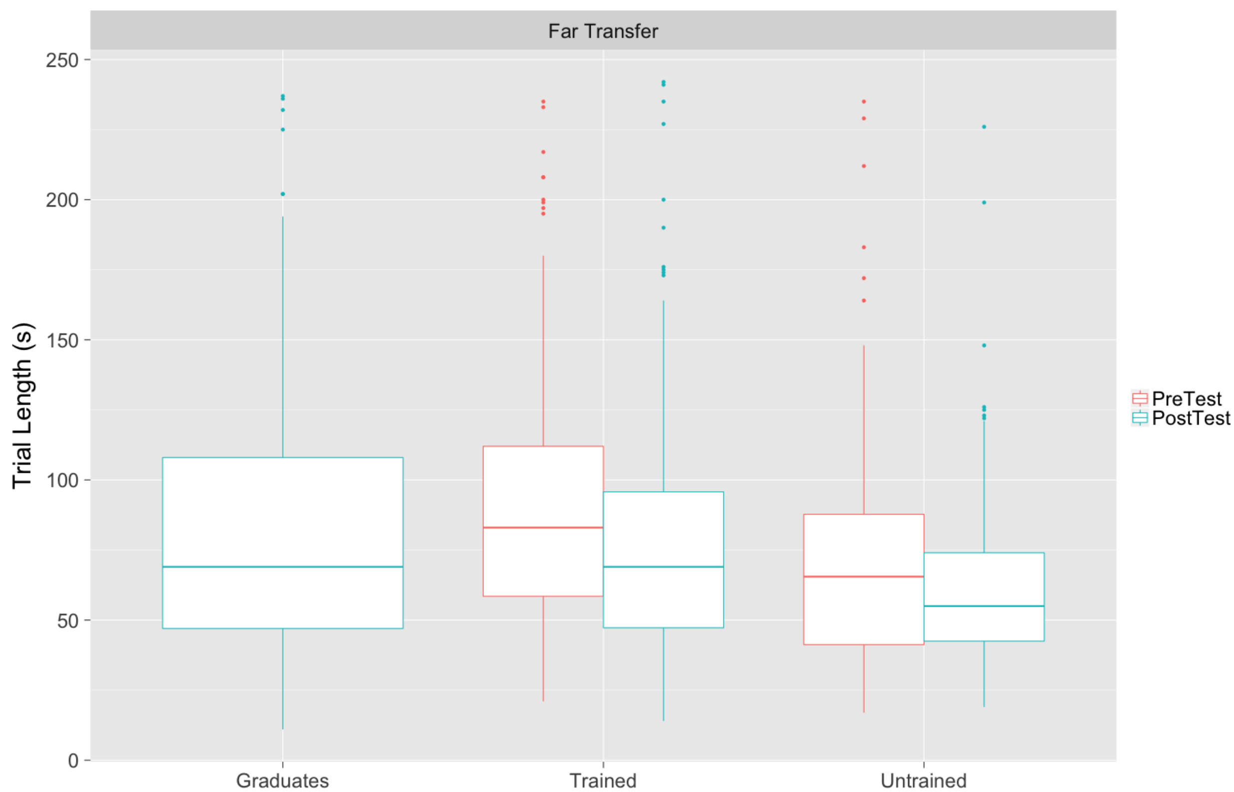


Figure 35: A box plot displaying the trial time as a function of the pre- and post-test by listener group.

A mixed-effects regression model was fit to the data using the *listener group* and *pre-/post-test* variables to predict *trial time*. The *subject*, *stimuli*, *trial time of day*, and *block order number* variables were input as random effects. The interaction effect between *listener* group and *pre-*

*/post-test* was also included. This model was a statistically significant fit to the data when compared to a baseline model ($x^2 = 17.21$, $p < .001$). Additional analysis showed that there was no interaction effect between *listener group* and *pre-/post-test* ($x^2 = 2.24$, $p = .13$) and no effect of *listener group* ($x^2 = 2.64$, $p = .10$). When pooling the two student groups together, there was a significant reduction in trial time from 82 s to 73 s ($x^2 = 12.52$, $p < .001$). The results of the previous three chapters led to the hypothesis that there would be no improvement in response time. This was not the case. Improvement was measured but it was similar for both student groups.

#### 7.3.1.2.1 Comparison to Graduates

To examine the relationship of the graduate's trial time to the students, a mixed-effects regression model was fit to the data using *listener group* to predict *trial time*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. This model was a statistically significant fit when compared to a baseline model ($x^2 = 34.52$, $p < .001$). Multiple comparisons of means using Tukey contrasts revealed that the graduates had a significantly longer trial time than the untrained students in the post-test ($p < .01$). There were no other statistically significant differences.

#### 7.3.1.3 Participant Comments

There were two participant comments from the trained group of students that indicated a self-perceived difference in ability. "…I knew more what I was listening for compared to the first time." "At first, I was scared of over compressing since I didn't not know how to listen for compressor artifacts. Now I know what is going to happen with each of the knobs and I can use the compressor effectively." These comments supported the hypothesis that the trained students would improve. There were no comments indicating a self-perceived difference in ability from the untrained students.

### 7.3.2 Secondary Research Questions: Additional Factors

#### 7.3.2.1 Ratio and Make-up Gain Mean Settings

In addition to participant consistency, the average ratio for each listener group for both the pre-test and post-test was examined. Figure 36 shows that both student groups preferred lower ratios (more dynamic range) in the post-test compared to the pre-test.

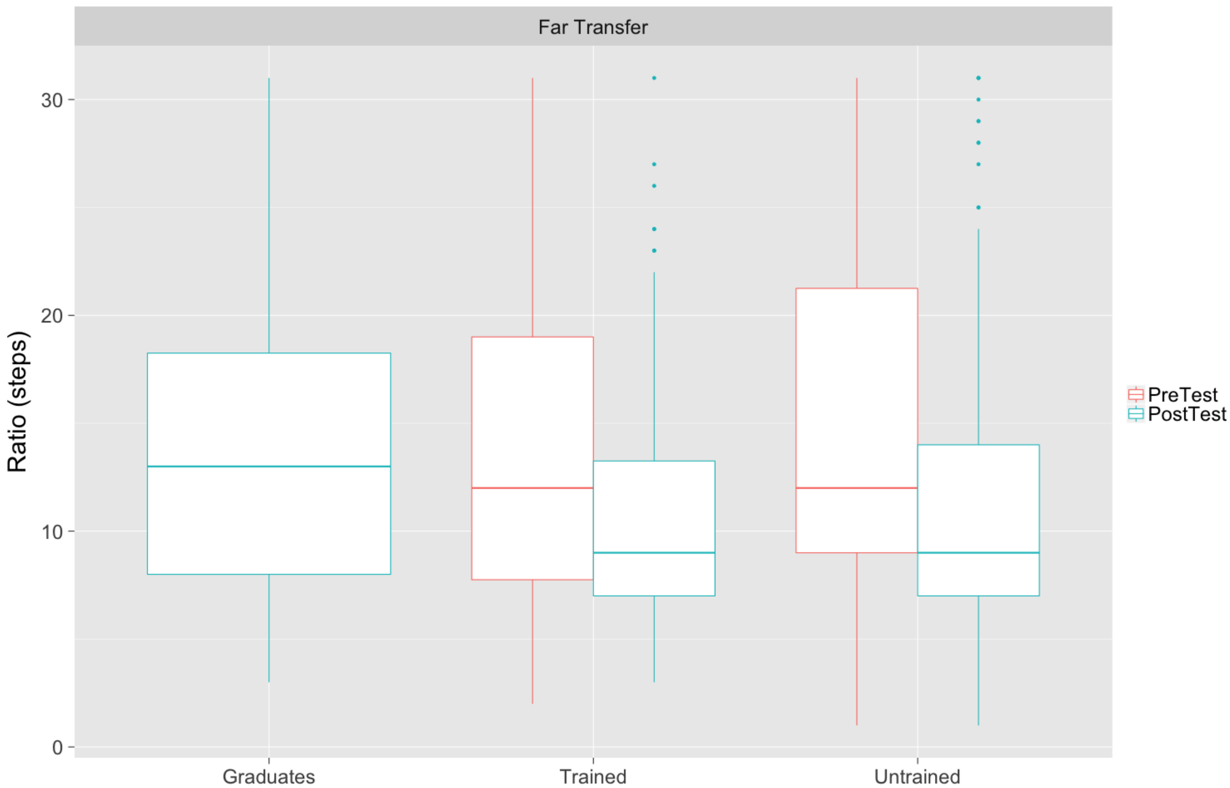


Figure 36: A boxplot displaying the ratio response in steps for each listener group for both the pre-test and post-test.

A mixed effects regression model was built using *listener group* and *pre-/post-test* to predict *ratio*. The interaction between the *listener group* and *pre-/post-test* was also examined. The *subject*, *stimuli, trial start time,* and *block order* were input as random effects. This model was a statistically significant fit when compared to a baseline model ($x^2 = 42.81$, $p < .001$). Further examination revealed there was not a significant interaction between *listener group* and *pre-/post-test* ($x^2 = 2.56$, $p = .11$). When pooling the data from the two groups together there was a significant reduction in the ratio ($x^2 = 39.41$, $p < .001$) from the pre-test (ratio = 3.9:1) to the post-test (ratio = 2.6:1). There was no significant effect of *group* ($x^2 = .54$, $p = .46$). Both student groups demonstrated a similar amount of reduction in the preferred ratio from the pre-test to the post-test. Since there were no previous results to compare to, there was no hypothesis and this result can be considered exploratory.

Figure 37 shows a similar examination of the make-gain setting. Both student groups appeared to mix the vocals a little bit quieter in the post-test than the pre-test.

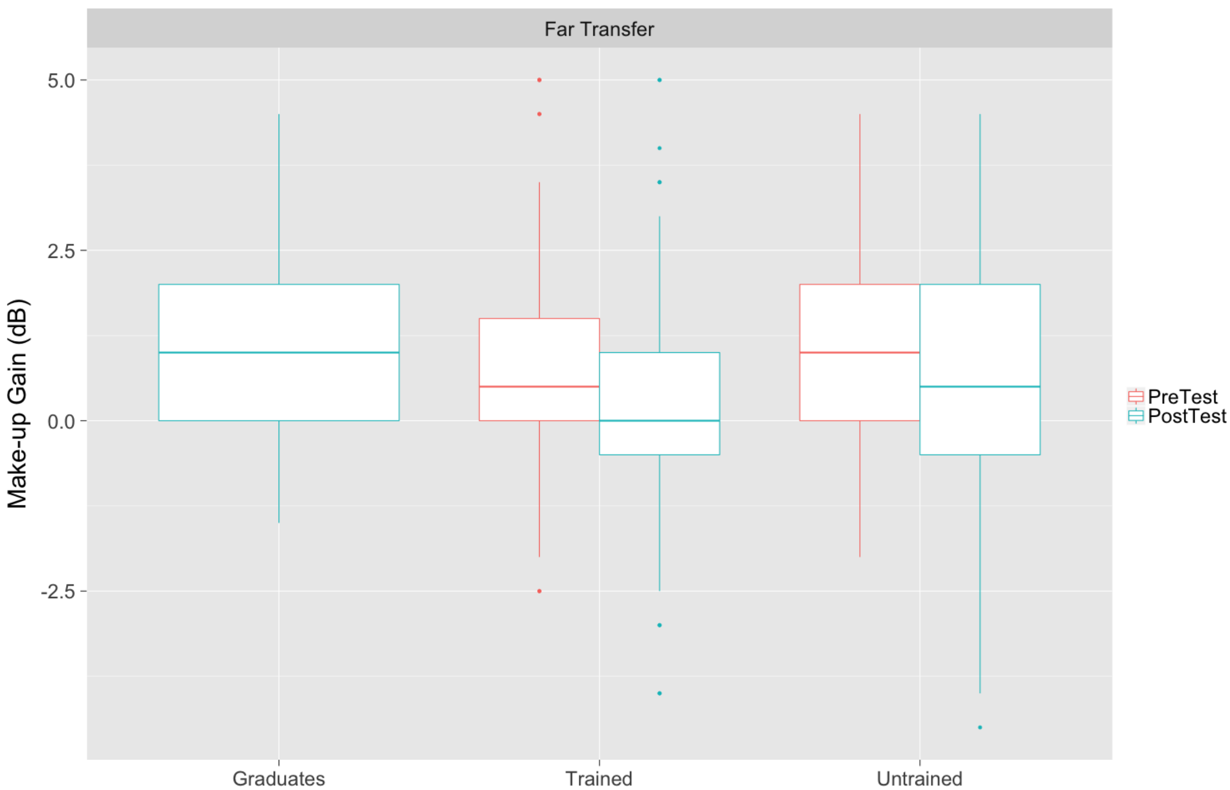


Figure 37: A boxplot displaying the make-up gain response in steps for each listener group for both the pre-test and post-test.

A mixed effects regression model was built using *listener group* and *pre-/post-test* to predict *make-up gain*. The interaction between the *listener group* and *pre-/post-test* was also examined. The *subject*, *stimuli, trial start time,* and *block order* were input as random effects. This model was a statistically significant fit when compared to a baseline model ($x^2 = 31.34$, $p < .001$). There was no significant interaction between *listener group* and *pre-/post-test* ($x^2 = 3.26$, $p = .07$).

When pooling the data from the two groups together there was a significant reduction in the make-up gain ($x^2 = 26.82$, $p < .001$) from the pre-test (0 .87 dB) to the post-test (0 .41 dB). There was no significant effect of *group* ($x^2 = 1.09$, $p = .30$). Both student groups demonstrated a similar amount of reduction in the preferred level of the vocal from the pre-test to the post-test.

##### 7.3.2.1.1 Comparison to Graduates

Looking back at Figure 36, the professional graduates appeared to prefer a ratio similar to the students during their pre-test. A mixed-effects regression model was fit to the data using *listener group* to predict *ratio*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. This model was a statistically significant fit when compared to a baseline model ($x^2 = 38.08$, $p < .001$). Multiple comparisons of means using Tukey contrasts revealed that the graduates set a significantly higher ratio (3.9:1) than the trained students (2.6:1) in the post-test ($p < .001$). There were no other statistically significant differences. The professional graduates had a similar preference for ratio as the untrained students in both tests and the trained students before training. The graduates preferred a higher ratio than the trained students in the post-test.

Looking back at Figure 37, the make-up gain preferences were examined with a mixed-effects regression model using *listener group* to predict *make-up gain*. The *stimuli, trial time of day*, and *block order number* variables were input as random effects. This model was a statistically significant fit when compared to a baseline model ($x^2 = 42.05$, $p < .001$). Multiple comparisons of means using Tukey contrasts revealed that the graduates set significantly higher make-up gain than the trained students in the pre-test ($p < .05$), the trained students in the post-test ($p < .001$), and the untrained students in the post-test ($p < .001$). The professional graduates had a similar preference for the vocal volume level as the untrained students in the pre-test. However, their preference was for a louder vocal when compared to all other student group/test combinations.

#### 7.3.2.2 Other Factors

Unlike the in-context and near-transfer listening tests, there were no significant effects of *stimuli, trial number, time of day,* or *block order* on participant performance (consistency). There was no support for the ongoing hypothesis that *time of day* (mostly likely related to fatigue) or *stimuli* (most likely related to difficulty) would be a factor.

#### 7.3.2.3 Participant Comments

The participant comments that did not discuss elements of learning or improvement were divided up into roughly three categories: 1) comments on varying difficulty, 2) comments about the listening task and its limitations, 3) comments about learning and motivation. Comments were converted to short form and can be seen in Table 16.

Table 16: Participant comments divided up into three categories. The number beside each comment indicates the number of times that comment was recorded.

| **Comments on varying difficulty** | **Listening Task and Limitations** | **Learning and Motivation** |
|---|---|---|
| Easier than the matching tests (2) | Wanted access to fader automation (3) | Learned a lot from the experience (5) |
| Some listening fatigue | Wanted control of all compressor settings | Enjoyed the listening task (4) |
| | | Enjoyed the stimuli (2) |

Overall, participants found the far-transfer preference task easier than the in-context and near-transfer matching tasks. A few participants mentioned they were not always able to find a completely satisfactory ratio and make-up gain setting. They mentioned they would need access to volume level automation to input specific changes at specific times. Many participants enjoyed the listening test and found it beneficial for skill development.

## 7.4 Discussion

Overall the results of the far-transfer listening test supported the main hypotheses. The trained group of listeners improved in their ratio response variance after the *dyntet* training. It was assumed that this improvement did not occur as a result of their ongoing education since the untrained listener group did not improve. In addition, both student groups had the same level of exposure to the testing interface, compressor algorithm, and stimuli. It was assumed that the improvement measured in the trained group was not due to familiarization. Therefore, it was concluded that this performance increase was due to their exposure to the *dyntet* training program. This listening test was designed to closely resemble a real-world mixing task. This suggested that skills learned in the *dyntet* training program were transferred. This finding was under the assumption that being able to set-up a compressor consistently is a desired skill in mixing. The trained listeners completed an average of 11.55 hours of training. Their ratio consistency improved from a standard deviation of 3.85 ratio steps to 3.31 ratio steps in that time. The mean ratio step was 13, with an equivalent ratio of 3.2:1. The equivalent improvement of the standard deviation of 0.5 of a ratio step is equivalent to about 0.15 when measured as a true ratio (x:1). For example, on the "Universe" vocal, 68 % of responses had a final loudness range between 11.5 LU – 8 LU before training and 68 % of responses had a loudness range between 11.2 LU - 8.1 LU after training. This improvement was interpreted as a small effect size, but still beneficial given the number of hours spent training.

Quesnel's result that trained students can outperform professional engineers held true. In this case, the professionals all had a great deal of past experience with TET interfaces and were not disadvantaged in that way. Also, the listening test consisted of a task, compressor design, and stimuli that no participants were familiar with. When looking at the group demographic data, the trained group differed from the graduate professionals in two main ways: 1) the trained group received the training, and 2) the graduates had an average of +7.2 more years of audio experience. Therefore, it seemed that the 11.55 hours the students spent training was more effective in reducing ratio variance than the 7.2 years of additional audio experience possessed by the graduate professionals.

Since automatic make-up gain was used in this study, the improvement in the variance of all students' make-up gain settings can be thought of as an improvement in the ability to set an overall vocal level consistently. This parameter did not affect the dynamics of the performance and was not involved in the user adjustment of dynamics. The change was not due to the *dyntet* training program since both student groups improved. Instead, it was likely due to their ongoing education or familiarization with the listening test. The untrained students had more variance in their pre-test responses than the professionals but less in their post-test. Therefore, it was assumed that the measured improvement was due to familiarization and not the audio education that all professionals would have also received. The effect size of this improvement was also small (SD 0.77 dB to 0.51 dB).

Contrary to the results in previous chapters, there was a significant reduction in the students' trial time from the pre-test to post-test. The improvement was not due to the *dyntet* training program since both student groups improved equally. The professionals performed the same as the students in the pre-test, but the students outperformed the professionals in the post-test. This meant the reduction in trial time was likely due to the same factors as the make-up gain variance: familiarization with the listening test. The measured improvement was a 9 % reduction (82 s to 73 s). That effect size was considered relatively large given the students did not spend more than two hours familiarizing with the listening tests.

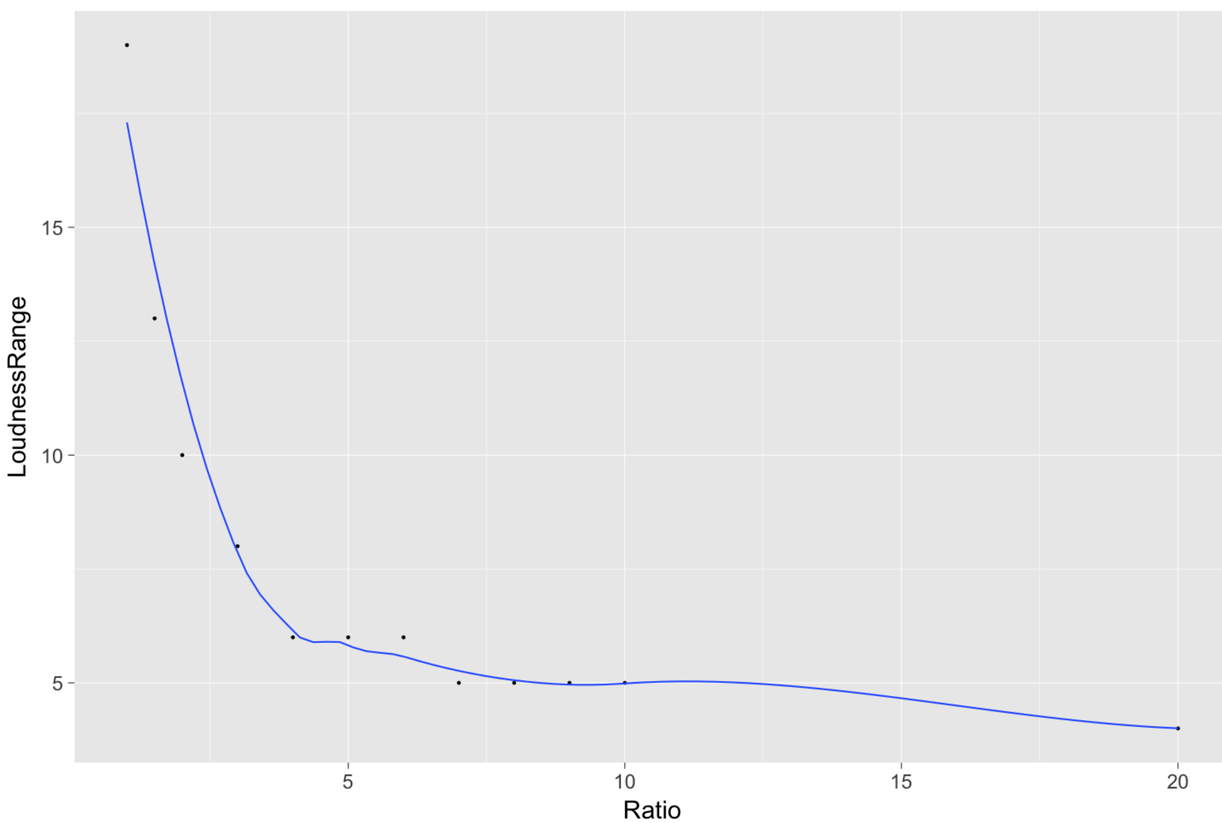


Figure 38: A scatterplot of the "Universe" vocal stem's Loudness Range according to the compressor ratio. The compressor threshold was set to -35 dBFS so that the entire vocal performance fit above the threshold. Attack time was 5.2 ms and release time was 125 ms.

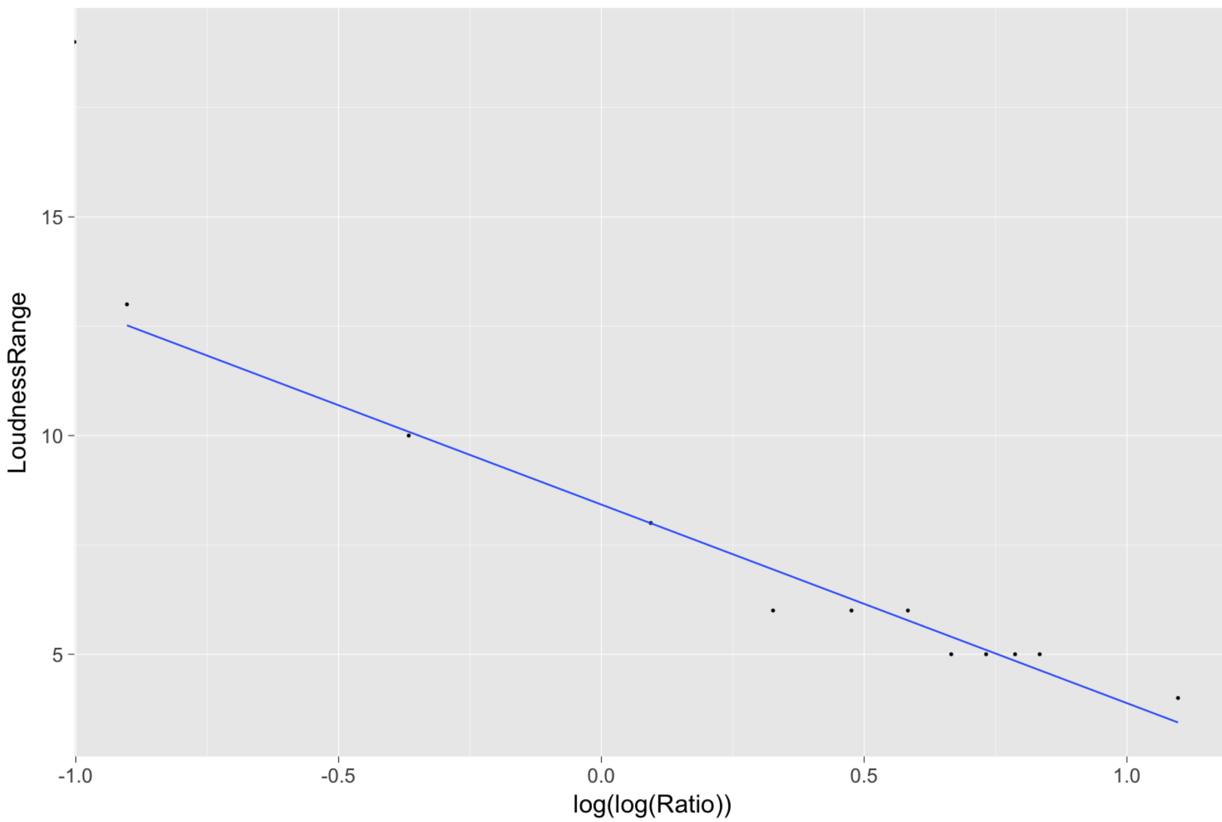


Figure 39: A scatterplot of the "Universe" vocal stem's Loudness Range according to the compressor ratio. The ratio has been transformed logarithmically twice (y = log(log(x))) to achieve a linear relationship. The compressor threshold was set to -35 dBFS so that the entire vocal performance fit above the threshold. Attack time was 5.2 ms and release time was 125 ms.

There was some evidence in the results that the ratio control was perceived logarithmically by the participants when listening for dynamic range. This support came from the comparison of the distribution of responses between the ratio steps and the true ratio. The ratio steps appeared to follow a more normal distribution. A normal distribution is typical of a linear variable measured in the real-world. A skewed distribution suggests a non-linear relationship. This logarithmic relationship can also be seen in Figure 38. The loudness range (LU, measured according to ITU-R BS.1770-4) of the "Universe" vocal track was plotted according to the ratio of the compressor [30]. It was clear that as the ratio increased the dynamic range of the vocal approached an asymptotic level (~4 LU). Also, the range of the ratio control that produced the largest change was from 1:1 to 4:1. Figure 39 shows that a double logarithm was needed to achieve a linear relationship between the variables. With this set of compressor settings, the loudness range could be predicted from the logarithm of the logarithm of the ratio. Going forward, it is suggested that this relationship be clearly explained when teaching audio students to work with compressors.

There was also evidence in the results to support the use of a rotation point design. This applies when the primary goal is to reduce the dynamic range of the audio material equally over its entire dynamic range. In this case, auto-make up gain was successfully applied. The user was freed to only adjust one control when manipulating the dynamic range. This function was only tested under this one scenario. In other DRC scenarios, such as peak limiting, it is still unexplored.

In the secondary analysis some interesting results emerged. Both student groups showed a preference for a larger vocal dynamic range in the post-test than the pre-test. Both groups experienced a similar preference change and the professionals' preference was similar to the students in the pre-test, so this was likely due to familiarization with the stimuli, testing interface, or compressor algorithm. The effect size was large (3.9:1 to 2.6:1). For the "Universe" vocal, this represented a change in loudness range from 8.6 LU to 10.5 LU. A very similar result occurred with the make-up gain parameter. Both student groups preferred a quieter vocal in the post-test and the professionals had a similar preference to the students in the pre-test. Once again, this change was likely due to familiarization. The effect size is relatively small (0.46 dB difference) when compared to the ~0.2 dB just noticeable difference for loudness [75]. These results may have implications in the context of preference listening tests which typically take place in short 30 minute to 1 hour listening tests. It's possible that if a preference listening test does not allow for adequate participant familiarization that the results could be inaccurate. There was evidence here that forming a stable preference takes time and requires lengthy familiarization with the stimuli or device under test.

To summarize, when teaching audio students about compression, the ratio control should be described as having a logarithmic relationship to the resulting dynamic range of the audio. There was evidence that the inclusion of a rotation point in compressor design provided effective automatic makeup gain for the user. Finally, the far-transfer listening test was successful in measuring the transfer of learned skills during the *dyntet* training program to a related listening task. This listening task was modeled on a typical mixing task. Skills learned during *dyntet* training could be useful in day to day audio work. The *dyntet* training program appeared to be time efficient when compared to acquiring skills "on the job".

# 8 Testing Training Effect V: Retrospective Think-Aloud Experiment

## 8.1 Introduction

In addition to the listening tests above, a retrospective think-aloud experiment was administered as an additional measurement of skills learned. This experiment was designed to assess whether the two student groups could clearly and logically explain their preferences for DRC settings on provided stimuli. In audio engineering work, DRCs are applied when the engineer identifies a problem, imagines the final corrected result, and then arrives at parameter settings that achieve the goals. It was acknowledged that high-level engineers might differ in their goals or approaches. In this study, the preference for various compressor settings was not the focus. Instead, the analysis was based on the students' retrospective explanations for their decisions. It was hypothesized that the trained students would perform better than the untrained students. Performance was measured as the clarity and completeness of the explanations, the proper use of technical language, and success in achieving the stated objectives.

## 8.2 Methodology

### 8.2.1 Listening Task

In this experiment, participants were asked to apply a DRC to a single musical instrument within the context of a complete musical work. Participants had access to all compressor parameters (threshold, ratio, attack, release, make-up gain) and could take as long as they needed to arrive at a response. A wide range of all parameters was available, including a ratio of 1:1 (no compression). Participants were asked to arrive at the settings "they preferred the most of all possible combinations". Immediately after submitting their response, their parameter settings were recorded, and participants were asked the following open-ended questions.

- If you applied compression, what did you hear in the uncompressed signal that you wanted to change? Describe how you wanted the signal to sound after applying compression.
- (5 questions) How did you arrive at the (1. threshold, 2. ratio, 3. attack, 4. release, 5. make-up gain) setting? What were you listening for as you adjusted this parameter?

### 8.2.2 Stimuli, Compressor Design, Testing Location

Each participant only completed two trials due to the long period of time required to complete the experiment. Two stimuli were used in randomized order. The stimuli were taken from the "near-transfer" listening test so that all participants had equal familiarity ("Vocal" and "Bass", see Chapter 6.2). The compressor and user interface were the same as the "in-context" and "near-transfer" listening tests. The same student listener groups were used again here (trained and untrained). The listening test was conducted within two months of completing the training. The listening test took place in the same CIRMMT Critical Listening Lab.

### 8.2.3 Response Grading

The analysis of the student responses was highly subjective and highly sensitive to experimenter bias. Therefore, two external graders were used to evaluate the response data. The graders were two Doctoral candidates in the Sound Recording program at McGill University and had no previous involvement in the research. Both graders had more than three years of teaching experience at the undergraduate level. Student identification was unknown to the graders. The student responses were graded on a continuous scale from 0-5. Several absolute anchors were provided along the scale (Table 17).

The graders were provided with the student responses (compressor settings, and written answers). They were also provided with the stimuli and compressor so that the student responses could be auditioned. The graders were instructed that the student's preferences were not be evaluated. Instead, the graders were told to read about a student's intent and evaluate the results against the student's stated goals. The graders were also instructed to take the student's discussion of the various compressor parameters into consideration. Evaluation of the overall discussion would have been considered high level if it was clear, thorough, and used technical language appropriately. Justifications for the various parameter settings should have matched the stated goals and not have been counter-productive or contradictory. The grading took place in the same listening lab and took each grader about 1.5 hours to complete.

Table 17: Grading scale absolute anchors provided to the graders.

| Grade Scale (0-5) | Description |
|---|---|
| 0 | No audio engineering knowledge. |
| 1 | Beginner, makes many mistakes, has many misunderstandings, and no noticeable practical experience. |
| 2 | Intermediate, makes some mistakes, has some misunderstandings, has some practical experience to back up the partially understood theory. |
| 3 | Experienced, makes a couple mistakes, may have some partial misunderstandings, starting to accumulate a decent amount of practical experience. |
| 4 | Advanced, rarely makes mistakes, should have no significant misunderstandings, has a significant amount of practical experience. |
| 5 | Professional, complete understanding, mastery of the tools and theory, has a large amount of practical experience. |

## 8.3 Results

Only six of the eight trained students and four of the eight untrained students were available to participate. The think aloud experiment took about 30 minutes for each student to complete.

### 8.3.1 Response Grades

The average student grades are plotted in Figure 40 and described in Table 18. Both groups received an average grade between "2 – Intermediate" and "3 – Experienced". The trained group received a slightly higher average grade of 2.73 than the untrained group, 2.41. The significance of this difference was examined with a three-way ANOVA using *listener group*, *stimuli*, and *grader* to predict *grade*. With this extremely small dataset (36 data points) and small effect size, no significant effects were found $F(3, 32) = 0.61$, $p = .62$.

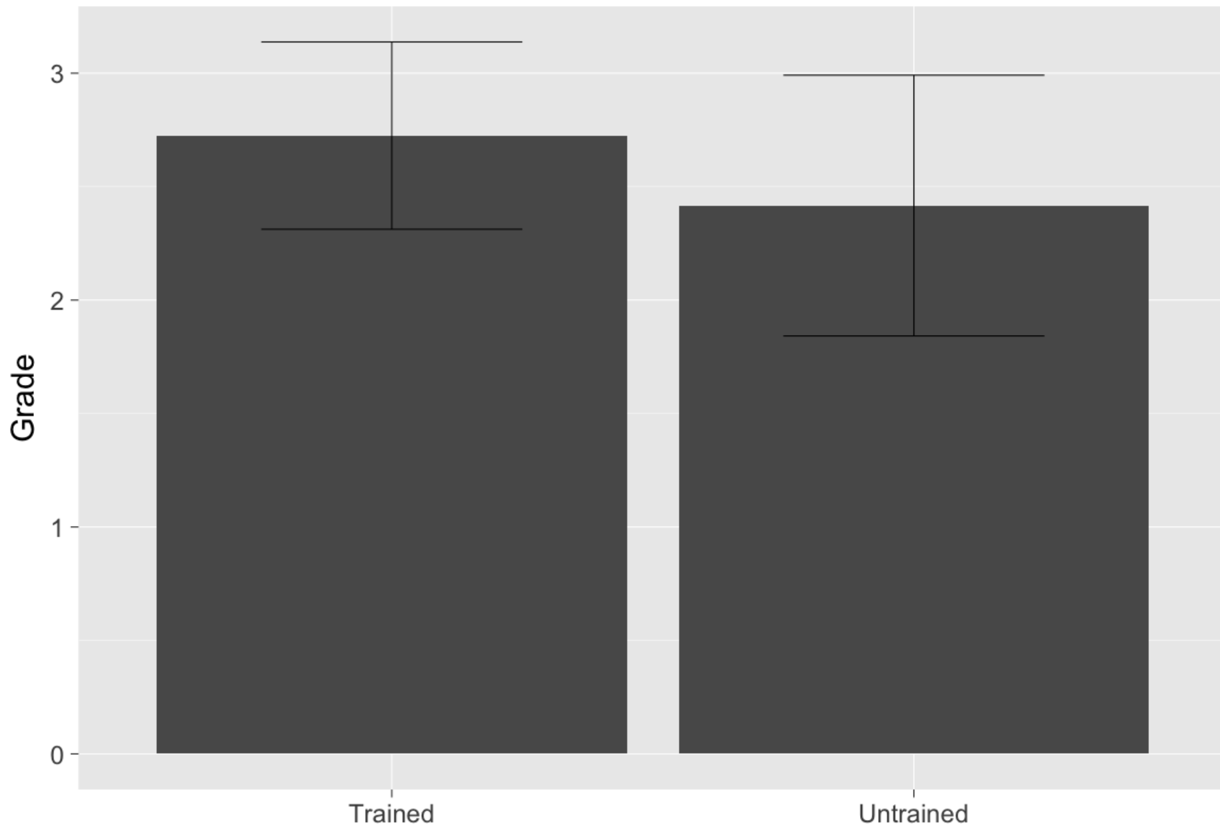


Figure 40: A bar plot displaying the average grades for each listener group.

Table 18: Descriptive statistics for the grades of each listener group.

| Group | Mean (SD) | Trimmed Mean |
|---|---|---|
| Trained | 2.73 (0.98) | 2.62 |
| Untrained | 2.41 (0.90) | 2.40 |

### 8.3.2 Example Responses

The following are the two highest and two lowest grade responses for the "bass" stimuli.

- [Average grade = 3.6 ] *"When I first heard this, I felt that I wanted to make the bass very consistent, more of a drone underneath the mix. Before compression there was an attack and then sustain, but the sustain was dying before the next note. After compression, I wanted less attack and more sustain, equal all the way to the next note. A threshold of -15 is my usual starting point, so I put it there and I totally forgot to adjust it, but I think it's ok. For the ratio,* *I wanted to make all the notes even, that's what I was listening for. Working my way up to 10:1 there was a point where everything was even. For the attack time, I started at the very fast end, it was filling the timbre a little bit so I adjusted it back and forth until it killed the attack enough for my liking but still a little bit left. For the release time, I just went up to the maximum length to make it sustain as long as it can."* (Trained participant #6)

- [Average grade = 3] *"I wanted to change the sustain, controlling the unevenness of the performance. The peaks as well, the second note is much louder than the first, I'm trying to control that, but it's mostly about the sustain. I wanted the bass to sustain from the kick drum notes. I started with a ratio of 9, because I know you can go pretty aggressive on bass, but it was too much, squashing way too much, even with a reasonable attack time. For the attack, I tried all the way up, it wasn't good, and too short was squashing too much, so I ended up somewhere in between, it sounded musical to my ears. Because it's not picked, it played with fingers, I like the smoothness of the attack, so you don't need to kill it. The release was timed to the sound, 100ms was long enough to sustain the bass note, more than 100ms was too much sustain, unnatural, you don't want the note to die, but I didn't want a sausage until the next note. 100ms sounds super natural with the tempo of the song."* (Trained participant #8)
- [Average grade = 2] *"In good pop fashion, I did not want to hear the bass. I wanted to find a balance where I could hear the pick of the note, but not really the bass in particular. The bass line itself is pretty nice with the scale up, I wanted to make sure you could hear that, but I didn't want the listener to focus on the bass. The threshold is set so that every note is hit pretty drastically, averaging -10dB gain reduction. For the ratio, I'm crushing it because I don't want you to hear the bass. For the attack, I set it a little bit longer so that you hear the front end of it before it gets crushed. For the release, I set it long enough so that you don't hear it pump back in."* (Untrained participant #14)
- [Average grade = 2.1] *"Because it's strings, there was an easier decay pattern to follow, the goal was just to even things out. The bass has a much faster decay rate than the other elements in the mix, and therefore feels slightly out of place. Given everything else that is going on, I wanted it to sweep smoothly between everything else. I think that is what the intention was besides the way it was played. That's the musical goal behind it. I ended up settling on a slightly higher ratio than I typically do. Just because it is much more dynamic. Within that, I know that with bass I like a longer attack to make sure there is some clarity still that's not being touched*

*by the compressor. For release, I kind of varied at points, I usually use a shorter release and that's where I started. Then there is this weird pulsing thing that happens at the end of the loop. I was trying to see if maybe I could fix that a little bit. However, I then ended up settling on a longer release to allow the decay to stretch out a little more. I kept going back and varying that with the make-up gain and the threshold to figure out how much of the low end I still wanted. I thought the low-end was a little overpowering, yet it would drop under everything at the end of the note."* (Untrained participant #10)

## 8.4 Discussion

There were no statistically significant differences between the grades given to the trained and untrained students. The effect size was small, therefore, a much larger sample size would have been required. Despite this result, it was encouraging that the results trended in the expected direction. It is also interesting to note that for the "bass" stimulus the two highest scores came from the trained students and the two lowest scores came from the untrained students. Having access to the student explanations was very useful as a reflection of the vocabulary, techniques, and values imparted on the students by the course instructors. Data from this type of experiment was hard to collect and analyze but was useful in "adding another piece to the puzzle".

# 9 Testing Training Effect VI: Post-Training User Survey

## 9.1 Introduction

The final measurement of learning was a post-training user survey. It was administered to assess the trained students' perceived levels of improvement. Learning from the students' point of view was used to provide a small amount of additional support for the overall hypotheses. It was expected that the students would perceive the benefit of the training program, report improvement within the context of the training environment, and provide examples of skill transfer.

## 9.2 Methodology

The post-training survey was administered to the trained group of students after the training had taken place. The survey was electronic, voluntary, and anonymous. Students could answer or skip any question.

### 9.2.1 Survey Questions

The survey questions listed in Table 19 were formed to examine a few things from the students' point of view: overall usefulness of the program, general listening skills, transferrable listening skills, and improvements to the training program. Most questions were open-ended and encouraged full-sentence answers. Question #1 asked for a rating on a five-point scale.

Table 19: Exit survey questions.

| Question # | Question |
|---|---|
| 1 | "I consider the benefits of using the dyntet training program to be worth the time I put in." |
| 2 | "Please comment on the reason for your rating above" |
| 3 | "What skills do you think you have developed using the dyntet training application?" |
| 4 | "Have the skills you've learned using the dyntet training program helped you with any other listening tasks? If yes, please give as many examples as you can remember." |
| 5 | "Has the dyntet training helped you in your day-to-day audio engineering work? If yes, please give as many examples as you can remember." |
| 6 | "Are there any improvements you think should be made to the dyntet training application? |

### 9.2.2 Analysis

Responses to the open-ended questions were reduced and coded to form the summary of results tables. If a student mentioned a specific topic, it was counted as one instance. The topics used in the coding are detailed in Table 20.

Table 20: Topic descriptions.

| Topic | Description |
|---|---|
| Best practices | Knowledge of how compressors should be applied. Being able to recognize and follow conventions and aesthetics in applying compression. |
| Compressor listening | Ability to discriminate small changes in the sound of compressors. The ability to associate changes heard with compressor parameters. |
| Compressor operation | Understanding of how a compressor works, what the controls do, and how they interact. |
| Confidence | Level of confidence when listening to and operating compressors. |
| Critical listening | Ability to listen to sound with an ear for technical parameters. |
| Listening tests | Perceived sensitivity when participating in any variety of controlled listening tests. |
| Mastering | The practice of mastering, as it applies to audio engineering. |
| Mixing | The practice of mixing, as it applies to audio engineering. |
| Speed | Ability to work quickly. Under the assumption that the quality of work remains the same or improves with improvements in speed. |
| Topic difficulty | Perception of difficulty on the topic of compression. |
| Transfer of skills | Ability to transfer skills learned to related listening tasks. |

Table 21: Comment valence descriptions.

| Valence Rating | Description |
|---|---|
| -2 | Strongly negative comment. |
| -1 | Moderately negative comment. |
| 0 | Neutral comment. |
| +1 | Moderately positive comment. |
| +2 | Strongly positive comment. |

Each statement instance was also assigned a valence, from -2 to +2. The average valence was calculated for each topic to determine if comments were mostly negative or positive. A description of the valence values is displayed in Table 21**.**

The coding of comments and assignment of valence values was executed by the primary researcher. A review of the analysis was conducted by two separate researchers, unassociated with the project, to confirm there was no obvious bias.

## 9.3 Results

Seven out of the eight trained students completed the post-training user survey. All seven students answered all questions. The survey was completed in a one-month period after the completion of the pre/post listening tests (two months after the completion of the training program).

### 9.3.1 Overall Usefulness of the Program.

All students "strongly agreed" that the benefits of the training program were worth the time they put in (5/5 rating). In addition, Table 22 summarizes the student comments in relation to this question. The most common topics were self-perceived improvement in learning to listen to compressors, the operation of compressors, and the transfer of those skills to other listening tasks. This was precisely the goal of the *dyntet* program and confirms the hypothesis. Comments were conflicting on whether or not the training program successfully addressed best practices (taste and aesthetics).

Table 22: Participant comments according to category, responding to "why" they gave the rating of "I consider the benefits of using the *dyntet* training program to be worth the time I put in." "n" represents the number of students who made that comment.

| Category | n | % of all responses | Average Valence | Sample comments |
|---|---|---|---|---|
| Compressor listening | 6 | 35 % | +1.5 | "This program has greatly increased my perception of compression…"<br>"It's a great program…for developing critical listening skills…" |

| | | | | |
|---|---|---|---|---|
| Compressor operation | 3 | 18 % | +1.7 | "(Before training, compression)…always confused me, I never fully understood how it worked. Now it is a lot more clear…" "This training clarifies how compression works…" |
| Transfer of skills | 3 | 18 % | +2 | "(*dyntet* has)…allowed me to work at a much faster rate…" "(*dyntet*)…has significantly improved the quality of my mixing" |
| Best practices | 2 | 12 % | 0 | "I'm still a little shaky on when to use compressors myself…" "(*dyntet teaches you*)…how compression should be applied to various things" |
| Topic difficulty | 2 | 12 % | +1.5 | "Compression is one of the hardest effects to understand and it is often misused" "(I)…didn't use much compression before *dyntet*…" |
| Confidence | 1 | 6 % | +1 | "I have observed a positive change in the amount of confidence (I have) in decision making about dynamic processing…" |

## 9.3.2 Skill Development

### 9.3.2.1 Base Skills

All students reported developing at least one skill while participating in the training program. Most students reported this development in listening to compressors and compressor operation. This result agrees with the previous survey question and the hypothesis.

Table 23: Participant comments according to category, responding to which skills the *dyntet* training helped them improve.

| Category | n | % of responses | Average Valence | Sample comments |
|---|---|---|---|---|
| Compressor listening | 6 | 40 % | +1 | "(*dyntet* helped me)…know what to listen for when I compress and how the different parameters will affect the sound." "(I now have)…better discrimination between settings on a compressor." |
| Compressor operation | 4 | 27 % | +1 | "(I now)…understand what all the knobs and stuff do in my compressor plugins." |
| Transfer of skills | 2 | 13 % | +1 | "(I now have)…more sensitivity to dynamic range in my own work." "(I can now)…use a compressor to change the timbre effectively in my own mixes." |

| | | | | |
|---|---|---|---|---|
| Confidence | 1 | 7 % | +1 | "(I now have)…more confidence in altering parameters for a particular outcome" |
| Speed | 1 | 7 % | +1 | "(*dyntet* helped me)…get the sound and result I want out of a compressor faster." |
| Critical listening | 1 | 7 % | +1 | "(*dyntet* helped)…develop my listening skills in general…I would say I hear smaller details after doing this training" |

### 9.3.2.2 Transfer of Skill

All students reported transferring skills learned in the training program to at least one other listening task. The most common area of application was in mixing. Secondary areas were mastering, critical listening, and participating in listening tests. While agreeing with the previous questions and hypothesis, this result gave some additional insight into where skills can be applied.

Table 24: Participant comments according to category, responding to which other listening task skills learned using *dyntet* transferred to.

| Category | n | % of responses | Average Valence | Sample comments |
|---|---|---|---|---|
| Mixing | 4 | 44 % | +1.5 | "I would definitely say that I apply compression to my mixes much better than before doing this training. I can find tune attack and release time, make compression less obvious when needed, or use it to achieve more specific goals"<br>"I would say I did get better at knowing when and how to use compression, making my mixes better, so making me a better engineer." |
| Mastering | 2 | 22 % | +1.5 | "(*dyntet* helped me with)…bus compression and mastering compression."<br>"(*dyntet*)…has especially helped me with the mastering gigs I have taken on, as an extreme sensitivity to dynamic range changes and compression is essential in these situations. I've heard a huge difference in the quality of my mastering work since beginning *dyntet*." |
| Critical listening | 2 | 22 % | +1.5 | "Compression is a great tool, I love it. But the training made me realize how much I hate certain compression settings. Music off YouTube is a good example. It is so crushed, but in a terrible way…hearing the compression has ruined a few listening |

| | | | | |
|---|---|---|---|---|
| | | | | experiences for me, but I'm glad I can hear it now and know why I don't like it." |
| Listening tests | 1 | 11 % | +1 | "(*dyntet*)…has aided me in several listening tests for research purposes…" |

### 9.3.3 Training Program Design Improvements

All students made at least one recommendation to improve the training program. The most common recommendation was for more stimuli. In addition, intermediate levels, audio loop control, and level ordering changes were recommended. All these changes were logical improvements that were also recognized by the course instructors.

Table 25: Participant comments according to category, responding to how the *dyntet* training program could be improved.

| Category | n | % of responses | Average Valence | Sample comments |
|---|---|---|---|---|
| More stimuli | 4 | 33 % | -1 | "It would be nice to have more excerpts. The training gets boring as we get to know every single excerpt really well"<br>"I wanted to try with additional tracks. With additional training, I would get to know the excerpts and there would be less of a challenge." |
| Intermediate levels | 2 | 17 % | -1 | "I would add more intermediate levels in between the current levels, for a more gradual introduction to each parameter."<br>"More levels in between those that we currently have would help…I feel like we flew by level one, and then got stuck in level 2.2 for a month…" |
| Audio loop control | 2 | 17 % | -1 | "…it would be nice to see a version of the waveform below, so we can select certain sections in order to loop a particular cymbal hit, snare hit, or vocal line."<br>"…if it had a play button and bar, then I could start from wherever." |
| Level order | 2 | 17 % | -1 | "I'm wondering if it would work better to go through the different parameters all together instead of spending 1 month on each and go to the next one after reaching a certain level. I just think focusing on a different element would make me understand previous elements better…" |

| | | | | |
|---|---|---|---|---|
| GUI glitches | 1 | 8 % | -1 | "…if you could get rid of that little glitch in the compressor meter where it flickers that'd be great." |
| Best practices | 1 | 8 % | -1 | "It would be cool, if possible, to have a level where you also practice using the settings in a mix situation. I liked the level where we could just read what other people would do when using compression in a mix." |
| Login system | 1 | 8 % | -1 | "(Could you)…have the system keep track of the user automatically?" |

## 9.4 Discussion

When completing the survey, there was likely a bias toward positive valence responses as a form of student response bias [22]. It was clear that the students of the training program reported improvements in their compressor listening skills. In addition, they reported improvement in compressor operation skills. Participants also reported transfer of learned skills to their mixing and mastering work. These were the primary goals of the training program. Therefore, the results were considered evidence of success. It was logical that participants perceived transfer of skills primarily to mixing over mastering since most training was done in this context. Only a couple training levels addressed master buss compression. The training program did not appear to effectively address best practices in the application of compressors. This was not a primary objective of the training program, but this was a frequently requested lesson in the Advanced Technical Ear Training class and should be addressed.

The primary recommendation for an improved training experience was additional musical excerpts. This would provide more variety in the listening experience, maintain participant interest, and promote the generalization of skills. The need for intermediate levels was also recognized in the Chapter 4 analysis. The suggestion of interleaving level types was also interesting (working on ratio one week, then timing the next). Compressor parameters work in combination with each other so a change in one parameter often affects the impact of another. An interleaved level approach would support this relationship instead of giving the impression of independent controls. The request for control over the audio excerpt loop duration and in/out points was assessed from two points of view. With full control, it was hypothesized that it would make the listening tasks easier for students. It would allow them to zero in on particularly informative parts of the excerpt. However, the loop points were very specifically chosen so that participants would be forced to listen over a significant portion of time. This was under the assumption that this sort of practice would improve and promote the use of short-term and long-term auditory memory (rather than iconic short-term memory). All of these improvements, minus the loop control, are slated to be integrated in the next version of the training software.

# 10 Conclusion

## 10.1 Summary

Audio engineering technology (AET) education will likely continue to be offered at post-secondary institutions. Many industry professionals agree that students need to learn how to use a variety of signal processors. This includes learning their operation (understanding, confidence, speed), learning to hear their effects (auditory sensitivity, isomorphic mapping, auditory memory), and learning how to apply them in an artistic way (best practices). Learning the audible effects of signal processors can be addressed using technical ear training (TET). Within the family of commonly used signal processors, dynamic range compressors (DRCs) are an important, complex, and somewhat neglected member.

Through the research reported in this dissertation the following tools have been created for DRC TET.

1. The *dyntet* application.
    a. Two training tasks (*matching* and *absolute ID*)
    b. An example level structure with increasing difficulty
    c. A basic feedforward compressor algorithm
    d. A diverse high-quality stimulus set
    e. A set of ecologically valid compressor settings (custom set for each stimulus)
    f. Mixing and Mastering contexts
2. A second compressor algorithm with a (unique) design based on a rotation point.

Through this work, the following hypotheses have been confirmed.

1. When using the *dyntet* application, students improved on the training task at a rate similar to previously reported equalizer TET applications. This improvement could have been a result of 1) familiarization with the testing application, 2) familiarization with the stimuli, 3) familiarization with the processor algorithm, 4) increased listening skill level due to their ongoing education, 5) increased listening skill level due to the training application. This result was supported by the training data (Chapter 4) and exit survey (Chapter 9).

2. *Cumulative time spent training* was a better predictor of participant improvement, and therefore a better predictor of the amount of training completed, than *cumulative number of trials completed* (Chapter 4).
3. As found in previous research, there were differences in the overall skill levels between participants (Chapters 4-9).
4. The trained students successfully transferred the skills learned to a real-world mixing task. It was determined that these transferred skills were a result of participating in the *dyntet* training program. This result was supported by the far-transfer listening test (Chapter 7 and exit survey (Chapter 9).
5. The trained students were more consistent in their compressor settings than the professional graduates during a real-world mixing task. An average of 12 hours training proved to be more beneficial than an average of seven years of production experience (Chapter 7).
6. The specified rotation point compressor design was effective in implementing automatic make-up gain as the compressor ratio was adjusted (Chapter 7).
7. With an ecologically valid attack and release time setting, the rotation point compressors' ratio had the following relationship to the resulting loudness range of a vocal (Chapter 7).
$$Loudness\ Range = \log(\log(Ratio))$$
8. Familiarization with the far-transfer listening test resulted in a change in preference among all student participants. During the second test, lower ratios and lower loudness levels were used on the vocals within the mixes. This result has important implications for preference listening tests. It revealed that short preference listening tests might be inaccurate. Instead, extended amounts of stimuli and device-under-test familiarization are likely required to achieve a stable measurement of participant preference (Chapter 7).

Within the circumstances of this study, several hypotheses could not be confirmed. This yielded the following results that are specific to this work and may not be repeatable.

1. Trained students did not improve their response time (Chapters 4-7).
2. A statistical fit for a logarithmic learning curve could not be found (Chapter 4).
3. Learning measured in the training data could be due to any of the 1-5 factors listed above. The in-context and near-transfer listening tests were not successful in systematically

eliminating Factors 2 and 4. (Familiarization with the stimuli, learning due to their ongoing education) However, elimination of those factors in combination was still effective in the far-transfer listening test (Chapters 5-7).

4. There was no clear pattern of changing participant performance according to time of day (Chapters 4-7).
5. There was no significant difference in participant performance between monitoring conditions (Chapter 4).
6. There was no difference in the grades between the trained and untrained students in the retrospective think-aloud experiment (Chapter 8).
7. There was no concrete evidence of student improvement in DRC best-practices (Chapter 9).

## 10.2 General Discussion

In executing this research project, it was confirmed that great care and attention to detail is required in the construction of a TET program. The design choices outlined in this text were drawn from previous research, teaching experience, professional experience, and student feedback. It was clear through the exit survey and informal discussions that the students noticed and appreciated this time and effort. This likely led to increased motivation and a higher overall skill level when completing the program. Analysis of the student training data and student feedback were particularly valuable in refining the program.

Another "take-away" was that user training data was easy to collect and offered a large number of data points for grading and monitoring of student progress. A simple statistical analysis package should be included with TET software. The training data can also serve as evidence of learning but may not be useful in isolating learning factors or measuring transfer of skills.

In developing standardized listening tests, it became clear that the dependent variable should be measured at the interval level and not the categorical level. Statistical power will often be lacking with typical time constraints (low number of data points) when combined with categorical data. In addition, other variables that might affect performance should either be collected or controlled (Ex: fatigue, stimulus difficulty). These factors were hypothesized as important factors in the lack of significance in the in-context and near-transfer listening tests and the success of the far-transfer listening test.

The Think Aloud experiment was a new technique in the context of TET research and showed great potential for future work. A higher number of data points should have been collected and it would have been interesting to record written feedback from the graders. Summarization can be difficult with a primarily qualitative data set, but the advantage is a great deal of available information. This type of test could be also be useful for instructors evaluating student success in a course.

## 10.3 Impact and Contribution of Knowledge

Not all measurements of learning yielded statistically significant results, however, three of the six experiments were successful and led to the following conclusions. Yes, DRC TET was effective. Yes, the learned skills can be transferred to other tasks. It was also concluded that students improved in their understanding of DRC operation and their ability to detect and quantify audible DRC effects. Learning to use DRCs is an important and difficult step for audio students. It is recommended to include a DRC TET program in beginner audio engineering courses. This type of training could also be useful for a variety of other professionals: researchers training listening test participants, acoustic engineers, electrical engineers, musicians, producers, composers, and others. Finally, this research has shed new light on the perception of various DRC parameters and has proposed two examples of effective DRC algorithm designs. This information could help electronic engineers and DSP engineers create better DRC tools in the future.

## 10.4 Future Work

The design of the *dyntet* training program was a starting place. There is no question that the software application, stimuli, and methodology could be improved for higher learning rates and more effective transfer. Specifically, the next version of the program will include more discrete steps for each parameter, a finer gradation in difficulty levels, more stimuli, a more advanced user evaluation method, exercises involving the compressor knee parameter, and different compressor algorithms. To achieve some of these goals, investigation into task difficulty is required. Some possible factors related to the stimulus may be dynamic range, spectrum, ADSR envelope, crest factor, or loop length. Eventually, the intention is to make the software application available for institutional use.

Further research into the efficacy of DRC TET programs is also needed. The findings reported here need to be confirmed. More data points, for a larger sample of students, over a longer training period could yield a better definition of the learning curve and more information about the intermediate levels of skill transfer. Another interesting approach might be to develop skill testing methodologies that take place over a longer time window, like several hours, days, or even weeks. It is hypothesized that highly refined and subtle evaluations of sound cannot always be achieved in short listening tests. Listeners may need time to “live with” stimuli in known listening environments in order to both recognize the full extent of the audible effects and come to an enduring conclusion.

One of the more interesting questions that resulted from this research is, to what degree does listener fatigue affect performance in critical listening tasks? Also, how can listener fatigue be measured? It is hypothesized that listener fatigue can have a large effect on listener performance in this context. If this is the case, the answers to these questions could be very useful for working professionals, educators, and researchers conducting listening tests. A second interesting question that resulted from this research is, how does a person’s preference evolve over extended listening? Also, how long does it take for preference to stabilize? There is preliminary evidence in this work that preference does indeed change over short periods of familiarization (1-2 hours). Answers to these questions could be very useful for researchers and product designers looking to quantify listening preference.

# 11 References

[1] Ballou, G. (2015). *Handbook for sound engineers*. Burlington, MA: Focal Press.

[2] Bech, S. (1992). Selection and training of subjects for listening tests on sound-reproducing equipment. *Journal of the AES, 40*(7/8), 590-610.

[3] Bech, S. & Zacharov, N. (2006). *Perceptual audio evaluation – Theory, method and application.* West Sussex, England: John Wiley & Sons Ltd.

[4] Bland, J. M., and Altman, D. G. (1995b). Calculating correlation coefficients with repeated observations: part 2 Correlation between subjects. *BMJ* 310:633. doi: 10.1136/bmj.310.6980.633

[5] Breslow, L. (2007). *Methods of measuring learning outcomes and value added.* Retrieved from: https://tll.mit.edu/sites/default/files/guidelines/a-e-tools-methods-of-measuring-learning-outcomes-grid-2.pdf

[6] Brixen, E. B. (1993). Presented at the 94th AES Convention: *Spectral ear training.* Berlin, Germany.

[7] Case, A. (2019). *AES education directory*. Retrieved from http://www.aes.org/education/directory/

[8] Cash-Jones, Laurel. (2002). Presented at the 113th AES Convention: *Finding a recording audio education program that suits your career choice*. Los Angeles, CA.

[9] Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education, 12*(2), 68-82.

[10] Cleveland, W., Grosse, R. & Shyu, W. (1991). Local Regression Models. In J.M. Chambers & T.J. Hastie (Eds.), *Statistical Models in S*. CRC Press: Boca Raton, FL.

[11] Corey, J. (2004). Presented at the 117th AES Convention: *An ear training system for identifying parameters of artificial reverberation in multichannel audio.* San Francisco, CA.

[12] Corey, J. (2007) Presented at the 123rd AES Convention: *Beyond splicing: Technical ear training methods derived from digital audio editing techniques*. New York, NY.

[13] Corey, J. & Benson, D. (2016). *Audio production and critical listening: Technical ear training* (2nd ed.). Burlington, MA: Focal Press.

[14] Corey, J. & Benson, D. (2016). *Technical Ear Trainer*. Retrieved from http://webtet.net

[15] Cowan, J. (2017). The potential of cognitive think-aloud protocols for educational action-research. *Active learning in higher education, 20*(3), 219-232.

[16] Demany, L. & Semal, Catherine. (2007). The role of memory in auditory perception. In Yost et al. (Eds.) *Auditory Perception of Sound Sources* (pp. 77-113). New York, NY: Springer.

[17] Ericsson, K. & Simon, H. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Boston, MA: MIT Press.

[18] Estep, M. & Meng, C. (2013). Presented at the 50th AES Conference on Audio Education: *Feedback frequency identification and reduction primer (FFIRP): Development of a training and application e-book*. Murfreesboro, TN.

[19] Everest, A. (1982). Presented at the 72nd AES Convention: *Instruction in critical listening.* Anaheim, CA.

[20] Everest, A. (1997). *Critical listening and auditory perception.* Vallejo, CA: Artistpro.com.

[21] Everest, A. (2006). *Critical listening skills for audio professionals*. Boston, MA: Cengage Learning.

[22] Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*(3), 385–400.

[23] Gorard, C. (2004). *Combining Methods in Educational and Social Research*. London, England: McGraw-Hill Education (UK).

[24] Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto: Higher Education Quality Council of Ontario.

[25] Huber, D. & Runstein, R. (2017). *Modern recording techniques*. London, England: Taylor and Francis.

[26] Indelicato, M. et al. (2014). Presented at the 137th AES Convention: *How critical listening exercises complement technical courses to effectively provide audio education for engineering technology students.* Los Angeles, CA.

[27] Interactive Audio Special Interest Group. (2011). *Game audio curriculum guideline*. Retrieved from https://www.iasig.org/index.php/projects/projects-menu/11-education-wg

[28] ITU-R BS.775-1. (1994). *Multichannel stereophonic sound system with and without accompanying picture*. Retrieved from: https://www.itu.int/rec/r-rec-bs.775/_page.print

[29] ITU-R BS.1116-3. (2015). *Methods for the subjective assessment of small impairments in audio systems.* Retrieved from: https://www.itu.int/rec/R-REC-BS.1116/en

[30] ITU-R BS.1770-4. (2015). *Algorithms to measure audio programme loudness and true-peak audio level*. Retrieved from: https://www.itu.int/rec/R-REC-BS.1770/en

[31] Iwamiya, S. et al. (2003). Technical listening training: Improvement of sound sensitivity of acoustic engineers and sound designers. *Acoustical Science and Technology, 24*(1), 27-31.

[32] Kaniwa, T. et al. (2011). Presented at the 8th Sound and Music Computing Conference: *Towards a personalized technical ear training program: An investigation of the effect of adaptive feedback*. Padova, Italy.

[33] Kassier, R., Brookes, T., and Rumsey, F. (2005). Presented at the Digital Music Research Network 1st Summer Conference: *A pilot study into listener training for spatial audio evaluation*. Glasgow, Scotland.

[34] Kassier, R., Brookes, T., and Rumsey, F. (2006). Presented at the 121st AES Convention: *A comparison between spatial audio listener training and repetitive practice.* San Francisco, CA.

[35] Kassier, R., Brookes, T., and Rumsey, F. (2007). Presented at the 122nd AES Convention: *Training versus practice in spatial audio attribute evaluation tasks.* Vienna, Austria.

[36] Kassier, R., Brookes, T., and Rumsey, F. (2006). Presented at the 120th AES Convention: *Designing a spatial audio attribute listener training system for optimal transfer*. Paris, France.

[37] Kawahara, K., et al. (2010). Presented at the 20th International Congress on Acoustics: *Case study of curriculum development for technical listening training for employees of an acoustic related company*. Sydney, Australia.

[38] Kim, S. et al. (2013). Inter-subject differences in personalized technical ear training and the influence of an individually optimized training sequence. *Acoustical Society of Japan Science and Technology, 34*(6), 424-431.

[39] Kim, S. & Imamura, H. (2017). An assessment of a spatial ear training program for perceived auditory source width. *The Journal of the Acoustical Society of America, 142*(2), EL201-EL204.

[40] Kim, S. & Olive, S. (2015). Presented at the 138th AES Convention: *Assessing influence of a headphone type on individualized ear training*. Warsaw, Poland.

[41] Kincheloe, J. (2004). *Rigour and Complexity in Educational Research*. New York, NY: McGraw-Hill Education.

[42] King, R., Leonard, B. & Sikora, G. (2010). Presented at the 129th AES Convention: *Variance in level preference of balance engineers: A study of mixing preference & variance over time*. San Francisco, CA.

[43] King, R., Leonard, B. & Sikora, G. (2011). Presented at the 131st AES Convention: *The effects of monitoring systems on balance preference: A comparative study of mixing on headphones versus loudspeakers*. New York, NY.

[44] King, R., Leonard, B. & Sikora, G. (2012). The practical effects of lateral energy in critical listening environments. *Journal of the AES, 60*(12), 997-1003.

[45] King, R. et al. (2014). Presented at the 137th AES Convention: *The role of acoustic condition on high frequency preference*. Los Angeles, CA.

[46] Letowski, T. (1985). Development of technical listening skills: Timbre solfeggio. *Journal of the AES, 33*(4), 240-244.

[47] Letowski, T. & Miśkiewicz, A. (1995). Presented at the 1995 International Congress on Noise Control Engineering: *Development of technical listening skills for sound quality assessment*. Newport Beach, CA.

[48] Letowski, T. & Miśkiewicz, A. (2013). Timbre solfege: A course in perceptual analysis of sound. In J.A. Adamczyk (Ed.), *Signal Processing in Sound Engineering* (pp. 83-95). IPPT-PAN.

[49] Leonard, B., King, R. & Sikora, G. (2013). Presented at the 134th AES Convention: *The effect of playback system on reverberation level preference*. Rome, Italy.

[50] Levine, S., Leonard, B. & King, R. (2013). Presented at the 134th AES Convention: *Binaural room simulation for acoustic testing*. Rome, Italy.

[51] Lewis, C. (1982). Using the 'thinking-aloud' method in cognitive interface design. *IBM Corporation Research Report*. RC-9265.

[52] Liu, Z., Wu, F., & Yang, Q. (2007). Presented at the 122nd AES Convention: *The practice and study of ear training on discrimination of sound attributes*. Vienna, Austria.

[53] Liu, Z., Wu, F., & Yang, Q. (2007). Presented at the 122nd AES Convention: *The training and analysis of listening discrimination of pure tone frequency*. Vienna, Austria.

[54] Marguerite, G., Lodico, D., & Voegtle, K. (2010). *Methods in education research: From theory to practice.* Hoboken, NJ: John Wiley & Sons.

[55] Martin, D. & Massenburg, G. (2015). Presented at the 139th AES Convention: *Advanced technical ear training. Development of an innovative set of exercises for audio engineers*. New York, NY.

[56] Martin, D., Massenburg, G. & King, R. (2017). Presented at the 143rd AES Convention: *Dynamic range controller ear training: Description of a methodology, software application, and required stimuli*. New York, NY.

[57] Martin, D., Massenburg, G. & King, R. (2017). Presented at the 144th AES Convention: *Dynamic range controller ear training: analysis of audio engineering student training data*. Milan, Italy.

[58] Martin, G. (2011). *Introduction to sound recording*. Retrieved from http://www.tonmeister.ca/main/textbook/

[59] Marui, A. & Kamekawa, T. (2013). Presented at the 21st International Congress on Acoustics: *Towards the development of objective difficulty measure in technical ear training tasks*. Montréal, Canada.

[60] Marui, A. & Kamekawa, T. (2018). Presented at the 144th AES Convention: *Does spectral flatness affect the difficulty of the peak frequency identification task in technical ear training?* Milan, Italy.

[61] Massenburg, G. (2019). *The METAlliance report: Critical listening and critical evaluation.* Retrieved from https://www.prosoundnetwork.com/pro-sound-news-blog/the-metalliance-report-critical-listening-and-critical-evaluation

[62] McKinnon-Bassett, M. and Martens, W. (2013). Presented at the 50th AES Conference: *Experimental comparison of two versions of a technical ear training program: Transfer of training on tone colour identification to a dissimilarity-rating task.* Murfreesboro, TN.

[63] Miskiewicz, A. (1992). Timbre solfege: A course in technical listening for sound engineers. *Journal of the AES, 40*(7/8), 621-625.

[64] Miśkiewicz, A. & Letowski, T. (2014). Presented at the Forum Acusticum: *Timbre solfege training in automotive industry*. Krakow, Poland.

[65] Moorefield, V. (2010). *The producer as composer: Shaping the sounds of popular music*. Cambridge, MA: The MIT Press.

[66] Moulton, D. (1993). *Golden Ears: Audio ear training program* [Audiovisual Material]. Sherman Oaks: KIQ Productions.

[67] Neher, T. (2004). *Towards a spatial ear trainer* (Unpublished doctoral dissertation). University of Surrey, Guildford, England.

[68] Neher, T., Brookes, T., & Rumsey, F. (2003). Presented at the 24th AES Conference on Multichannel Audio: *Unidimensional simulation of the spatial attribute 'ensemble depth' for training purposes – part 1: pilot study into early reflection pattern characteristics*. Banff, Canada.

[69] Neher, T., Brookes, T., & Rumsey, F. (2004). Presented at the 116th AES Convention: *Unidimensional simulation of the spatial attribute 'ensemble depth' for training purposes – part 2: creation and validation of reference stimuli*. Berlin, Germany.

[70] Neher, T., Brookes, T., & Rumsey, F. (2006). A hybrid technique for validating unidimensionality of perceived variation in a spatial auditory stimulus set. *Journal of the AES, 54*(4), 259-275.

[71] Neher, T., Rumsey, F., and Brookes, T. (2002). Presented at the 112th AES Convention: *Training of listeners for the evaluation of spatial sound reproduction*. Munich, Germany.

[72] Neher, T., Rumsey, F., Brookes, T., & Craven, P. (2003). Presented at the 114th AES Convention: *Unidimensional simulation of the spatial attribute 'ensemble width' for training purposes*. Amsterdam, The Netherlands.

[73] Nishimura, A. (2013). Presented at the 50th AES Conference on Audio Education: *Ear training for non-technical students using a web-based training system*. Murfreesboro, TN.

[74] Nishimura, A. (2015). Presented at the 138th AES Convention: *Effects of ear training on education on sound quality of digital audio for non-technical undergraduates*. Warsaw, Poland.

[75] Norcross, S., Soulodre, G. & Lavoie, M. (2003). Presented at the 115th AES Convention: *The subjective loudness of typical program material*. New York, NY.

[76] Olive, S. (1994). Presented at the 97th AES Convention: *A method for training listeners and selecting program material for listening tests.* San Francisco, CA.

[77] Olive, S. (2001). Presented at the 110th AES Convention: *A new listener training software application*. Amsterdam, The Netherlands.

[78] Olive, S. (2003). Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study. *Journal of the AES, 51*(9), 806-825.

[79] Pritts, R. (1988). Education and the AES. *Journal of the AES, 46*(1/2), 88-90.

[80] Quesnel, R. (1990). *A computer-assisted program in timbral ear training – a preliminary study* (Unpublished masters thesis). McGill University, Montréal, Canada.

[81] Quesnel, R. (1996). Presented at the 100th AES Convention: *Timbral ear trainer: Adaptive, interactive training of listening skills for evaluation of timbre differences.* Copenhagen, Denmark.

[82] Quesnel, R. (2009). *Timbral ear training: A computer-assisted method for training and researching timbre memory and evaluation skills.* Saarbrücken, Germany: VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG.

[83] Quesnel, R. & Woszczyk, W. (1994). Presented at the 96th AES Convention: *A computer-aided system for timbral ear training*. Amsterdam, The Netherlands.

[84] Reba, C. (2010). Presented at the 38th International AES Conference: *Sound quality education: Pedagogical issues, concepts and practices*. Piteâ, Sweden.

[85] Reiss, J. & McPherson, A. (2015). *Audio effects: Theory, implementation and application*. Boca Raton, FL: CRC Press.

[86] Rościszewska, T. & Miśkiewicz, A. (2015). Presented at the 138th AES Convention: *Timbre solfege: Development of auditory cues for the identification of spectral characteristics of sound*. Warsaw, Poland.

[87] Rumsey, F. & McCormick, T. (2014). *Sound and recording: Applications and theory*. Burlington, MA: Focal Press.

[88] Schaller, T. & Burleigh, I. (2005). Presented at the 26th AES UK conference on Audio Education: *Using web audio for ear training of sound engineers*. Glasgow, Scotland.

[89] Shively, R. & House, W. (1998). Presented at the 104th AES Convention: *Listener training and repeatability for automobiles.* Amsterdam, The Netherlands.

[90] Sievertsen, H., Gino, F. & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences, 113*(10), 2621-2624.

[91] Tough, D. (2010). Presented at the 129th AES Convention: *Shaping audio engineering curriculum: An expert panel's view of the future*. San Francisco, CA

[92] Watson, C. (1980). Time course of auditory perceptual learning. *Annals of Otology, Rhinology & Laryngology, 89*(5_suppl), 96-102.

[93] Woods, R. & Barrow, R. (2006). *An introduction to philosophy of education*. London, England: Taylor & Francis.

[94] W3C. (2015). *Web Audio API*. Adenot, P. & Wilson, C. (Eds.) Retrieved from https://www.w3.org/TR/webaudio/

# 12 Glossary of Terms and Abbreviations

**Audio Engineering Society (AES) –** The Audio Engineering Society is a worldwide professional society focused on audio technology.

**Audio Engineering Technology (AET) –** Audio Engineering Technology is a recently standardized term to describe audio technology fields surrounding audio engineering in the broad sense of the term.

**Buss –** A buss is an audio path inside a mixing console carrying an audio signal from one location to another. It is typically used to describe a mix of source audio signals that have been combined.

**Delay –** A delay is an audio processor the delays the incoming signal by a set amount of time before outputting it. In practice, the delayed portion of the signal is typically added back to the original signal at a quieter volume to create a special effect.

**Distortion –** Distortion occurs when an audio signal's amplitude increases beyond the available headroom in an amplifier. It results in a modification of the output waveform adding higher harmonics that can be described as "color" or "brightness" in their subtle form all the way to "crunchy" or "dirty" in their exaggerated form.

**Dynamic Range Compressor (DRC) –** A dynamic range compressor is an audio processor that reduces the difference between the quietest and loudest parts of a signal, making the overall volume more consistent.

**Equalizer –** An equalizer modifies the balance of a signal's spectrum. Commonly referred to as the "tone control", it can modify a signals bass, mid-range, and treble response.

**Master Buss –** The master buss is the final buss in a mixer that contains a combination of all signals. Audio effects applied to the master buss will affect the entire audio signal as a whole.

**Mastering –** Mastering is the final stage in the audio production process when all media is combined into one cohesive unit, ready for market. It typically involves the application of audio effects to the master buss.

**Mixing** – Mixing is the middle stage of the audio production process when individual source audio signals are combined in appropriate proportions.

**MP3 –** An MP3 is a type of lossy perceptual codec used to significantly reduce the size of audio files. Extreme compression settings often result in audible artefacts.

**Multi-track –** A multi-track recording refers to a recording where all source signals are recorded separately so that they can be adjusted or changed in the future.

**Octave Bands –** Octave bands are a standardized set of frequency bands used to divide up the audible spectrum. They are typically used as reference points on an equalizer.

**Pan Pot –** A pan pot is an audio processor used to spatialize an audio signal continuously from left to right on a stereo playback system.

**Reverberator** – A reverberator is an audio processor used to simulate the acoustic signature of a large variety of indoor and outdoor spaces.

**Technical Ear Training (TET) –** Technical ear training is a type of perceptual learning focused on developing critical listening skills.

**Track –** A track is a signal source audio signal, typically one of many in a multi-track recording.

**Volume Fader –** A volume fader is an audio processor used to adjust the loudness of a signal.