Technical and Clinical Approaches for Implementing a Vision Screening Tool

Cameron Kline-Sharpe

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Computer Science and Software Engineering

University of Washington

2023

Committee:

William Erdly, Chair

Jeffery Kim

Alan Pearson

Program Authorized to Offer Degree:

Computing and Software Systems

University of Washington

Abstract

Technical and Clinical Approaches for Implementing a Vision Screening Tool

Cameron Kline-Sharpe

Chair of the Supervisory Committee:
William Erdly
Computing and Software Systems

Detecting vision problems in children is a challenging task, especially in large populations. This is in part due to the difficulty of obtaining useful indications of vision problems which may cause a child to be sent to an eye doctor. Modern vision screening approaches, intended to solve this problem, are either hard to scale, expensive, or limited in applicability. The aim of this thesis was to continue the development of and clinically test a vision screening mobile application aimed at wide distribution among Washington state school nurses and determine future development and testing plans based on the results of those tests. The QuickCheck vision screening application was tested against the FrACT vision screening application and near vision screening cards using accuracy, specificity, sensitivity, and other statistical measures. Using QuickCheck's best testing policy the application is currently able to detect subjects with any vision problems with an average sensitivity of 0.91 ± 0.11, although performance by eye is much lower (average sensitivity of 0.83 ± 0.06). Accuracy by subject is 0.95 ± 0.07, and again average accuracy by eye is lower, at 0.84 ± 0.09. Additionally, QuickCheck is worse at detecting distance vision

problems (sensitivity 0.83 ± 0.11) than near vision problems (sensitivity 0.875 ± 0.10), although this is more than offset by a corresponding difficulty with detecting a lack of near vision problems as compared to distance vision problems. A modification plan for QuickCheck, including methods to decrease the application's false negative rate for distance vision tests and decrease acuity test time by half was established. However, given the small sample size of the clinical test, further testing of the QuickCheck application is required to demonstrate its effectiveness to a degree allowing widespread distribution. In addition to gathering more data to confirm the results of this work, future research directions include an examination of how the suggested changes improve performance and how well QuickCheck can detect the vision problems of very young (kindergarten-aged) children.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

Thank you to Dr. William Erdly, Dr. Alan Pearson, and Dr. Jefferey Kim for serving on my committee. An additional thank you to Dr. Alan Pearson for helping during data collection efforts and offering time in his busy schedule to discuss clinical designs with me. Thank you also to my family and friends for supporting me through the thesis process.

Chapter 1: Introduction

Children with visual problems may face challenges when learning to read and write, which are key skills for further academic growth and meeting their goals in life. These vision problems, including myopia or hyperopia (near- and far-sightedness), astigmatism (distortion of lens shape), color blindness, strabismus (misalignment of the eyes sometimes preventing binocular fusion), and others can limit children's learning by making reading, following written directions, maintaining focus on classwork, and many other important academic tasks more difficult. Vision problems that hinder academic development are relatively common, as approximately 25% of Washington students have some vision issue that has a negative impact on their education [1]. Fortunately, the most common visual conditions are eminently treatable once diagnosed. Therefore, Washington state law mandates that all public schools shall conduct "distance vision and near vision acuity screening of children" in "kindergarten and grades one, two, three, five, and seven" [2].

Detection is one of the key challenges in the treatment of nearly all forms of visual problems. There are many different reasons for this; for example, visual impairment often occurs gradually, making changes in visual perception hard to notice, even for the individuals who are experiencing significant loss of visual acuity (or other similar symptoms). Furthermore, many begin to lose acuity at a young age, so a significant portion of those who need treatment may be unable to communicate their visual impairment effectively. Vision problems may also have been present from birth, or visual acuity may have never fully developed, so children may be unaware that they have vision problems, as they may never have known what 'good vision' looks like. It

is also possible for an individual to suffer visual impairment in only one eye, which may mask the loss of acuity even in adults.

These factors together mean that loss of visual acuity can sometimes go unnoticed, especially in children. Because vision problems can hinder academic performance, the swift and easy detection of vision problems in children can help improve their long-term academic outcomes. Developing tools that allow for easy identification of vision problems is therefore an important part of improving academic development, particularly among those children with lower socio-economic status, who live in rural areas, or who otherwise have limited access to healthcare.

QuickCheck is a mobile vision screening application designed to quickly and easily identify vision problems in children. The end goal is for QuickCheck to provide easy-to-use, inexpensive, and reliable vision screenings for school nurses to provide for their school children. QuickCheck is not intended to replace a full visual examination, but rather, it is intended to determine if a child needs a visual examination in the first place. However, before distribution can occur, the clinical reliability of the application must be verified; that is, the ability of the application to detect vision problems must be tested.

Prior work on the QuickCheck application had brought it to a nearly testable state, however, several components of the application needed further development before testing could begin. For example, the “results” screen could not be displayed during testing, as that would provide subjects with a possibly incorrect result. Additional modifications included ensuring the security of data generated by clinical tests by implementing an authorization system.

The primary goal of this thesis’ work was in verifying that QuickCheck was able to identify vision problems at the level required by Washington State law for school vision screenings. For

this to occur, QuickCheck was also developed to ensure that the application was ready for clinical testing. Some of the work on QuickCheck (namely, integrating the application with Amazon's Cognito authentication service) was completed with the assistance of Wooyoung Son, a former undergraduate student at the University of Washington, Bothell.

1.1 Project Positioning: Vision Screening in Children

A key challenge in treating many forms of visual impairment is detection [3]. There are many reasons for this, especially when it comes to diagnosing acquired (non-congenital) eye conditions in children. Visual impairment is often gradual, meaning that changes in perceptive ability can be difficult to notice, even for those experiencing these problems firsthand. Children who do not initially develop normal visual abilities can lack a baseline of "good" vision to compare their experiences against and so may not realize that they have vision problems [3] [4] [5]. Also, some conditions, such as limited color vision or convergence insufficiency, can be subtle and difficult to detect, especially because young children may be unable to effectively communicate about the vision problems they have [5]. These problems are compounded by the co-occurrence of visual impairment with other developmental conditions, including learning disabilities and conditions such as cerebral palsy and epilepsy [6].

Because visual impairment is best treated early, is often difficult for a lay person to detect (even in themselves or their children), and can cause difficulties throughout a child's life, vision screening is an important part of pediatric healthcare. Vision screening is so important that it is sometimes mandated by law; as noted above, Washington State requires that public schools conduct vision screens on their students to determine if any students require a vision examination.

A vision examination is different from a vision screening [7]; vision exams typically include various techniques for measuring the health of a patient's eyes in ways other than just measuring visual acuity. These techniques can include taking photographs of a patient's retinas or measuring intraocular pressure. Vision exams also include an examination of visual acuity for the purpose of prescribing glasses or contacts. Unlike vision screenings, an optometrist or ophthalmologist must be the one who performs a vision exam. Additionally, exams often require expensive medical equipment and can usually only be performed within a specialized clinic or hospital setting.

On the other hand, a vision screening test is a fast and easy-to-perform test that determines if a patient is likely to have specific visual conditions as a first step towards diagnosis and treatment. A vision screening may examine visual acuity (the subjective ability for a subject to discriminate between objects they can see) or a handful of other easy-to-test-for conditions which have known (and usually easy to acquire) treatments. Vision screening tests are not diagnostic in the sense that they cannot determine an official diagnosis of a vision problem; thus, they are not intended to provide a prescription for eyeglasses or similar corrective equipment. Vision screenings can also be performed to evaluate whether the current set of vision-correcting equipment in use by a particular person is sufficient. The latter situation is usually easier to address than the former because individuals who already use visual aids (eyeglasses or contacts) are aware that they have vision problems and are thus more likely to notice any additional problems that occur.

Screening tests are used to divide their subjects into those who need further evaluation for a vision problem and those who do not. Therefore, vision screening tests need not represent the 'final answer' about a particular patient's vision or eye health. Instead, screening tests and

techniques are focused on filtering out patients who definitely do not need a vision exam. Standard screening tests are intended primarily to avoid false negatives even if that means increasing the number of false positives. That is, a screening test should first and foremost detect as many people with vision problems as possible, even if that means increasing the number of subjects who are falsely identified as having vision problems to some degree.

1.1.2 Current Methods of Vision Screening

A number of vision screening methodologies are currently in use. The most common of these methods is using optotypes (letters of specific shapes and sizes) through techniques such as Snellen Charts (Figure 1). These charts require subjects to read optotypes from specific distances and use the smallest optotype a person can read as an indicator of their visual acuity. Snellen Charts, and most similar techniques, require that a subject can read English letters to work properly, although similar charts that use shapes or a single rotated letter are also used. These charts are widely used not only because they have been in use for so long (the first Snellen charts were produced in the 1800s), but they are also non-invasive, inexpensive, present no risk to the subject, and are easy to use (provided the subject can read English letters).

However, Snellen Charts are comparatively slow to test with, and do not provide possible reasons why a subject is unable to read letters. For example, if a subject cannot read letters close to their face, that problem could occur due to hyperopia, astigmatism, or presbyopia (an inability to focus on nearby objects due to aging). Additionally, very young children (or those with difficulties communicating) may not be able to make use of a Snellen chart due to an inability to understand instructions or read the optotypes. Therefore, other methods of vision screening are sometimes used to replace or supplement traditional optotype-based techniques.

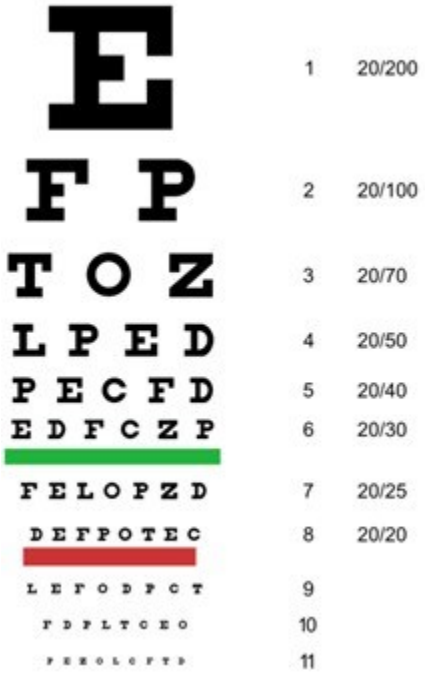

Figure 1: Snellen Chart used for Vision Screening *[8]*

Another category of vision screening tools are autorefractors, devices which automatically detect refractive and other optical errors of a subject's eyes [9]. Photoscreening, a technique commonly used in autorefraction devices, works by taking one or more photographs of the eyes, and using information gathered therein to determine if a subject has any vision problems. Modern photoscreening devices can detect where each eye is looking, distortions in lens shape that may cause refractive errors, and some can even examine nerves inside of the retina to determine if a subject has amblyopia.

As a photoscreener merely needs to take photographs of a subject's eyes, individual photoscreening tests take relatively little time, and do not require the subject understand English letters, both of which are limitations of the Snellen chart (and similar optotype-based vision screening techniques). However, photoscreening devices are far from perfect. They are often

expensive, and therefore difficult to obtain in economically disadvantaged communities, which is where vision exams are often the most needed. Additionally, photoscreening is not a perfect technology; while devices are typically good at detecting whether problems exist, their results often vary quite a bit based on factors unrelated to vision, such as the angle of a subject's head in relation to the photoscreening camera or the ambient light.

1.1.3 QuickCheck in Context

The QuickCheck mobile vision screening application is intended to fill in the gaps left by photoscreening and physical-chart-based screening tests. QuickCheck uses digital optotypes to measure a subject's visual acuity to determine if they need to go to an eye doctor. Additionally, by using the Convergence Insufficiency Symptom Survey (CISS), QuickCheck can determine if any symptoms of convergence insufficiency (CI, which can be a precursor to strabismus and amblyopia) are present [10] [11]. The application is intended to be cheap and easy to use. It is not intended to require expertise or medical training, so users may test on their students (or their children) without requiring medical training. Further, as a mobile application, distribution would be straightforward and inexpensive, thus increasing access tremendously. This is in contrast to photoscreening devices, which are often expensive, and are therefore difficult to obtain in under-resourced communities.

However, these goals require a tradeoff in design: because QuickCheck is designed to be easy to use by many people, there are limitations on the type of information QuickCheck can obtain or determine. For example, detailed information about refractive errors, such as those provided by using a photoscreening device, is beyond the ability of QuickCheck to detect. Furthermore, a single test using the application is not as fast as a photoscreening test, because it requires that a subject read several optotypes rather than just taking a photograph.

1.2 Ethical and Legal Considerations

Because testing the QuickCheck application requires testing on children, certain ethical and legal requirements must be followed. First, approval of an ethics committee (for this thesis, that board is the University of Washington Institutional Review Board (IRB) under the Human Subjects Division) was required before beginning any clinical tests. Additionally, the University of Washington provides guidelines and training for working with children, which were completed along with a background check before clinical testing began. Unfortunately, the training is largely oriented towards those who participate in community programs involving children, such as summer camps. This means that the bulk of the training was not very applicable to this study. However, some of the guidance provided was useful, including information about how to construct a safe environment for children and how to communicate effectively with young people. Rules about how and when physical contact is permitted, as well as rules about safe supervision of children (namely, no adult can be alone with a child) were also applicable to this work. Finally, there are unique concerns about collecting data from minors. These concerns are addressed by the minimal data gathering non-functional requirement of the QuickCheck application as well as the design of clinical trials, in that QuickCheck gathers as little information as possible and cannot gather any personally identifying health information.

In addition to ethical rules determined by the UW IRB, there are legal requirements surrounding data collection and use that QuickCheck follows. As QuickCheck is intended to be used in school screenings in Washington state, it must follow the laws surrounding the types of vision tests to be used on children of various grade levels. Additionally, the key federal laws regulating applications like QuickCheck are HIPAA (the Healthcare Insurance Portability and

Accountability Act) and FERPA (the Family Educational Rights and Privacy Act). HIPAA governs the storage, manipulation, sharing, and use of personal health information (PHI), which includes all medical information that contains personally identifiable information. Any potential subject's PHI is protected under HIPAA, because by its nature, QuickCheck provides a medical service (or, in the context of this thesis, being used in medical testing). FERPA, on the other hand, protects the data of students. This means that FERPA's restrictions can only apply to QuickCheck in some contexts; namely, when QuickCheck is used at a school on students in their capacity as students. The legality and minimal data gathering non-functional requirements of QuickCheck were created and followed to meet the standards set out by HIPAA and FERPA.

1.3 Stakeholders

The specific stakeholders of this project are:

- The EYE Research Group (ERG), the research group headed by Dr. William Erdly, Ph.D., which is responsible for the development and testing of the QuickCheck application.
- The Near Vision Institute (NVI) and Dr. Alan Pearson, OD PhD FCOVD, who are assisting the ERG in developing the QuickCheck application.

More general stakeholders of this project are parents of young children, school nurses, teachers, and others who need or want to know if their children's vision has any problems, but who do not have ready access to vision screening tools such as photoscreening devices. Importantly, one of QuickCheck's major goals is to provide easy vision screening technology to rural and underserved communities. The University of Washington has received several grants from the Seva Foundation so that, in concert with NVI, mobile vision clinic events can be brought to

Native American tribes. While these grants are not directly applicable to the QuickCheck application (as those grants are for NVI's vision clinic, not for QuickCheck itself), one of the Near Vision Institute's long-term goals is to help these tribes and other underserved communities more easily screen their populations using technologies like QuickCheck.

1.4 Contributions, Project Goals, and Criteria

This thesis' main contribution is the initial clinical validation of the QuickCheck application, including obtaining IRB approval, study design, the study itself, and data analysis. Additionally, the development of QuickCheck was completed to bring the application to a clinically testable state.

The project's main goals were as follows:

1. Developing the QuickCheck application until it is in a clinically testable state.
2. Test the QuickCheck application against traditional screening tools, including Optotype-Based vision screening techniques.
3. Analyze the results of the clinical test and determine how well the application performs compared to traditional screening tools. This analysis should examine how well QuickCheck detects vision problems and other factors that affect the use of QuickCheck, such as ease of use and testing speed.
4. Create a plan to improve the QuickCheck application based on the results of the analysis performed as part of goals 2 and 3.

This thesis' criteria of evaluation are:

1. The production of a clinically testable QuickCheck which meets the following non-functional requirements:
    a. Safety and Security – the application does not harm its users. Private data is not exposed to the public, and there is little expected deviation in vision condition prediction from traditional methods of vision screening.
    b. Testability – the application must produce verifiable results associated with an identifiable individual who has also been tested with a traditional screening approach.
    c. Minimal Data Gathering – the application gathers the minimum amount of information required to make a good estimate of vision quality.
    d. Reliability – the application must be reliable enough for several hours of use on several devices.
    e. Ease of Use – the application is easy to use for the purpose of clinical testing.
    NF1. Legality – all tests must follow legal rules around the use of medical information and medical exams. QuickCheck must comply with both HIPAA and FERPA. Further, the application must fulfill the legal criteria for what is required to be covered by a basic school vision screening under Washington state law.
2. An IRB application for testing the application is accepted.
3. A clinical test comparing QuickCheck to optotype-based vision acuity tests is performed.
4. Data gathered in the clinical test is used to determine what changes to make in the QuickCheck application and the next steps to take in future studies.

1.5 Outline of Thesis

This master's thesis is organized as follows: Chapter 2 covers work related to this thesis, and provides a background for understanding vision screening techniques, clinical testing, and application development tools used in detail. Chapter 3 discusses QuickCheck's design, and the design of the clinical test used to verify the application. Chapter 4 reviews the results of the study and suggested changes to the QuickCheck application and future study designs. Chapter 5 is a conclusion summarizing the findings, as well as discussing limitations and next steps.

Chapter 2: Related Works

This chapter provides background on the treatment of visual problems, the use of mobile applications as medical tools, and the clinical testing of applications to determine their validity. This section begins with a discussion of visual problems (2.1), with a focus on visual problems in children and the effects they have on academic development. Second, this section examines the use of mobile applications as medical tools and in medicine-adjacent tasks (2.2). Following that discussion, the clinical testing of medical devices and software products is reviewed (2.3). Section 2.4 covers an overview of clinical testing of software and new screening technologies, and finally section 2.5 reviews legal requirements for vision screening applications under federal and state law.

2.1 Overview of Vision Problems in Children

There are an enormous range of vision problems that can occur in children. Because vision problems can prevent academic development, and the most common visual conditions are usually treatable once diagnosed [3], Washington State law mandates that all public schools shall conduct “distance vision and near vision acuity screening of children” in “kindergarten and grades one, two, three, five, and seven” [2]. The most common problems affecting vision are called refractive errors [6] [12]; that is, errors that occur due to physical and optical properties of the system that directs light towards the light sensitive cells within the eye. In children, these conditions are usually divided into myopia, hyperopia, and astigmatism [12]. Myopia refers to a loss of distance vision (near-sightedness), hyperopia refers to a loss of near vision (far-sightedness), and astigmatism refers to an irregularity in the lens or cornea that causes visual blurriness at all visual ranges [12] [13]. These conditions vary in their severity; on the less severe

end, a diagnosis of "normal hyperopia" means that an individual has optical hyperopia but has normal near vision functionality. That is, while a subject's eyes have optical differences from 'normal' eyes, they have no deficit in visual acuity or function. With more extreme hyperopia, a child may have extreme difficulty learning in a classroom due to a lack of near vision or due to discomfort caused by the effort required to see objects or read at close range.

Another relatively common vision problem is convergence insufficiency (CI), a condition where binocular vision (observing the same object with both eyes simultaneously) is unstable or difficult to sustain at near distances, such as when reading. This can lead to significant vision symptoms such as blurriness or double vision. Such conditions can be progressive and are often most treatable at a young age, which means that early detection is key. Convergence insufficiency in particular can be difficult to detect, as children with CI can often read text at close range for short periods of time (although they often experience discomfort such as eye strain and headaches when doing so). Additionally, because CI cannot be detected without a specialized test, children who struggle with CI in school may simply be thought of as "poor students" rather than as children with a medical problem. Further, CI cannot be treated with normal eyeglasses, and must instead be treated with a series of orthoptic eye exercises. These exercises can be completed at home or in a clinic, although a 2011 meta-analysis of non-surgical interventions demonstrated that at-home exercises were less effective than supervised exercises as measured with the Convergence Insufficiency Symptom Survey (CISS) [14] [10].

Schoolchildren with untreated vision problems can experience a wide range of negative effects on their academic performance. Many of these are easy to imagine. For example, nearsighted children can face challenges in school due to an inability to understand a teacher's writing on a

whiteboard (or similar front-of-the-room displays); farsighted children can find learning to read difficult when they cannot discriminate words in their book or when they avoid near vision activities due to discomfort [15]. More subtly, poor vision can be mistaken for (or be disguised by) other conditions that affect academic performance. For example, although they are distinct conditions, students may mistakenly be thought of as having dyslexia or dysgraphia when they actually have vision problems, or vice versa; a situation which can only be detangled by a vision exam [16]. To make matters more complicated, vision screening or exams which rely on reading letters or words can be challenging to implement in young children or children who have difficulty reading English, such as children with dyslexia or who are English language learners. Additionally, disruptive behavior patterns and vision problems can be associated with each other. This can be due to a variety of causes, but of most relevance for this work is a lack of ability to engage in lesson plans due to difficulty reading at close distances or follow written instructions. Of course, some of these behavior patterns may be due other conditions affecting behavior, such as ADHD (which is correlated with vision problems), and may therefore not be resolvable through treatment of visual errors [17]. Whatever the reason, it is well established that vision problems are associated with decreased academic performance, and that prompt detection and treatment of said problems is key to learning in many children [4].

Reported rates of refractive visual problems vary depending on measurement location and exact methodology used [18]; for example, a 2013 study in the United States found a 42.2% rate of myopia in children aged 10 to 15 (n=370) [19], but a 2015 study on 9,884 Indian schoolchildren (mean age 11.6) found a rate between 12.5% and 13.8%. Some of this variation is due to difference in the proportion of children screened for vision problems, and some is due to actual differences in rates of occurrence; for example, it is well known that some types of visual

problem (such as myopia) are more common in Asia compared with Europe [18]. Additionally, the reported rate of vision problems has increased over time, and this is believed to reflect a genuine upward trend in the number of children with vision problems, particularly myopia [20].

2.2 Manual Vision Screening Techniques

Section 2.1 covers the use of 'manual' vision screening techniques, that is, techniques that use no computers to screen subjects. These techniques are the longest standing ways of performing vision screenings available. Of these, a particularly widespread way of measuring visual acuity is the use of optotype charts, where a screening subject is asked to read letters of specific sizes from a chart or poster. This body of techniques is what the near and distance vision acuity tests in the QuickCheck application are based on. These techniques often use vision charts, or collections of optotypes (letters) of various sizes that are read by a subject. Different charts use different optotype sizes, dimensions, and letters. HOVT charts (that is, charts where only the letters H, O, V, and T are used) or LEA symbols (where subjects are asked to distinguish between symbols rather than between letters) are a common way to measure the acuity of individuals who do not read English fluently.

There are generally two types of techniques used to measure visual acuity: thresholding tests and criteria-based tests. Thresholding tests generally present a large number of sub-tests (e.g., various optotypes of dynamically controlled sizes), and use statistical techniques to determine precisely what visual acuity a subject has. For example, a threshold-based test could determine that a subject has exactly 20/34 vision, not 20/33 or 20/35. On the other hand, criteria-based tests provide a fixed number of fixed sub-tests and can only determine a subject's visual acuity based on the sub-tests that are present. For example, if a criteria-based optotype test only uses

optotypes corresponding to 20/20 vision, that test could determine at most that a subject has better or worse vision than 20/20.

2.2.1 Snellen Charts

The measure of visual acuity through optotype charts became widely popular after Dr. Hermann Snellen published his visual acuity charts in the early 1860s [21]. In addition to Snellen charts, other systems such as the tumbling Es chart and HOVT charts are popular ways to measure visual acuity. In the modern era, Snellen charts are still (with some modifications) a very common way to measure the acuity of distance vision. As discussed in Section 1.1.2, each optotype size in a Snellen chart (see section 1.1.2 Figure 1) corresponds to a different level of visual acuity, with larger optotypes corresponding to lower levels of distance vision. The chart is used by asking the subject (the individual receiving a vision screening) to read from the chart at a specific distance (usually 10 feet) until the smallest size of letter they can read is discovered. This means that Snellen charts are criteria-based tests.

Using a Snellen chart system, the metric of visual acuity is an estimate of how far away a person with healthy vision could see the smallest letter a subject can read. For example, if a subject has 20/100 vision, that means that the smallest letter they can read from 20 feet could be read at 100 feet by an individual with healthy vision. Letter sizes are determined by the size of the arc they subtend (that is, if you drew lines from the top and bottom of the optotype to the center of your eye, the angle the two lines meet at would be a set number of degrees). Under this system "normal vision" (20/20) is defined as being able to read a letter subtending 5 arcminutes (5/60ths of a degree) at 20 feet (6 meters) [22]. Using Equation 1 (h representing optotype height, d representing the distance an optotype is read at, and theta is the subtended angle), this means that a standard 20/20 optotype read at 20 feet must be approximately 8.75mm tall.

$$h = 2\left(d * \tan\left(\frac{\theta}{2}\right)\right) \quad \text{[Equation 1]}$$

For larger optotypes (representing progressively less sharp vision), the angle they subtend grows progressively larger, although the distance stays the same. In equation 1, this represents increasing theta (θ), but fixing d. As Snellen optotypes are square, the height and width of each optotype is the same, so only the height needs to be calculated.

2.2.2 LogMAR Charts

In addition to the system used by Snellen charts, another popular optotype chart is the LogMAR chart (Logarithm of the Minimum Angle of Resolution chart). Rather than estimating the distance at which the smallest letter a subject can see could be seen by a person with 'healthy vision', the LogMAR system instead uses the logarithm of the subtended angle of the smallest letter a subject can read to measure visual acuity (Equation 2).

$$\text{LogMAR} = \log_{10}(\theta) \quad \text{[Equation 2]}$$

In Equation 2, theta (θ) represents the smallest subtended angle of an optotype that a subject can read. A Snellen score of 20/20 (or 6/6 in meters) corresponds with a LogMAR score of 0; LogMAR scores above zero therefore represent 'worse than healthy' vision. There are a variety of charts used to estimate a subject's LogMAR score, including the ETDRS (Early Treatment Diabetic Retinopathy Study) chart shown in Figure 2 [23]. Note that the ETDRS chart uses rectangular optotypes with a height to width ratio of 5:4, although this is not true of all LogMAR charts.

Figure 2: Early Treatment Diabetic Retinopathy Study's LogMAR Optotype Chart *[23]*

2.2.3 Near Vision Cards

While Snellen and other vison charts are useful for determining vision acuity at a distance, near vision acuity requires a different set of tools to determine. There are a number of near vision acuity measurement tools, but one of the most common are near vision cards. These cards present a smaller vision chart on a small card (usually slightly larger than an index card, though sizes vary). Figure 3 displays one such card that uses Lea symbols to allow those without the ability to read English letters to be tested. These cards must be kept a fixed distance (usually 40cm/16in) from the face, which can be challenging, especially when test subjects have difficulty reading from the card. These near vision cards are therefore sometimes accompanied by a string which is used to maintain the card's distance from the subject's face by asking the subject to hold one end of the string by their eyes, then keeping the string taut as the subject is tested. There are a large variety of near vision cards using various types of vision charts to measure near vision acuity [24] [25] [26].

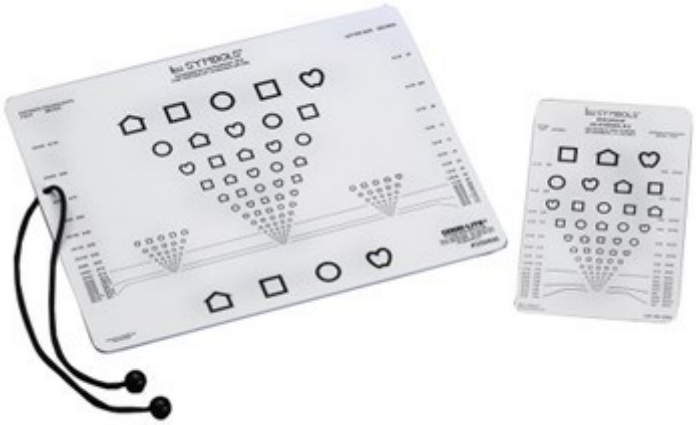

Figure 3: Near Vision Card With Lea Symbols (left: front) *[24]*

2.2.4 Non-Acuity Vision Problem Detection

There are a variety of other vision problems that might be tested during a vision screening. See Table 1 for a list of a few of the more common vision problems that might be tested in a typical screening. The number of these conditions that are actually tested for during a screening depends on the available time, age of the subject, expertise of the screener, and many other factors.

Condition	**Description**	**Screening Test**	**High Risk Subjects**
Convergence Insufficiency	Inability to focus both eyes on a nearby object simultaneously	Convergence Insufficiency Symptom Survey (Table A.B.1)	Young Children

Condition	Description	Screening Test	High Risk Subjects
Strabismus/Amblyopia	Eyes do not look in the same direction	Autorefractors (§ 2.3.2), light reflex, prisms	Young Children
Color Blindness	Cannot perceive the usual range of colors	Color vision tests	Men of all ages
Stereoscopic Problems	Inability to see in 3D through binocular fusion (stereopsis)	Stereoscopic tests (Figure 5)	Children
Cataracts	Cloudy eye lens distorts vision	Examination of the eye	Elderly

Table 1: Example Non-Refractive Vision Problems Tested in Vision Screenings

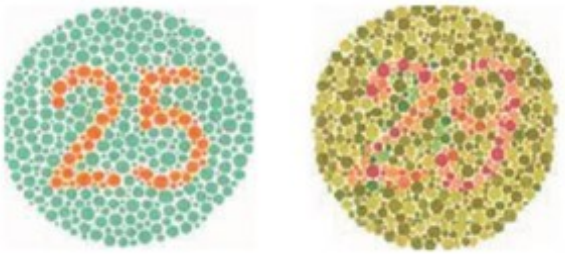

Figure 4: Ishihara Tests for Color Blindness *[27]*

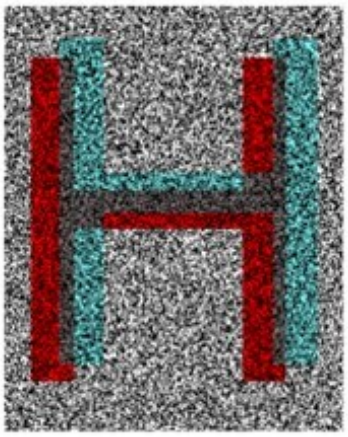

Figure 5: Red-Blue Stereoscopic Test

2.3 Vision Screening Devices and Applications

Vision screening performed by non-medical experts often involves the use of tools which are designed so that people who are not medical doctors can still perform vison screenings. This commonly involves devices which automatically perform the process of discovering refractive error in vision, called Autorefractors or Autorefractor devices. Other applications adapt more traditional optotype-based techniques, such as the FrACT screening application [28]. An alternate avenue of vision screening tools is to adapt more common devices to a medical purpose, usually a smartphone. Vision screening using these devices can be performed by school nurses, by concerned teachers, or by professionals brought into the school by outside actors [3] [29]. Training testers can improve their reliability when it comes to detecting vision problems; however, these tools are designed to allow non-experts to increase screening accuracy and the range of conditions which can be detected in a vision screen without extensive training [29].

2.3.1 Crowding Bars

One, perhaps unexpected, result of using usual vision screening charts is the “crowding effect”, a phenomena where subjects with eye conditions such as amblyopia experience greater difficulty reading groups of adjacent letters than they do reading letters which are distant from each other [30]. This effect is an important part of what a typical vision chart measures, so electronic tools which attempt to use optotypes to measure visual acuity may require the use of placeholders to measure this effect when optotypes are displayed one at a time. One technique used to create the conditions for the crowding effect to take place is the use of crowding bars, or bars which surround an optotype (Figure 6). Although these bars do not cause the effect to occur as strongly as when letters are placed in close proximity to each other, they are still a useful way to measure the crowding effect [30].

Figure 6: Crowding Bars around an "O" Optotype

2.3.2 Autorefractor Devices

Photoscreening, the process of examining the refractive properties of the eyes though the use of a specially designed camera, is often used as a method of autorefractoring. This is because photoscreening requires only that a small number of special ‘photographs’ of a subject be taken to determine if they have refractive vision problems. As refractive errors make up the majority of vision errors even in children, photoscreening devices are a useful way to perform vision

screenings quickly. However, photoscreening measures different vision features than using optotypes. Vision charts measure the subjective experience of discriminating between and recognizing objects in the visual field, while photoscreening measures the exact way light passes through a subject's eyes. The Eye Research Group has anecdotal evidence (from an informal survey of Washington State school nurses at bi-annual meetings of the school nurse association of Washington) that most Washington schools are adopting photoscreening devices to measure the visual health of their students because they are fast and relatively reliable.

An example of a photoscreener is the Spot Vision Screener (SVS), see Figure 7 [31]. Designed to eliminate the difficulty for non-experts to detect vision conditions, even for relatively subtle conditions such as some forms of strabismus [32], these tools have the upside of ease of use and (relative to the baseline of non-medically trained testing personnel), accurate performance, even in young children [32] [33]. Furthermore, photoscreening tools like the SVS can provide automated recommendations to those they test, allowing for more accurate and in-depth suggestions to parents of children with poor vision. Finally, the SVS contains useful features for organizing and storing data about vision screenings, which allows it to serve as an 'all in one' vision screening device for schools.

Figure 7: A Welch Allyn Spot Vision Screener

However, photoscreening tools are usually expensive; the spot vision screener often costs over $6,000 per device [34] [35] [36]. Additionally, even though using a single device on a single child is fast (usually about as fast as taking a photograph, plus set up time [32]) using a single device on a large population of children can be a time consuming affair if the physical devices are limited in number. If a school district does not have the resources to acquire a large number of vision screening devices to test children in parallel, they may be required to use older technology or techniques, which may be less reliable than a photoscreening device. This is especially true of schools in rural or less wealthy areas, where vision screening is rarely a top priority when it comes to the allocation of school funds [37]. However, it is often in exactly such situations that providing vision screening and optometric healthcare is most needed and most impactful.

2.3.3 Vision Screening Applications Overview

Devices built for the explicit purpose of being employed in vision screenings are just one part of the wide world of vision screening tools; software applications designed for vision screening (which can be run on different types of devices) are also common. Unlike the SVS and other purpose-built autorefractors, such applications are usually not meant to be used in a vision screening event for a large number of subjects at once, as in school-side vision screenings. Instead, such applications are usually meant to allow an individual to, in their own time, test their visual acuity. These applications have a number of different purposes and examining their designs with their purposes in mind granted insight into how to adapt the QuickCheck application for clinical testing, as well as how to adjust the design with insights from said clinical tests. While applications for vision screenings are most applicable to this thesis, a brief examination of other types of healthcare applications is also provided in section 2.3.4.

2.3.3.1 The FrACT Vision Screening Application

The FrACT (Freiburg Vision Test) Vision Screening Application is a web application that uses a Bayesian approach to determine a subject's vision acuity (measured with LogMAR and standard Snellen scores) [28] [38] [39]. The most recent version of FrACT is FrACT_{10}, shown in Figure 8, which was last updated in February of 2022 as of writing [28].

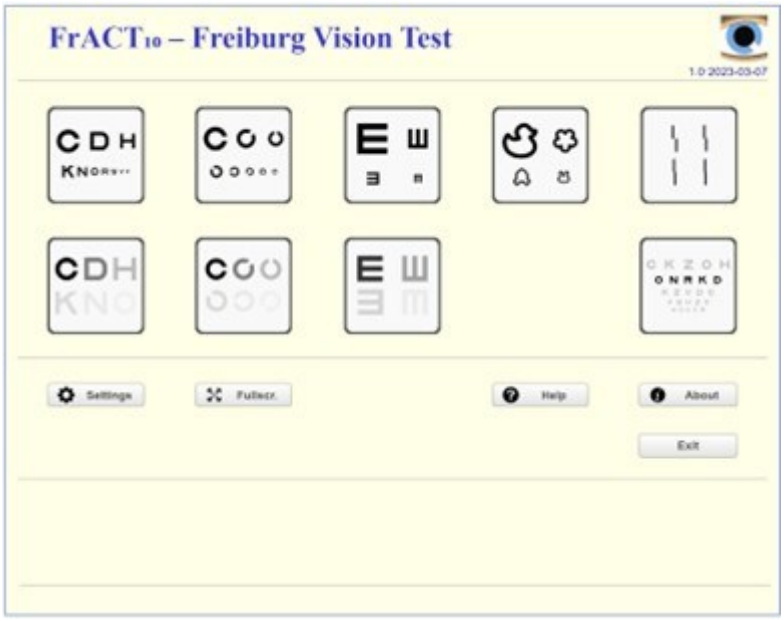

Figure 8: FrACT Testing Menu Displaying Various Test Types *[28]*

There are several different test types implemented as part of FrACT$_{10}$. Acuity Runs (ARs), which measure visual acuity, have the following general format:

1. The device running FrACT (which can be any platform with a sufficiently large screen, smartphones are too small) is placed a certain distance from the subject and set to full screen mode.
2. The subject is presented with a single optotype or symbol on the screen at a time. The subject attempts to read the optotype or symbol, and the tester enters their guess into the system. Typically, this means typing the guessed letter on a keyboard, but different tests have different guess types.
3. After the subject's guess is entered, an optotype/symbol is again presented to the subject, but this optotype/symbol is now larger or smaller than the first one.
4. Step 3 is repeated with varying optotype/symbol sizes until FrACT$_{10}$ has determined a subject's vision acuity.

Contrast Runs (CRs) have a very similar format, except that they measure the subject's ability to distinguish an optotype or symbol from the background. For CRs, the contrast between the letter and the background is what is changed, not the optotype size as in ARs. Because FrACT can dynamically change optotype size and use statistical techniques to precisely determine a subject's vision, it is a threshold-based test. See Table 2 for a list of tests available in $FrACT_{10}$.

Test Name	**Description**	**Subjects Guess**	**Purpose**
AR Sloan Letters	Sloan optotypes of various sizes are displayed one by one	The letter displayed	Standard vision acuity test
AR Landolt Cs	Landolt Cs (Cs with various rotations) of various sizes are displayed one by one	The C's rotation	Useful for those who cannot read
AR Tumbling Es	Tumbling Es (Es with various rotations) of various sizes are displayed one by one	The E's rotation	Useful for those who cannot read
AR TAO*	Symbols (outline of a duck, helmet, flower, and rabbit) of various sizes are displayed one by one	Which of the four symbols is displayed	Useful for young children
Hyperacuity Run	Two bars are displayed on top of each other; the top bar is offset from	The direction which the bar is offset	Very sensitive measure of acuity

Test Name	Description	Subjects Guess	Purpose
	the bottom in various directions by various amounts		
CR Sloan Letters	Sloan optotypes of various contrasts are displayed one by one	The letter displayed	Standard contrast acuity test
CR Landolt Cs	Landolt Cs of various contrasts are displayed one by one	The C's rotation	CR useful for those who cannot read
CR Tumbling Es	Tumbling Es of various contrasts are displayed one by one	The E's rotation	CR used for those who cannot read
Lines of Optotypes	Optotypes of various sizes are displayed line by line	The letters displayed	Standard vision test showing multiple letters simultaneously

Table 2: FrACT$_{10}$ Vision Test Types

*TAO: The Aukland Optotypes, a set of four symbols

FrACT was named one of the “Recommended Methodologies for Assessment of Visual Acuity” by the HOVER (Harmonization of Outcomes and Vision Endpoints in Vision Restoration Trials) taskforce [40]. It has been cited in over 1300 studies over the last two decades, including in trials of subretinal electronic chips which allow blind patients to read letters [28] [41] [42].

2.3.3.2 Warby-Parker’s Virtual Vision Test

The Warby-Parker “Virtual Vision Test” (VVT) application (Figure 9) is designed by Warby-Parker (a large eyeglass and contact retailer) to allow users to test their vision using their glasses [43]. This application is used to verify that the current prescription of the user is still valid. The application flow is as follows: the user starts the application and provides information regarding themselves, then, the user performs tests of visual acuity with their glasses (or other corrective equipment) on. Finally, a medical professional evaluates the results remotely, and within 48 hours the user is told whether or not they need a new prescription (incurring a fee of $15).

Figure 9: Warby-Parker Vision Screening Application Introductory Screen

The VVT application employs a digital version of optotypes, or letters of fixed size used in vision screenings and exams; see Figure 10 for an example of what these optotypes look like. The use of optotypes instead of auto-refracting means that the Warby-Parker results must be interpreted by an expert. In other words, the choice of vision screening technique forced other design choices.

This is unlike standard vision screening tools in a few different ways. First, there is no second party using the application; the subject is both tester and tested. Secondly, the results generated by the tests are not evaluated in place by the application and displayed as easy to understand results; the results generated by VVT are evaluated by an eye doctor remotely. Additionally, those using VVT are already aware that they have visual problems, unlike many who undertake a standard vision screening (especially children). Together, these factors make VVT useful for individual adults who want to test their eyes on their own time, but less useful as part of a vision screening event or for use with children.

Figure 10: Warby-Parker VVT Optotypes *[43]*

2.3.3.3 GoCheck Kids

Similar to VVT, GoCheck Kids (GCK, see Figure 11) is a mobile application; like the SVS, GCK provides vision screening through photoscreening technology. In that sense, GCK attempts to combine the portability of a smartphone application with the speed and ease-of-use of the SVS. However, unlike VVT and QuickCheck, GCK is marketed as a cheaper alternative to photoscreening tools like SVS for pediatric clinics specifically [44] [45]. This means that while GCK is a mobile application, it is not intended for widespread distribution to individuals. Instead, it is intended to serve as a tool for clinicians to use on their patients in situations where autorefractor devices like the SVS are not available or are too expensive to obtain.

Unfortunately, as GCK is proprietary, and use requires a verified account, a deeper analysis into the design of the application was not possible. However, clinical tests of the application have been performed to test GCK's performance, and GoCheck Kid's also self-reports some details of their application's features and performance [46] [47]. For example, they claim that their specificity is 91% and their sensitivity is 81% (compared to the SVS at 85% specificity and 80% sensitivity), although what exactly these statistics mean is impossible to know without more context than is readily obtainable [48]. Similarly to the SVS, GCK reports that they have a "full digital workflow", which includes integration with a practice admin portal and electronic health records (systems which medical professionals use to manage patient information) [48] [49].

Figure 11: Go Check Kids Screening Tool Testing Screen (left) *[45]* and Results Screen (right) *[44]*

2.3.3.4 Summary of Vision Screening Applications

The three vision screening applications reviewed differ from the QuickCheck application in both structure and goals. Table 3 shows these differences on several key points: the application's purpose (what it was designed to do), its userbase (who it was designed to be used by), the medium (what method it uses to distribute itself), the cost to use, and any limitations each application has.

Comparison Criteria	**QuickCheck**	**FrACT** (§2.3.2.1)	**Warby-Parker VVT** (§2.3.2.2)	**GoCheck Kids** (§2.3.2.3)
Purpose	Test if vision exam is needed	Measure visual acuity	Check if new glasses are needed	Cheaply detect and measure refractive errors

Comparison Criteria	**QuickCheck**	**FrACT** (§2.3.2.1)	**Warby-Parker VVT** (§2.3.2.2)	**GoCheck Kids** (§2.3.2.3)
Userbase	School nurses, teachers, etc.	Medical professionals, researchers	Everyone, but requires medical professionals to evaluate results	Pediatricians
Test Methods	Optotypes, CISS	Optotypes	Optotypes	Autorefractor via photoscreening and optotypes
Medium	Mobile App	Web App	Mobile App	Mobile App*
Cost	Free	Free	$15 per use	$129 per device per month
Limitations	No specific acuity measurement	Designed for medical professionals, cannot detect convergence insufficiency, time consuming tests,	Cannot determine vision status without external aid	Available only to medical professionals, subscription fees

Comparison Criteria	**QuickCheck**	**FrACT** (§2.3.2.1)	**Warby-Parker VVT** (§2.3.2.2)	**GoCheck Kids** (§2.3.2.3)
		cannot be used on phones		

Table 3: QuickCheck and Vision Screening App Comparison

*GCK runs on iPhones but delivers those phones to their subscribers

The primary difference is that QuickCheck is not intended to present its users with any in-depth information about the kind of vision problems they have, aside from the distinction between near acuity, distance acuity, and convergence insufficiency symptoms. Additionally, unlike the other mobile applications, QuickCheck is free to use. Finally, unlike FrACT, QuickCheck measures symptoms of convergence insufficiency and can be run on mobile devices.

2.3.4 Non-Vision Screening Healthcare Applications

In general, applications tend to fall into three categories: applications used to create a treatment schedule or manage health information, applications used to assist in gathering and storing biometrics, and applications used to address condition-specific needs (like those considered in sections 4.2.1 – 4.2.4). While many of these applications are not perfectly analogous to the adaptation of the QuickCheck application to clinical trials and the improvement of QuickCheck after said trials, there were still valuable lessons that the design and use of these applications had to teach. However, though some review was performed of Android applications, the bulk of the review in this section was performed on mobile applications available on the Apple App Store. Finally, many healthcare applications make use of proprietary tools, designs, or techniques. This

meant that analysis of many commonly used applications (including those requiring paid subscriptions, such as Weight Watchers' mobile app) could not be analyzed in depth.

Mobile applications built to manage calendars, plan out schedules, and build to-do lists are exceedingly common, and applications for managing healthcare information are nearly as abundant. These applications come in many different forms: some are specific to certain areas of medicine, such as applications used to track food intake, like MyPlate (Figure 4) [50]. Others provide tools to manage the (seemly endless) documentation produced by interactions with the healthcare system. For example, the Virginia Mason "Virtual Mason" application manages appointments with doctors and provides information regarding filling and taking prescribed medications [51]. These applications differ from QuickCheck in several ways: they are generally meant to be used continuously over time, they do not usually provide diagnostic information, and they are intended to be used by a single person. However, examining their design (and the reviews provided by their users) can still provide insight into the successes and failures of healthcare related mobile applications.

One important lesson is the relative importance of various functional and non-functional properties in the context of mobile applications. Based on reviews of the application, users of MyPlate focus on the applications' user interface, touching on the ease-of-use non-functional requirement. On the other hand, users of Virtual Mason, while there are some complaints about the application's UI, focus more on the safety, scalability, and reliability non-functional requirements, with worries about data security, complaints regarding network connectivity, and ability to retrieve important medical information reliably all being common. These are just two examples of a wider trend in medical applications: applications that proport to address healthcare

concerns require a greater level of trust than other types of mobile software. For QuickCheck, this means that design choices that increase user trust, such as an emphasis on the software's reliability when considering tradeoffs or presenting information about the diagnostic ability of the application to the user, are all important considerations that need to be taken into account when improving the application based on clinical data.

2.4 Clinical Testing of Software

Real-world testing of software occurs for a variety of reasons. Testing is often done to determine a product's readiness for deployment, such as tests that ensure an application's backend can handle the expected user load before deployment. But in the world of software used in healthcare applications, there is another reason for testing: clinical verification of a product's usefulness and safety as a medical or diagnostic tool. Luckily, QuickCheck provides a non-invasive screening of a subject's vision, and so does not require the same level of scrutiny that, say, software used in surgery might. However, verification of QuickCheck's ability to detect vision problems is still required if it is to be used in school screenings. Additionally, legal requirements about healthcare information needed to be addressed before the application could be tested at all. Therefore, this section (2.4) focuses foremost on the design of clinical studies and statistical methods used to verify the validity of software similar to QuickCheck.

2.4.1 Testing Vision Screening Devices

In the field of medicine, the highest level of statistical verification is generally thought to be the double-blind randomized control trial (RCT) with a large, representative study population. However, the style of testing where a control is administered in place of treatment is often impractical, unethical, or inapplicable to a specific tool or medicinal device being tested. In the

case of QuickCheck, there was no control that could be utilized to provide an unbiased, perfectly accurate ground truth against which the application's performance could be compared. Instead, the application needed to be compared to the screening tools and techniques that do exist, which means that instead of testing QuickCheck's ability to detect vision problems, any study testing QuickCheck is measuring the application's ability to predict what using some other tool (e.g., near vision cards or FrACT) will result in.

Fortunately, this is not a new problem for the medical field. Often, new diagnostic tools must be compared, not to some absolute truth, but to the results of imperfect historical tools. Therefore, examining the study design and statistical techniques used by past clinical tests can provide guidance for designing a test for QuickCheck that will convincingly verify the application's ability to detect vision problems.

2.4.1.1 Case Study 1: The Plusoptix Screener Study Design

The first study examined was "Vision screening in children by Plusoptix Vision Screener compared with gold-standard orthoptic assessment" by Dahlman-Noor et. al. [52]. The Plusoptix vision screener (PVS) (Figure 12), like the spot vision screener discussed in Section 2.3.2, is an autorefractor photoscreening device designed and produced specifically for vision screening in children. The study aimed to establish whether or not the Plusoptix vision screener produced clinically actionable results, as well as the types of vision problems it could successfully diagnose. The core design of the study is as follows:

0. The author's obtained "ethics committee approval" [52, p. 342].
1. The parents of children aged 4-7 were asked to allow their children to participate in the study; with the aim of gathering 300 participants, 379 were invited and 288 took part.

2. Each study participant was tested with the PVS, and distance acuity testing, cover tests, extraocular movement measurements, and a prism test were performed by an orthoptist who did not know the results of the PVS screening.
    a. If certain criteria were met, subjects were referred to a hospital for further pediatric care, including manual cycloplegic refraction (MCR), the most accurate procedure used to measure refractive errors [52, p. 343].
3. The authors performed statistical comparison of the traditional methods against the PVS results, including testability and calculating the spherical equivalent of the measurements obtained from PVS and MCR.

Figure 12: Plusoptix Vision Screener Front Screen (top right) and Back View (bottom left) *[53]*

This case study gives several key points of insight: first, the statistical methods used are relatively straightforward. Comparing the results from PVS to those performed by an orthoptist was performed via comparing sensitivity (ability to designate an individual with a vision problem as having a vision problem) and specificity (ability to designate an individual without

vision problems as not having vision problems). However, several of the statistical calculations used are not relevant to the QuickCheck application. For example, calculating spherical equivalence is not possible for QuickCheck, as the application does not provide estimations of level of visual impairment. Additionally, comparing against a gold standard (MCR in this case) were not possible; as previously noted, this presents a strong limiting factor on the ability to verify QuickCheck's performance, because there is no "ground truth" to compare against.

A final consideration is the pre-calculation of several statistical benchmarks given the number of participants in the study. With 300 participants, given the occurrence rate of vision problems in the study population, the authors estimated a maximum precision for sensitivity (at a 95% confidence level) to be ±5% for specificity and ±16% for sensitivity. Performing this sort of pre-study calculation provides a framework for understanding the scope of possible results before analysis of the results begins and is a good way to improve confidence in the results of a study.

2.4.1.2 Case Study 2: Google Play Snellen-Chart-Based Vision Screening Application Testing

In "Validation of the Smartphone-Based Snellen Visual Acuity Chart for Vision Screening", Gupta et. al. examined 10 different smartphone applications which test visual acuity using vision charts [54]. The goal of this study was to identify and validate smartphone vision screening applications on the google play store. The methodology of the study was split into phases: first, researchers searched through the Google Play store for applications which test vision and filtered those applications for those which could be studied. In Phases II through IV, the smartphone applications selected for testing were validated against a standard ETDRS (Early Treatment of Diabetic Retinopathy Study) chart (see Figure 2 for a similar vision chart).

The study progressed as follows:

- Phase I consisted of the "exploration, screening, review, and calibration" of the vision screening applications selected for the study [54].
    - The initial exploration found 1,396 non-unique search results in the Google Play Store using a variety of terms related to vision screening (e.g., vision screening, vision test, LogMAR chart, etc.).
    - After duplicates were removed and application names were screened for applications that were likely to actually test vision acuity, the number of applications set for further investigation was 167.
    - Following a deeper review, including an investigation of language used (only applications in English were accepted), application purpose, and testing methods, only 10 applications were considered for inclusion.
    - Of the 10 applications considered for inclusion, all but one was excluded. Exclusions were for a variety of reasons, including the use of non-Sloan optotypes and a lack of detail regarding optotype size.
- Phase II was the testing of the luminance of the testing room and visual acuity chart, measured with a digital lux meter. This was performed because the application screen is 'self-illuminated' (lit from within) and the ETDRS chart used needed to be illuminated manually, so comparing the two approaches could lead to a false comparison (e.g., if the room was too dark, the application could outperform the ETDRS because the latter was difficult to read).
- Phase III consisted of the first test of the selected application ("Snellen Chart") against the ETDRS. Subjects who met the inclusion criteria (LogMAR vision acuity of at least 1.0) were tested with both approaches on the same day by different researchers.

- Phase IV was a test-retest repeatability experiment; a subset of the subjects in phase III were re-tested with the application and the ETDRS to determine if the results of Phase III could be reaffirmed during a repeat trial.
- The results of Phases III and IV were analyzed using the limits of agreement, mean difference, and the coefficient of repeatability.

This study presents several important lessons in terms of study design and statistical analysis. In terms of study design, this study shows the importance of controlling as many factors as possible to compare the reference measure of visual acuity and the QuickCheck application's measure. Although not all of the methods used to control external conditions can be used in this study (for example, this thesis does not have access to a high-fidelity luminance measurer), the strategies used for controlling external factors could be adapted for purpose in this thesis.

In terms of statistical analysis, the use of the limits of agreement is useful in terms of comparing two different measures of visual acuity. The limits of agreement estimates the interval within which a given proportion of differences between measurements lie, and includes information both about random errors (low-precision errors) and more systematic mistakes (bias in a measurement technique) [55]. In other words, this statistical measure is useful because, when comparing the 'agreement' between two measuring techniques where one is treated as a reference, the limits of agreement can be used as a way of measuring the total error of the non-reference measurement method. Unfortunately, this method is not usable for QuickCheck, because it relies on continuous rather than categorical predictions (and QuickCheck produces a binary outcome). Therefore, a different measure of agreement must be used.

As QuickCheck produces a categorical (binary) outcome, one method of measuring the agreement between it and a reference vision acuity measurement technique is Cohen's Kappa [56]. This measurement was introduced by Dr. Jacob Cohen in 1960 to overcome the limitations of the then-standard percent agreement metric, which did not take into account agreement caused by random chance [57]. Cohen's Kappa ranges from -1 to 1, with numbers closer to 1 indicating stronger agreement. Cohen's original paper determined that, between two 'raters' (measurement techniques), kappa of 0.40 to 0.60 indicate measurements that are moderately in agreement and kappa of 0.60 to 0.80 indicate measurements that are strongly in agreement [57]. However, later work has called this into question, with McHugh arguing in 2012 that "accepting 0.40 to 0.60 as "moderate" may imply the lowest value (0.40) is adequate agreement" [58]. Instead, a higher standard of 0.60 and above indicating agreement was introduced. Cohen's Kappa also has some key limitations; most relevant here is the fact that it does not differentiate between types of errors, so false negatives and false positives are treated identically (even though a false negative is worse in the context of QuickCheck).

2.5 Legal Requirements for Vision Screening

This section describes the legal requirements at both a national and state level placed on vision screening in general and vision screening applications specifically. Section 2.5.1 covers the law for vision screenings in Washington state schools, and sections 2.5.2 and 2.5.3 cover federal statutes and their effect on vision screening regulations.

2.5.1 Vision Screening Under Washington State Law

Washington State law (Chapter 246-760) mandates that schools must test vision acuity in children in "kindergarten and grades one, two, three, five, and seven" and that those children

who do not meet the criteria set in 246-760-071 WAC (see Table 4) must be referred for a vision examination [59]. As of 2017, Washington state law requires all students to be screened for both near and distance vision problems. QuickCheck uses these criteria to establish thresholds for positive (vision problems) and negative (no vision problems) visual acuity tests. This means that using QuickCheck would allow a school nurse (or similar personnel) to determine if they needed to refer a student to a vision exam by law.

Purpose of Screening	**Grade**	**Screening Tools**	**Rescreening and Referral Criteria**
Distance Vision	Kindergarten	LEA vision test: Single LEA symbol (at 5 feet), or HOTV letter	Visual acuity worse than 20/40 in either eye
Distance Vision	Grade one	LEA vision test: Single LEA symbol (at 5 feet), or HOTV letter	Visual acuity worse than 20/32 in either eye
Distance Vision	Grades two and above	LEA vision tests: LEA symbols or numbers, or HOTV letters, or Sloan letters	Visual acuity worse than 20/32 in either eye
Near Vision Acuity	Kindergarten	LEA vision tests: LEA symbols near vision, HOTV, or Sloan letters	Visual acuity worse than 20/40 in either eye
Near Vision Acuity	Grade one and above	LEA vision tests: LEA symbols near vision, HOTV, or Sloan letters	Visual acuity worse than 20/32 in either eye

Table 4: Vision Acuity Referral Criteria for Washington State Schoolchildren *[59]*

2.5.2 Vision Screening Under HIPAA

HIPAA, or the Healthcare Insurance Portability and Accountability Act, is the core federal law which regulates the use, access, and disclosure of personal health information (PHI). Under HIPAA, all covered entities (including healthcare organizations as well as insurance providers and healthcare clearinghouses and business associates) must put in place PHI access and sharing policies as well as security controls to ensure that those policies are met. Violations of HIPAA result in a variety of penalties, including fines and loss of licensure. However, the core regulations of HIPAA only apply to PHI, not all possible medical information. Therefore, under the legality and minimal data gathering non-functional requirements of QuickCheck (see Section 1.4), the clinical testing version of QuickCheck does not make use of PHI. In order for healthcare information to not be considered PHI, it must not contain the following 14 data elements or qualities [60]:

- Names
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
    - The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
    - The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
- All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and

all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older

- Telephone numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Fax numbers
- Device identifiers and serial numbers
- Email addresses
- Web Universal Resource Locators (URLs)
- Social security numbers
- Internet Protocol (IP) addresses
- Medical record numbers
- Biometric identifiers, including finger and voice prints
- Health plan beneficiary numbers
- Full-face photographs and any comparable images
- Account numbers
- Certificate/license numbers
- The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

There is an exception to the above, where data may contain a unique identifying number, characteristic, or code, as permitted under the "re-identification" rule [60]. One reason for this

exception is to allow data to be reinforced by the addition of further data about its subjects gathered by future surveys, tests, or trials.

2.5.3 Vision Screening Under FERPA

FERPA, or the Family Educational Rights and Privacy Act, is a federal law which protects the privacy of students and the records of their education. FERPA gives both parents and their children rights regarding access to and control of their educational data. Once a student turns 18, all rights are transferred to the student only. These rights include the right to "inspect and review" their education records as maintained by their school(s), although schools are not usually required to provide free copies of records [61]. Additional rights include restrictions on how educational data can be shared with outside personnel. The term 'education records' is interpreted quite broadly, for example, FERPA covers information about:

- Whether or not a specific student attends a given school or class.
- A given student's presence at school on a given day.
- A student's grades or other performance records.
- A student's attendance records.
- A student's participation in any school-hosted or sponsored clubs, including school sports.
- Any other records generated by a school about a student.

Schools may present non-specific information about the general performance of their student populations, for example, they may discuss information about the average GPA of their students, or the number and size of school clubs, so long as that information does not personally identify students or could reasonably lead to such identification. In situations where both HIPAA and

FERPA can apply, such as a student's records of access to a school nurse, which rules school personnel must comply with become quite context dependent, but in general all applicable restrictions from both laws must be met in such cases [62].

In the context of QuickCheck's clinical testing, FERPA does not place many restrictions as the application is not being used in the context of education. However, the application should still not request any 'education records' as defined by FERPA in order to ensure compliance. Post-deployment, when QuickCheck itself creates 'education records' (such as when it is used to screen children by a school nurse), QuickCheck must comply with FERPA's rules of data access. This means that in the future, QuickCheck data will need to be stored securely, and it must be available on request by school personnel and should not allow data access to unauthorized personnel.

Chapter 3: Methods

This section covers the design of both the QuickCheck application as well as the design of the clinical test used to evaluate the application. Section 3.1 covers the base design of the application (before this thesis began) as well as changes made to the application, including both general changes and modifications for clinical testing. Section 3.2 covers study design, and covers both the initial study design, as well as secondary clinical test designs in other settings.

3.1 QuickCheck Application Design

Much of the design and development of the QuickCheck application was performed this thesis began. For the sake of clarity, notes are made where development or design work was performed as part of this thesis alongside the base design. Additionally, some of the work done before this thesis began had errors or deviations from QuickCheck's designs, so notes are made where those deviations were corrected.

As QuickCheck is designed to be easy for even inexperienced or young people to use, the original design includes a straightforward linear flow from one scene to another. In Unity, this progression is constructed by dividing the application into "scenes" which flow from one into another. In general, the application can be divided into 4 phases: the identification phase (containing the student information, initial settings, and calibration scenes), the testing phase (including both near and distance vision tests), the survey phase (including the CISS and ending survey), and the results phase (containing the results scene).

In a typical use case, the application begins with the identification phase, and moves through the testing and survey phase into the results phase. However, there are several alternate scenes that a

user can run as needed, such as the calibration and settings scenes. The calibration scene automatically runs the first time the application is used on a given device, after which calibration data is saved both locally and in the database. See Figure 13 for the general flow of the application. Note that each box within represents a distinct component of the application but does not necessarily correspond to an individual scene. See Section 3.1.2 for a breakdown of the scene-by-scene flow.

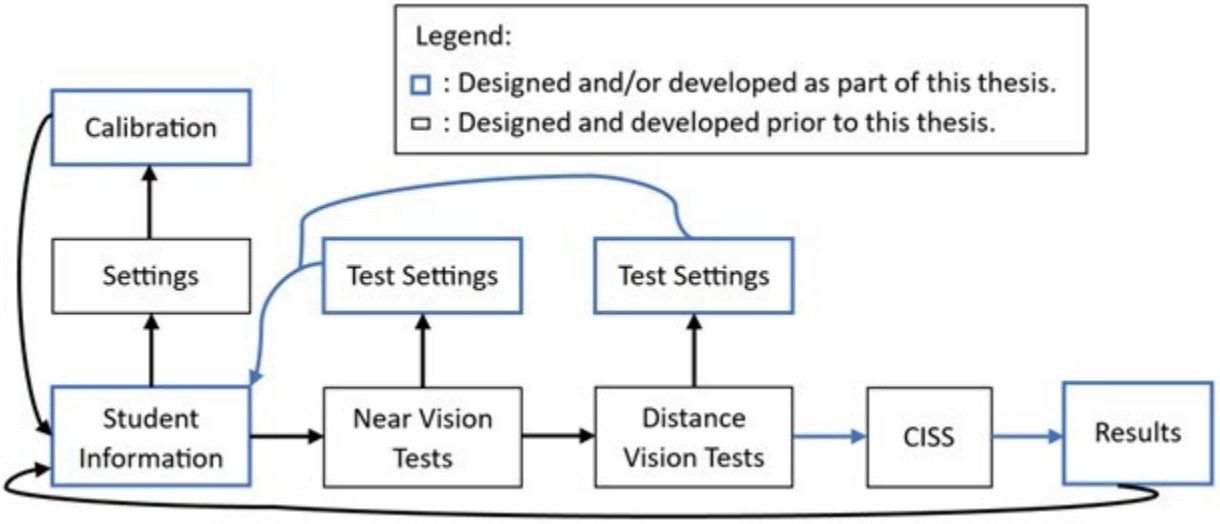

Figure 13: Simple Flow of the QuickCheck application

The application was built in the Unity engine, which has its benefits and drawbacks. Using Unity to build the application meant that application development experience was less necessary to contribute to the project, especially in the early phases of the project. Additionally, a wealth of documentation and online help is available for those developing new projects using Unity. On the other hand, it has meant that experience in producing applications using other tools has been

less transferable to the project. Additionally, Unity uses C# as a programming language, which is not as commonly used as other languages such as C/C++ or Java, which means that contributing to the project required learning a new application development tool and a new programming language.

3.1.1 QuickCheck Non-Functional Requirements

The QuickCheck application has a number of both functional and non-functional requirements, the vast majority of which were determined before this thesis began. See Appendix A for the list of functional requirements. Non-functional requirements are listed in the Introduction under thesis goals, but are repeated here for convenience:

NF2. Safety and Security – the application does not harm its users. There is no leaking of private data and little expected deviation in vision prediction from traditional methods of vision screening.

NF3. Testability – the application must produce falsifiable results associated with an identifiable individual who has also been (or will also be) tested with a traditional screening approach.

NF4. Minimal Data Gathering – the application gathers the minimum amount of information required to make a good estimate of vision quality.

NF5. Reliability – the application must be reliable enough for several hours of use on several devices.

NF6. Scalability – the application must be scalable (i.e., it must be usable on many devices simultaneously).

NF7. Ease of Use – the application is easy to use for the purpose of clinical verification by non-software experts.

NF8. Legality – all tests must follow legal rules around the use of medical information under HIPAA and FERPA. Further, the application must fulfill the legal criteria for what is required to be covered by a basic school vision screen.

3.1.2 Scene-by-Scene Overview

As much of the design work was performed before this thesis' work began on the QuickCheck application, many of the scenes were "inherited" from previous work. Where additions to the design were made or where development needed to be performed to correct deviations from the original design, those additions or changes are described alongside the base design. A key design decision made before this thesis began was how to split up design components into scenes; additionally, these scenes can vary quite a bit in amount of content. See Figure 14 for a scene-by-scene flow. In general, the scene flow matches fairly closely to what is expected from Figure 13; however, the settings components are a part of each individual scene and are not separated out. Additionally, the calibration and tutorial scenes link directly from the student information scene. Finally, the two testing scenes and the results scene all link back to the student information scene via their settings page.

Figure 14: Scene-by-Scene Flow of QuickCheck's Base Design

3.1.3 Student Information Scene

The student information screen (Figure 15) is responsible for gathering information regarding the subject being tested by the application. As per NF3 (minimal data gathering), only a limited number of features are collected. These are:

1. The grade of the student, ranging from Pre-School through to 12th grade, and including both “College” and “Adult”.
2. The age of the subject, as calculated by the subject entering their month and year of birth.
3. The Biological Sex of the participant.
4. The zip code of the subject. This is intended to keep track of where the application is being used when it is eventually released publicly.
5. Whether or not (and how often) the subject currently uses eyeglasses.

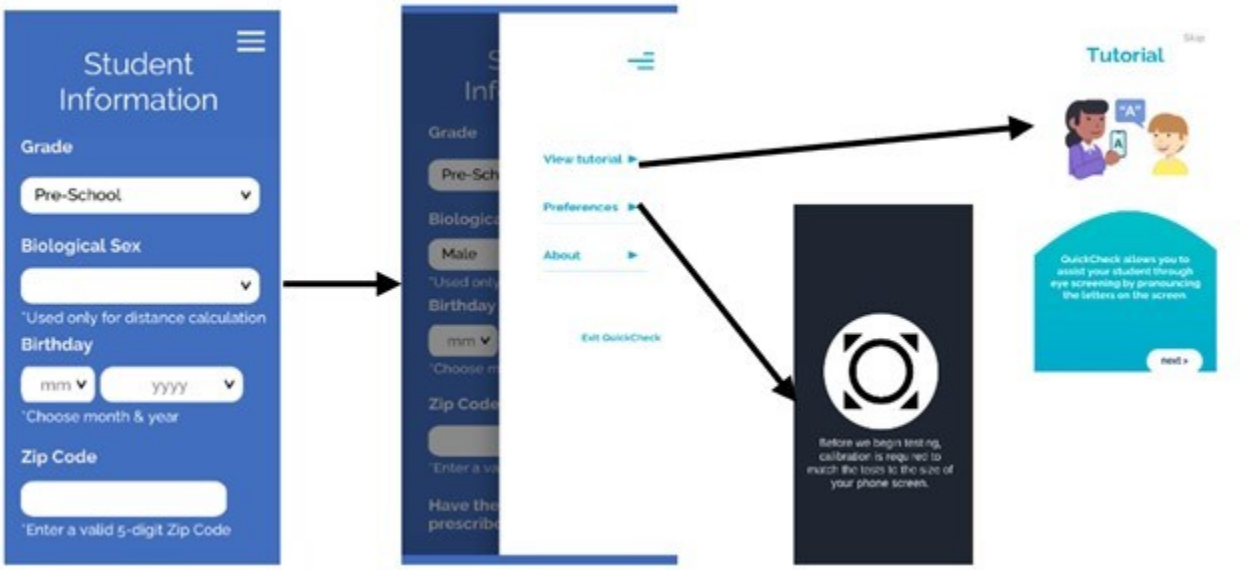

Figure 15: Student Information Scene

When the student information scene transitions to the Near Vision Testing scene, the application sends information about the subject and session to the backend. This data is sent in a custom data structure sent via HTTPS in the JSON format. Further communication from the current session maintains this session ID so that later scenes can send information to the backend that is stored alongside the information gathered in this scene.

While most of this work was done prior to the start of this thesis, some of the design choices were not fully implemented. Implemented designs include:

1. Implementing resizing to fit a wider range of device screens.
2. Enabling scrolling down through the questions so that the button linking to the next screen was always visible when it should be.
3. Fully implementing the settings page, including connecting the settings page to the calibration scene and tutorial scene.

4. Connecting the student information scene to the backend correctly. There were several errors that needed to be fixed for this to work, including correcting the data format used by this scene to communicate with the backend. Additionally, the encryption scheme used to send information securely needed to be adjusted slightly to a more secure version.

The design of the Settings scene is intended to allow a school nurse, doctor, or other healthcare provider to quickly obtain and enter the information needed to complete a vision screening. Under the minimal data gathering non-functional requirement, the application should ask for no more information than is necessary to complete a vision screening. Past UI and useability work showed that auto-populating the information from a student information database improved speed and usability, but this was not possible during clinical testing because no such database existed for all potential subjects during clinical testing [63].

3.1.4 Student Information Changes Made for Testing

During the clinical testing phase, changes needed to be made in the student information scene to better fit in a testing environment. These changes were:

1. Producing a new Student Information scene specific to clinical testing, so that both versions could be accessed through the same application.
2. Implementing a new field to capture an anonymous ID tag used to link QuickCheck results to results generated by more traditional eye screening tools and techniques.
3. Removing fields that gathered personally identifying information as defined by HIPAA's safe-harbor de-identification rule, including biological sex and month of birth.
4. Increasing spacing between elements to prevent the "next scene" button from being pressed by accident during the initial interview with subjects.
5. Correcting word choice where appropriate to ensure readability.

These changes were not without their drawbacks, however. Altering the student information scene for clinical testing required producing a second copy of the scene which was independent of the initial version. This meant that any changes that needed to be made to the student information scene, such as the changes made to properly connect to the backend, needed to be copied over to both scenes so that both versions would work properly. This style of change increases development time in the long term and degraded the program's modifiability. However, as the new scene was only necessary for the testing process, it was judged that the benefits in terms of initial reduction in development time were worth the longer-term cost. Figure 16 shows the clinical testing version of the student information scene.

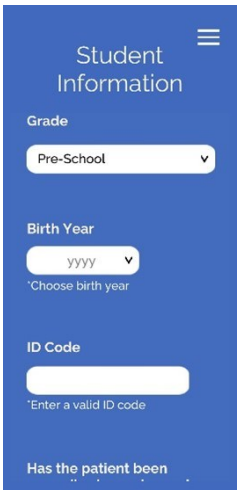

Figure 16: Clinical Version of the Student Information Scene

3.1.5 Tutorial

As shown in Figure 17, the tutorial scene gives a brief overview of the tests that are performed in the application. Data generated by the tutorial is not sent to the backend. Once the tutorial has been completed, the user is returned to the student information scene. Previous usability testing uncovered concerns that subjects (particularly younger children) would require time and practice

to understand how to use the application [63]. Additionally, widespread implementation could mean that there will be too many users to provide individualized guidance to each of them. The tutorial attempts to solve both of these problems by teaching subjects how to be tested (and users how to perform the tests) without storing the information for later use. That way, confused users can experiment with the application without needing to worry about their results. Also, very young study subjects who have difficulty understanding a verbal explanation of the application can be quickly shown how it works.

While the design of the tutorial was finalized before this thesis began, development of this scene had not been completed. The tutorial did not correctly return users to the student information screen and could not always be accessed when required. Furthermore, the scene did not scale well to different screen sizes, which also needed to be corrected. Finally, the optotype-phase did not always end at the right time. Each of these issues needed to be correctly fixed before clinical testing could begin.

Figure 17: Tutorial Scene and Transition to Student Information Scene

3.1.6 Calibration Scene

The calibration scene (Figure 18) is used to scale the size of the vision tests so that the correct sizes are displayed by the near and distance vision tests. Initial work determined that the size of pixels on a phone varied too much for device-reported size to be useful across a wide range of mobile phones, so instead users are asked to scale the size of the image of a quarter so that it aligns with an actual quarter when pressed against the screen. After calibration, the near and distance vision tests have their sizes adjusted according to the size of the phone being used. This information is saved remotely in the QuickCheck database (see Section 3.1.13) so that the device does not need to be recalibrated each time it is used.

Figure 18: Calibration Scene with Transition to Student Information Scene

However, at the time this thesis started, the calibration scene did not properly calculate the size of the letters to display for near and distance vision tests. The goal of the tests is to present each subject with an image of a letter which is 5 arcminutes tall and wide. Due to changes in height and visual field as children age, the actual size of letters needs to vary as the age of a subject increases for the letter to take up the same proportion of the visual field. Changing the calibration scene involved recalculating the scaling factors for the near and distance vision tests so that each

test correctly displayed the correct size of letter. Table 5 shows the desired letter size for each age group and test type. The size of letter stabilizes beyond second grade, because at that point QuickCheck begins testing all subjects at the same distances and for the same acuity thresholds, per Washington state requirements for school vision screening (see Table 4) [59].

Grade	Desired Letter Size at 16in (mm)	Desired Letter Size (distant, mm)
Pre-School	2.13	8.00 (at 5 feet)
Kindergarten	2.13	8.00 (at 5 feet)
1st	1.70	6.38 (at 5 feet)
2nd	1.70	12.77 (at 10 feet)
Adult	1.70	12.77 (at 10 feet)

Table 5: Optotype Sizes for Vision Testing for Various Grade Levels

In order to ensure that the sizes demonstrated in Table 5 were met, the following procedure was used:

1. Recalculate the size of optotypes on the screen (using the Snellen criteria from Table 4 and the size calculation in Equation 1) to ensure that the sizes in mm supplied in Table 5 were correct.
2. Re-scale the optotype sizes to be 10 times larger than their current sizes so that the differences between near vision optotype sizes could be easily measured.
3. For an Android phone and 3 different android simulators:
    a. Perform calibration using a United States' quarter.

    b. Measure the size of each age and distance category.
    c. Repeat after re-calibrating with 2 more quarters.
4. Calculate the average size of each optotype at each distance using the data gathered in step .
5. Calculate the conversion factor needed to take the current scaled optotype size to the correct scaled optotype size.
6. Implement the conversion factors calculated in step 4 by multiplying the current optotype sizes by those factors.
7. Undo the scaling done in step 2.
8. Re-measure the optotypes to ensure the correct size was used.

Table 6 shows the data gathered in steps 3, 4, and 5. The initial optotype sizes were relatively close to correct for near vision tests but were off by more than a factor of three for all distance tests. However, there were significant differences in optotype size, particularly for distance vision. This is for two main reasons: first, not all U.S. quarters are of the same size (older quarters tend to be worn down around the edges or bent slightly), and secondly, the process of matching a quarter to an image on the screen is not particularly precise. Part of the issue is that Quarters are ridged around the edge; not only do these ridges wear down over time, but they also allow some of the image to be seen from behind the quarter. This makes precisely calibrating optotype sizes a challenge.

In practice, this imprecision in calibration was worse for distance optotypes than for near vision optotypes, because a small change in calibration value resulted in a larger absolute change for distance optotypes than near optotypes. During the phase of re-calculating calibration factors, the

distance optotypes varied by as much as 3 or 4mm. This means that it is likely that, in the real world with users being perhaps less careful in performing the calibration size, and with a wider range of Quarter sizes used for calibration, it is very possible that the distance vision optotypes could vary in size by 1 to 2 mm, and the near optotypes by 0.1 to 0.2 mm. With particularly imprecise calibration, those numbers could increase further.

School Year	Target Near (mm)	Actual Near Average (mm)	Conversion Factor (Near)	Target Distance (mm)	Actual Average Distant (mm)	Conversion Factor (Distant)
Preschool	21.279	23	0.925	79.796	22	3.627
Kindergarten	21.279	23	0.925	79.796	22	3.627
1st	17.023	19	0.896	63.837	17	3.755
2nd	17.023	18	0.946	127.674	35	3.648

Table 6: Average Actual and Target Optotype Sizes in mm, Scaled by a factor of 10.

3.1.7 Near and Distance Vision Test Scenes

In the base design of QuickCheck, each vision test is performed by displaying a series of optotypes to the subject until 4 optotypes are read correctly in a row or ten optotypes are attempted by the subject. In other words, a subject who had read 3 optotypes correctly in a row with 7 optotypes attempted would still need to read one more optotype correctly or three incorrectly before that test would be complete. The subject's distance from the phone screen varies by test; in distance vision tests, the tester should stand 10 feet away from the subject, while in near vision tests, the phone must be 16 inches (40cm) away from the subject's face.

Additionally, in order to maintain the size of each letter in the subject's visual field across different ages and distances, the size of each letter by test type and subject age must vary (Table 5). As the application limits the number of unique optotypes presented to four to improve

readability for young children, the only possible letters that can be supplied by the application are T, O, H, and V, though subjects were not informed of this limitation. Additionally, each letter the subject is presented with is selected completely at random at test time by the application; this means that some subjects may see fewer than four letter types in their tests, though this is unlikely.

When using the application, the tester presents the subject with the letter displayed on the screen and asks the subject to read the letter aloud. If the subject reads the letter correctly, the tester swipes the screen to the right, prompting the application to present a new letter. If the subject reads the letter incorrectly, the tester swipes to the left, again prompting the application to display a new letter. Whenever an optotype is 'submitted' through swiping, the application provides haptic feedback through vibrating. This style of feedback was implemented to meet three main goals:

1. Speed – Simply swiping left or right allows the subject's guess to be entered into the system.
2. Limiting feedback – By using tactile inputs like swiping, it is harder for a subject to guess (especially for younger subjects) whether they guessed correctly or incorrectly.
3. Limiting information leaks – Unlike initial designs for QuickCheck, the swiping approach does not give away what letters can be shown to the subject [63].

In addition to these goals, the haptic feedback ensures that the tester is aware of whether or not they have swiped far enough to submit a given optotype. See Figure 19 for a comparison of a 2019 design of a QuickCheck testing scene and the current base testing scene. As Figure 19 shows, a key benefit of moving to the newer UI design was a lack of information leak, as the

prior design clearly displayed the possible options for what the displayed optotype could be. However, swipe direction being the only way to ‘submit’ an optotype means that it is easier for user error to result in an incorrectly submitted optotype. Furthermore, the new design does not allow users to record what letter was guessed, which means that systematic errors (e.g., always mistaking a T for a V) may not be noticed.

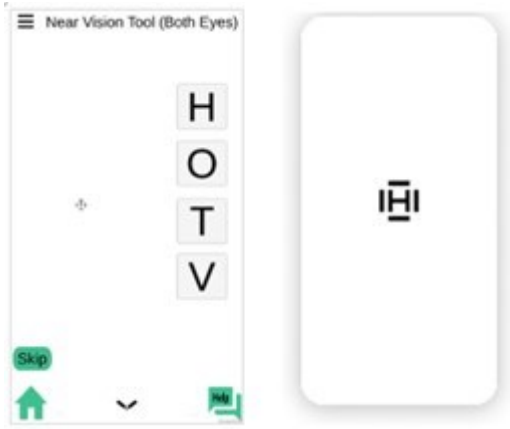

Figure 19: Prior Acuity Test Design (Left, *[63]*) and Current Base Acuity Test Design (Right)

In the base design, the trial ends when the subject reads 4 optotypes correctly in a row (‘passing’ the test, no vision problems) or attempting 10 optotypes without reading 4 correctly in a row (‘failing’ the test), so the next round of testing is presented. For each distance category, three testing rounds are completed; one where the subject uses both eyes, and two where the subject uses only one eye each. See Figure 6 for a visualization of the near test scene.

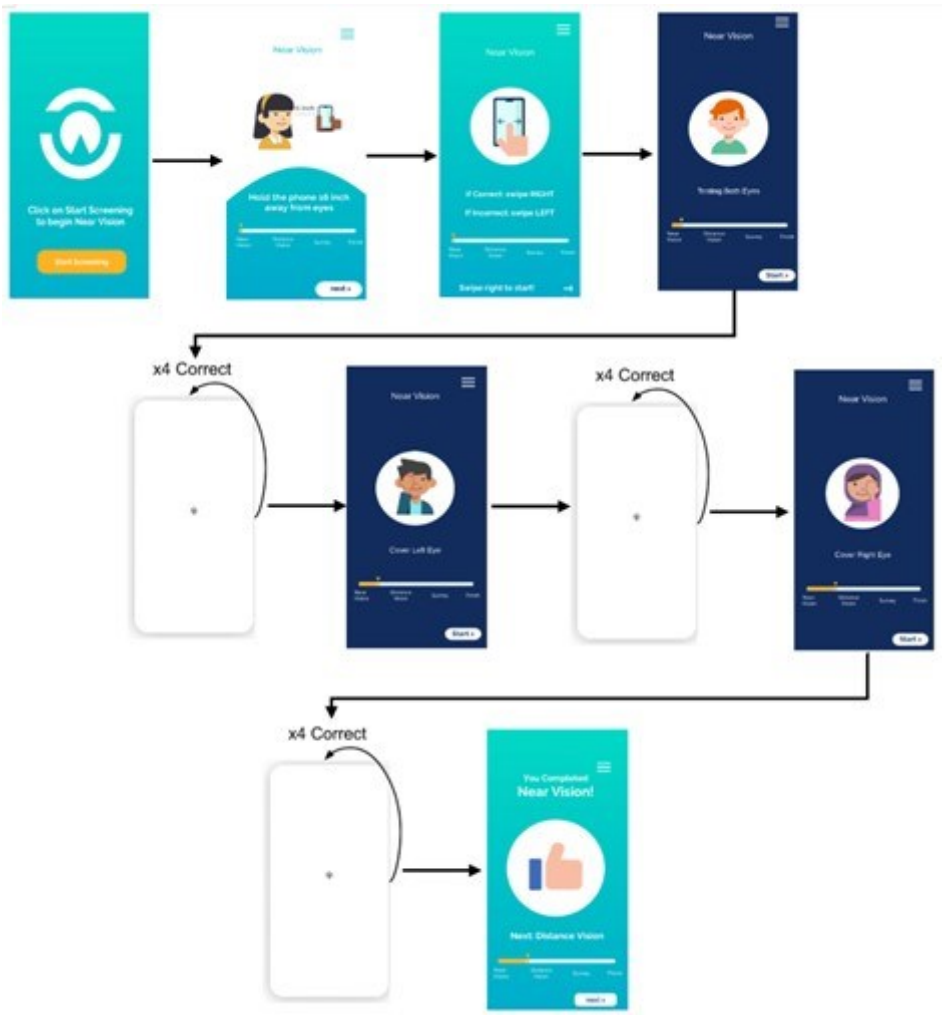

Figure 20: Base Near Vision Test Scene

Because each letter shown by QuickCheck is of the same size (for a given distance and grade level of subject), QuickCheck uses criteria-based rather than threshold-based acuity tests. While this means that the application is less precise in its determination of visual acuity, QuickCheck is intended to determine if a subject must be referred to a vision exam under Washington State law. Therefore, in order to decrease testing time, QuickCheck can be restricted to only testing for the required vision acuity criteria without violating any requirements.

The distance vision test scene is the same as the near vision scene except in the following two ways:

1. The scale of the letters to read is different. See Table 5 for the size of the letters for different tests and subject ages.
2. The near vision test moves into the distance vision test, but the distance vision tests moves into the CISS scene once complete.

3.1.8 Changes to the Near and Distance Vision Test Scenes for Clinical Tests

Two key changes needed to be made to the vision testing scenes in order to prove their efficacy. As with the student information scene, copies of the two testing scenes were made and then modified. The modifications were:

1. The type and number of tests taken was changed. Instead of requiring 4 correct optotypes to move to the next test, each test for each eye at each distance displayed a sequence of 10 optotypes every time. This means that during clinical testing, ERG volunteers presented 30 optotypes each for the near and distance vision testing scenes subjects for a total of 60 optotypes shown to each subject.
2. The ability to reset the test was removed. As part of the goal of the testing process is to gather information about application failures; keeping data about when a test failed, rather than overwriting it, was important. Instead, attempting to reset the test brings the user back to the home scene.

Additionally, there was a small error which could result in the optotypes not being displayed in a random order, which was corrected.

3.1.9 The CISS Scene

The CISS (Convergence Insufficiency Symptom Survey) is designed to test for convergence insufficiency, a condition where a subject's eyes cannot simultaneously focus on the same object, usually when said object is close to the face. The CISS used in this application consists of 15 questions (see Appendix B for a list of CISS questions), and the QuickCheck application adds an additional question ("When was the last time you had an eye exam with an eye doctor?", with possible answers "Less than one year ago", "More than one year ago", "More than two years ago", and "Never"). See Figure 21 for an example of a CISS question and the non-CISS question.

Figure 21: CISS Sample Question (left) and non-CISS Question (right)

The CISS survey results (as well as the results of the after-survey question) are sent to the backend via HTTPS as JSON after the survey is complete. Both contribute to the results displayed in the results scene displayed directly following the after-survey question. For example, any student who has had an eye appointment more than two years ago receives a recommendation to visit an optometrist for an eye examination.

Most of the work on the CISS scene had been completed prior to this thesis' start. New additions included:

- Correctly implementing the scene transitions to and from the CISS scene.
- Recording CISS results so they could be displayed to the screen to ensure data correctness.
- Correcting several communication and display errors, which together resulted in the CISS data not being correctly sent, received, or stored correctly in the QuickCheck server or backend. Several of these bugs were on the front end, but some were in the server and database.

3.1.10 The Results Scene

The results scene, as the name implies, summarizes the results of the prior tests performed by the QuickCheck application. In general, there are 5 result types that can have an auto generated message. These are:

1. No indications of vision problems were found, and the survey questions did not reveal anything worrying about convergence insufficiency of lack of eye exams in the recent past.
2. The near vision tests indicate a vision problem.
3. The distance vision tests indicate a vision problem.
4. The CISS survey indicates a problem with convergence insufficiency.
5. The subject reports that they have "Never" seen an eye doctor, or that their last visit was "more than two years ago."

Case 1 cannot co-occur with any of the other cases, but cases 2 through 5 can be displayed together, as shown in Figure 22, where both case 2 and 3 occur.

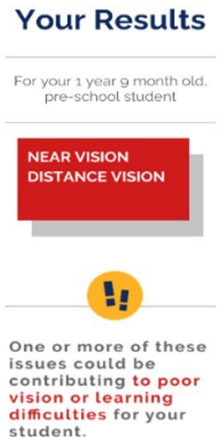

Figure 22: The Results Scene Displaying Concerns about a Subject's Near and Distance Vision

3.1.11 Changes to the Results Page for Testing

Although the results page is useful for non-testing use cases, in clinical testing, the base design of the results page is not useable. This is because presenting results that have not been clinically validated is, at best, ethically dubious. Additionally, there is little utility in displaying said results, as test subjects need to have their vision screened with more traditional methods so that those results can be compared with the results found by QuickCheck. Therefore, using the initial design of the results page provides at best no new information for study subjects, and at worst provides inaccurate results and recommendations. Therefore, a new results screen was needed.

The clinical testing results scene (shown in Figure 23) borrows from the design of the first sub-scene of the base results scene. However, instead of displaying a message thanking a subject for completing the tests, the clinical results scene displays an encoded version of the results of their tests.

Figure 23: Clinical Testing Results Scene

The encodings are created as shown in Table 7. Once the encodings have been created, they are concatenated together with the result category abbreviation (see Table 7), and then all encodings are concatenated together. This scheme was not meant to be an encryption of the results; rather, it was intended to provide a short summary of the test results. Not only did this mean that subjects would be unable to understand what QuickCheck had determined about their results, but it also meant that the results could be written down as a backup in case the internet service at a testing location was unexpectedly unavailable.

Test Category	**Abbreviation**	**Encoding Scheme**
Near Vision, Both Eyes	NB	Correctly read optotype = 1 Incorrectly read = 0 Concatenate readings together into a binary number, then convert that
Near Vision, Right Only	NR	
Near Vision, Left Only	NL	
Distant Vision, Both Eyes	DB	
Distant Vision, Right Only	DR	

Test Category	Abbreviation	Encoding Scheme
Distant Vision, Left Only	DL	number into a decimal number*
CISS Answers	CISS	Encode answers as digits 0 (never) to 4 (always), concatenate digits together in order.

Table 7: Results Encoding Schemes by Test Category

*e.g., reading all 10 optotypes in a test category correctly becomes 1111111111, which becomes 1023

3.1.12 Entry Scene

In order to prevent unauthorized access to the QuickCheck application, as required by the IRB, the QuickCheck application's clinical testing version required a basic access control system for the front end. In order to comply with this requirement, an "entry scene" with a basic password system was implemented. To begin screening in the clinical testing version of the application, the user is required to enter a password. Since there is only one tester of QuickCheck for this thesis, the inclusion of a username-password pair was deemed unnecessary, however, both versions were developed.

The new scene, shown in Figure 24, is now the first scene a user sees when starting the application. Once the user has successfully entered their password, they can press "next" to move to the "Home scene" (see Figure 14). If the user has entered the wrong password, they are prompted to re-enter their password and are not taken to the Home Page scene.

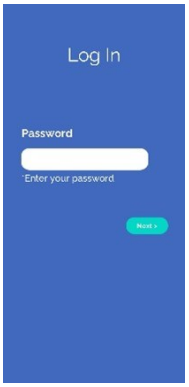

Figure 24: The Clinical Testing Entry Scene

3.1.13 QuickCheck Stack and Cognito Authentication

QuickCheck uses a typical three-layer system, with a frontend built in Unity, an Apache server built in Node.js, and a MySQL database hosted on Amazon Web Services using the Relational Database Service. Prior to this thesis, this three-tiered client-server design pattern had only these three components (see Figure 25). To ensure that the application can still function even in rural areas with little internet connection, the QuickCheck application stores data locally when it cannot be sent to the backend. Then, once the application is started with an internet connection, the unsent data is resent to the backend, and removed from local storage. An extremely useful feature that was implemented prior to the start of this thesis is that, in the event that the application cannot connect to the server, data will be saved locally. Then, whenever the application is next opened with a successful connection to the server, the saved data will be sent to the server for remote storage.

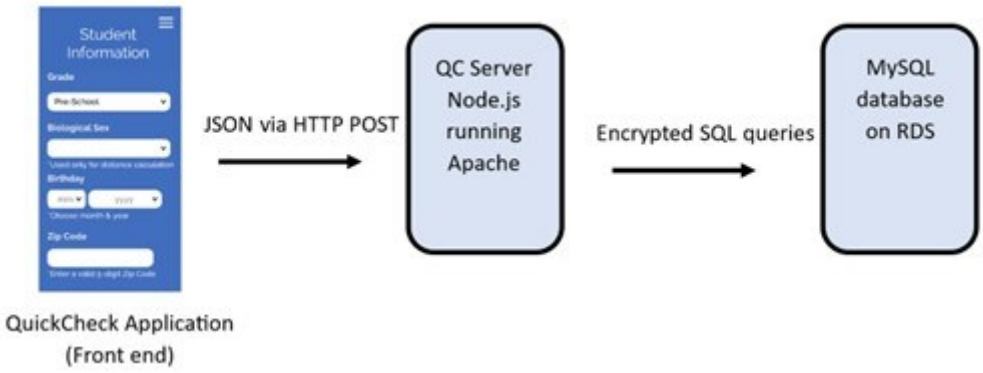

Figure 25: Initial QuickCheck 3-Tiered Client-Server Design

However, as part of the requirements set out in the IRB request (see section 3.2.1), additional user authentication was required. Therefore, with the assistance of Wooyoung Son, the front end and server were configured to use AWS Cognito authentication service. Wooyoung Son was primarily responsible for configuring the server to use Cognito, and I was primarily responsible for configuring the front-end application to use Cognito. The new QuickCheck architecture is as shown in Figure 26, where both the QuickCheck front end and server connect to the Cognito service. The Cognito authentication service works by requiring the QuickCheck frontend to present identity verification. Upon receiving this information, it grants a temporary access token. QuickCheck then adds this token to its communication with the server, and the server can then query the Cognito system to ensure the token is valid at the time of communication.

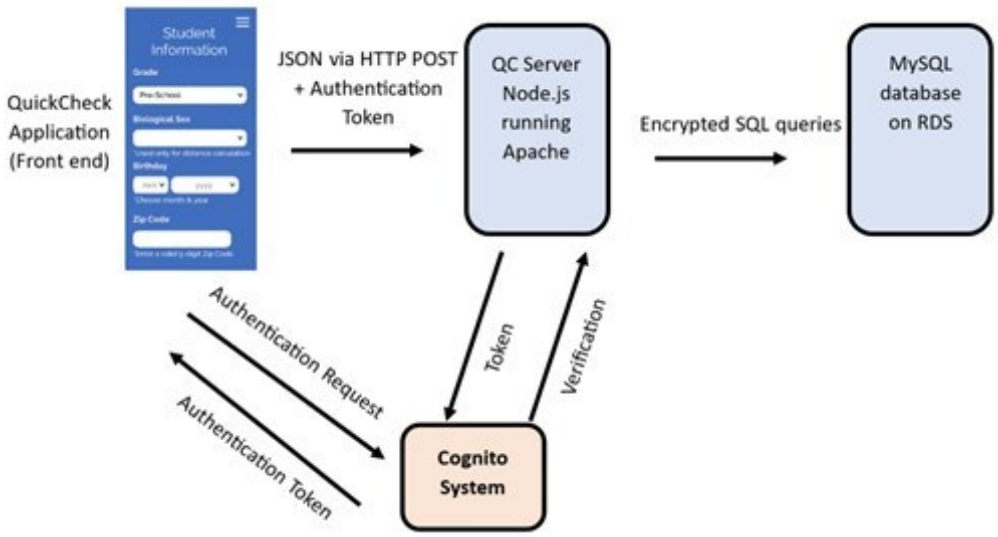

Figure 26: Current QuickCheck Design

3.2 Clinical Study Design

Section 3.2 covers the design of the clinical test performed to determine QuickCheck's effectiveness. Section 3.2.1 discusses obtaining ethics board approval for clinical research, section 3.2.2 covers clinical study design, and 3.2.3 briefly discusses plans for statistical analysis made prior to the collection of data.

The goal of clinical testing was to examine the correctness of the vision screening results as produced by the QuickCheck application. Secondary goals included streamlining the application's tests, examining ease of use (NF6), and providing a stress test for the application's backend (NF4). Therefore, the QuickCheck application must be tested against traditional vision screening tools. For a description of the tools compared against QuickCheck, see sections 2.2.3 (HOVT cards) and 2.3.2.1 (FrACT).

3.2.1 IRB and YPS Approval Process

Because testing QuickCheck required working with human subjects, the approval of an ethics board was required before testing could begin. For the University of Washington, the relevant ethics board was the Institutional Review Board (IRB) run by the Human Subjects Division (HSD). In order for the board to approve this thesis, they required:

1. An application form detailing the study design, data collection and storage, safety controls and risk management strategies, and more.
2. A Data and Safety Monitoring Plan.
3. The consent form to be signed by prospective subjects or their parents.

Additionally, once IRB approval was granted, each researcher who would interact with minors in the course of this study was required to complete the following by the University of Washington's Youth Protection Services:

1. Undergo (and pass) a criminal background check.
2. Sign a contract detailing appropriate and inappropriate behavior when interacting with minors.
3. Complete training for how to interact with minors.
4. Complete training for detecting child abuse in the course of research.

Each of these steps was completed, however, there was some delay in completing the criminal background check that reduced the amount of time available to gather data. Additionally, due to the requirement of getting consent from parents for their children to participate in the study, many of the Near Vision Institute's mobile vision clinic events that could have been used to gather data could not be, as no parents would be present.

3.2.2 Initial Clinical Test Design

In order to properly test the QuickCheck application, the EYE Research Group (ERG) collaborated with the Near Vision Institute (NVI), led by Dr. Alan Pearson. NVI provides vision screening services to schools and community centers. To determine if QuickCheck can successfully detect vision problems, NVI invited volunteers from the EYE research group to test QuickCheck on subjects that had already been screened or would soon be screened by traditional methods. The first clinical test and later clinical tests took place under different circumstances; the first clinical test occurred at an NVI vision screening event in Darrington, WA. The general strategy of the first clinical test was as follows:

1. NVI and the ERG volunteer set up a portable vision screening system at the screening event.
2. NVI set up 4 photoscreening stations (this is the default screening procedure for NVI, everyone at such events gets the photoscreening service). NVI additionally had a more traditional distance acuity (using FrACT) and near acuity optotype screening (using a near vision card). These additional screening services provided a point of comparison against QuickCheck.
3. A subject would approach the onboarding station, and receive photoscreening and other onboarding tests, such as a color vision test. The subject would then be directed to the QuickCheck substation.
4. The ERG volunteer at the substation asked for consent to screen the subject, and upon receiving consent from the subject (and their parent, if the subject was a minor), would test QuickCheck as well as FrACT (at 10 feet) and a near vision card (at 16in/40cm) on

the subject. If the subject wore glasses, they would be asked to remove their glasses prior to testing QuickCheck and the other tests.

5. To anonymously connect the QuickCheck results to the photoscreening and traditional optotyping results, the tester entered an anonymous ID number into the application in place of a zip code.
6. After the event, a table linking the ID numbers and the subject's names was given to NVI so that NVI could share de-identified data regarding the results of the screening with photoscreening and traditional optotypes.
7. Each individual subject had their QuickCheck results compared with their results from the other screening methods using the anonymous IDs to link across data sets. To avoid breaching HIPAA data security requirements, only the de-identified data was provided to ERG. This data consisted only of the anonymous ID and the testing results for QuickCheck, FrACT, and the near vision card.

Note that all subjects were required to meet the inclusion criteria (and not meet the exclusion criteria). The inclusion criteria were:

- Subjects must be older than 18 or be under the age of 18 with a parent/guardian present to provide informed consent.
- Subjects (and their parent/guardian(s) if the subject is under 18) must speak English and must be able to read letters on a screen.
- Subjects must be present at a Near Vision Institute (NVI)-run vision screening event.

The formal Exclusion Criteria were:

- Subjects younger than 18 who do not have a parent/guardian present.
- Subjects (or parents/guardians) who do not speak English at a level needed to understand the instructions and consent information that are provided in English only.

3.2.3 Secondary Clinical Test Designs

In order to gather additional data, further clinical tests were undertaken. These clinical tests used a similar design to the first test but were undertaken in a different setting. There were two types of later tests: those undertaken at a vision clinic, and (later) those undertaken with adult volunteers. Section 3.2.3.1 covers the former, and section 3.2.3.2 covers the latter.

3.2.3.1 Clinical Tests at EYE SEE Clinic

Unlike the initial trial, which was at a mobile clinic event, later tests were at the EYE SEE clinic in Bellevue, WA (NVI's pediatric vision clinic). The general strategy of these clinical tests was as follows:

1. The ERG volunteer set up QuickCheck, FrACT, and near vision card tests in the EYE SEE clinic.
2. A clinic patient would approach the EYE SEE clinic's front desk to check in for an appointment. At this time, the front desk would briefly describe the presence and purpose of the QuickCheck clinical test.
3. After the subject's eye exam with an EYE SEE optometrist or vision therapist, the ERG volunteer would approach the subject and their parent(s) and ask for permission to test QuickCheck. If the subject and their parent(s) consented, and met the inclusion criteria, the subjects would be tested.

4. To anonymously connect the QuickCheck results to traditional results, the tester entered an anonymous ID number into the application in place of a zip code.
5. After testing QuickCheck (or if the subject and/or their parent(s) did not consent to test QuickCheck), measurement of visual acuity using FrACT for distance vision (10 feet) and a near vision card for near vision (at 16 inches) was performed. These served as the reference measurements to compare QuickCheck against.
6. After the event, a table linking the ID numbers and the subject's names was given to NVI so that NVI could share de-identified data regarding the results of the screening with and traditional techniques.
7. Each individual subject had their QuickCheck results compared with their results from the other screening methods using the anonymous IDs to link across data sets. To avoid breaching HIPAA data security requirements, only the de-identified data was provided to ERG. This data consisted only of the anonymous ID and the testing results for QuickCheck, FrACT, and the near vision card.

The study design of the initial trial and later trials was intentionally similar in order to prevent systematic differences in trial design from producing bias in the data. The key differences in trial design from the initial to later tests were:

1. Setting: The initial test was at a mobile clinic event, but later tests were in a clinic.
2. Test order: In the initial test, the QuickCheck test came before seeing an optometrist. In later tests, the order was reversed.

3. Glasses: some subjects did not wish to test QuickCheck with their glasses off, and so used their glasses while testing QuickCheck. However, no subjects were tested with and without their glasses.

3.2.3.2 Tests with Adult Volunteers

After tests at the EYE SEE clinic were completed, there were a few adults who volunteered to participate in this thesis outside the context of a vision exam. The same study design was used for these adults as for adults during a visit to the EYE SEE clinic, except that no testing was done by NVI, and no ID-name table was given to NVI about these adults. The main differences between these tests and others were:

1. Lack of previous tests: as mentioned, no tests were performed by NVI, so the subjects were "fresh" at vision tests (at least with regards to very recent history).
2. Lighting: the lighting conditions were very different for some of the adult volunteers, with a few occurring during mostly natural light with less artificial illumination than in other trials.

3.2.4 Study Restrictions and Considerations for QuickCheck

Despite the relatively simple study design in use to test QuickCheck, there were a few additional considerations that needed to be taken into account for testing to succeed. First, as this study required working with children, an IRB approval process, including creating a Data Safety and Management plan, was required. Also, all EYE and NVI members participating at a screening event needed training to handle the screening devices. Additionally, EYE members needed training to work with children per the University of Washington's APS 10.13 policy. Additional training in detecting and proper reporting of child abuse was also required. All those working at

the screening event required two hours of training to ensure that the process would flow smoothly.

Furthermore, there were some changes that needed to be made to the design of the QuickCheck application to ensure that it could be used in clinical testing. Application-level changes were made primarily in the student information scene, the near and distance testing scenes, and the results scene. See Sections 3.1.3, 3.1.6, and 3.1.8 for more details on the scene-specific changes made for testing purposes.

As noted in Section 3.1.6, changes were made to the near and distance testing scenes in order to ensure that the vision tests are as efficient as possible. The base design of QuickCheck required that the vision tests continue until either 4 correct or 4 incorrect letter choices are made for each eye in each distance test. However, during clinical testing, the testing scenes were modified so that 10 letters were displayed for each test instead, which meant an increase in the number of tests performed on each eye. This change was made so that various different testing policies could be tested against the traditional screening techniques. For example, the base testing policy could be compared against a policy that required only 3 correct or 3 incorrect letters read before a test ended. In order to limit the possibility of unintentional bias though multiple hypothesis testing, all testing policies were determined before the final data was analyzed. See Table 8 for the list of all tested policies.

Policy Name	Normal Vision Requirements	Impaired Vision Requirements	Additional Notes
Four-Correct	4 correct letters	Reached 10 optotypes without achieving normal vision.	—
Two-Correct	2 correct letters	Reached 10 optotypes without achieving normal vision.	—
Five-Tests	3 out of 5 correct	3 out of 5 incorrect	5 tests performed per eye
Ten-Tests	6 out of 10 correct	6 out of 10 incorrect	10 tests performed per eye

Table 8: Tested Vision Policies

As none of the tested policies require more than 10 tests, all possible policies could be simulated using the test results generated by the vision screening event. Additionally, the first three policies could be further tested by randomly shuffling the order of letters read by the participant.

Finally, while the initial test design called for testing at a school event, this proved impossible due to ethical concerns. Primary among these was the lack of adults to ask for consent at an in-school screening. This meant that instead of an in-school event where only children attend, testing needed to be held at a community event where both children and adults would attend, visiting the EYE SEE clinic to gather data from children with their parents present, or gathering data from adult volunteers.

3.2.3 Statistical Analysis

Once the data from the QuickCheck application and the traditional screening techniques was gathered, statistical analysis was performed to determine if the QuickCheck application was capable of determining the vision status of its users. First, the performance of QuickCheck was measured using accuracy, specificity, and sensitivity of each QuickCheck policy, using both traditional optotyping results as the ground truth. This method is commonly used when comparing two test results, as seen in Dahlmann-Moor et. al.'s 2008 paper testing the Plusoptix vision screener [52].

Additionally, the three policies that did not use all available tests (that is the base policy, the fast policy, and the fixed 5 tests policy) were further examined via a shuffling method described in section 3.4.2. Each possible testing policy of QuickCheck was compared against the others. The goal of this test was to determine how closely matched the different methods were in general terms. Additional work to determine what caused different types of errors was then performed based on the initial performance metrics. Finally, an analysis of the two surveys contained within QuickCheck was performed, although the statistical analysis of the CISS was limited due to the lack of reference measurement about whether subjects actually had convergence insufficiency. Statistical analysis was performed in Python and recalculated using R to ensure correctness.

3.2.3.1 Example Study Result and Shuffling

Table 9 displays an example set of results of a test on a single eye at a single distance for a single subject. Note that these results were generated by hand and do not represent the actual results of any individual subject. For the sake of example, let us say that the following results are from a near vision test on the subject's left eye.

Letter Number	Letter Displayed	Correctness
1	T	Correct
2	H	Correct
3	O	Correct
4	T	Incorrect
5	H	Incorrect
6	V	Correct
7	T	Incorrect
8	O	Correct
9	H	Correct
10	T	Incorrect

Table 9: Example Vision Test Results for One Eye at One Distance

To determine whether this individual passes each test according to their policy, a given policy was applied to this individuals' results, ignoring any tests that occurred after the individual would have passed or failed a vision test using that particular policy. See Table 10 for details on when and why each policy was passed or failed by the example subject. These results would then be compared with the ground truth (either FrACT for distance vision tests or a near vision card for near vision tests). The threshold for passing a QuickCheck test is set at 20/32, so vision worse than that should not pass a vision acuity test (positive reference test result) and any vision better than that (negative reference test result) should pass a vision acuity test.

Policy Name	Normal Vision Requirements	Impaired Vision Requirements	Test Results
Four-Correct	4 correct optotypes	Reached 10 optotypes without achieving normal vision.	Passed on letter 6
Two-Correct	2 correct optotypes	Reached 10 optotypes without achieving normal vision.	Passed on letter 2
Five-Tests	3 out of 5 optotypes	3 out of 5 incorrect	Passed on letter 5
Ten-Tests	6 out of 10 optotypes	6 out of 10 incorrect	Passed on letter 10

Table 10: Example Test Results by Policy of Example Subject

From Table 10 it is evident that all test policies pass this example subject. However, for the first three tests, there were additional results that were ignored by the policy. For these policies, that ignored data can be considered by randomly shuffling the order of letters read by the subject and applying the policy to the new shuffled ordering. This additional step was performed for several reasons. First, it expanded the size of the data useable in examining QuickCheck's performance. Secondly, it eliminated some of the random chance that occurred when letters were being selected for an individual to read by the QuickCheck application. Additionally, this shuffling allowed examination of any bias that could result from ending tests early. For example, subjects who experience high eye strain from attempting to read the letters presented during a test may begin to read letters incorrectly more often as a test progresses. On the other hand, subjects may

artificially improve their reading ability as a test progresses by memorizing all possible letters even if they were initially difficult to differentiate. By comparing the shuffled results to the real results of many subjects, these sorts of biases can be examined, at least on the scale of a single test.

As Table 11 shows, all policies that can be shuffled demonstrate a pass ratio of above 0.5, indicating that the example test results were not likely to be a fluke caused by letter order. However, note that the fast test policy is closest to 0.5, which is an indication of a high variance policy. By performing such shuffling tests on all results generated by the three shuffle-able policies, the policies could be evaluated not just on the results that occurred, but also on many results that could have occurred if the application had randomly presented a different letter order.

Policy Name	**Proportion Passed**
Base Policy	0.83
Fast Tests	0.67
5 Tests	0.74

Table 11: Example Proportion of Vision Tests Passed With 10000 Random Shuffles

Chapter 4: Results

This chapter discusses the results of the clinical tests performed with QuickCheck. Section 4.1 is a discussion of the population sample of this thesis. Section 4.2 provides a quantitative comparison between the results of QuickCheck in recognizing refractive errors, Section 4.3 covers error analysis, and Section 4.4 covers the results of the CISS and doctor visit frequency question. Moving to qualitative results, Section 4.5 covers qualitative findings from clinical trials, including problems encountered and future research questions of interest, and finally section 4.6 discusses changes that could be made to improve the design of QuickCheck.

Throughout this section, visual acuity will be measured using the Snellen fraction denominator with a fixed numerator of 20. While using a different acuity metric, such as LogMAR, would have been more precise, the comparison tool used for near vision (near vision cards) only used Snellen fractions to measure acuity, so that is the measurement that must be used.

4.1 Population Statistics

Before analysis of QuickCheck can begin, the population of the study must first be understood. Given the relatively small population size of this study, there are limitations in how detailed population analysis can be, and subgroup-level analysis is even more limited. However, understanding the study population is still an important component of analyzing the results of QuickCheck. Section 4.1.1 covers population age, section 4.1.2 covers population vision and visual acuity, and section 4.1.3 covers population by testing modality.

4.1.1 Population Age

As QuickCheck is largely intended to be used on minors, the best possible study sample would have included only (or mostly) minors. However, given limitations on when and how data on minors could be gathered, a large portion of this study's population are 18 or older. Importantly, age within this study was measured as "time since Jan 1st on the year the subject was born", as month and day of birth were not collected. Therefore, the recorded age for any given subject may be 1 year older than that subject's actual age. There were 24 adults and 16 children in this study, although 3 of the 26 adults were exactly 18 years old. The mean age of the population is 26 years old (median 20), with a standard error of 19 years. Considering only children, the mean age is 10 years old, with a standard error of 1.75 years. The oldest subject was 71, the youngest was 7. See Figure 27 for a boxplot displaying the distribution of ages of children within the study, and Figure 28 for a boxplot displaying the distribution of adults within the study.

Notably, the ages of children tended to be between 9 and 12, with an inter-quartile range of 3.25 years. This indicates that most of the children in this sample are within the range of ages that Washington state law mandates receive vision screenings. For the adults, the ages are much more spread out, with an inter-quartile range of 23 years, although given that children can only be 17 years old, that is a largely mechanical difference. The oldest person in this study was 71 years old, and the youngest was 7, so overall a wide range of ages are represented. However, none of the subjects in this study were in kindergarten, so the 5 ft testing distance for children of that age could not be examined.

Grouped by eye, among eyes with good vision, the mean age is 25.6 years old, with 47% being eyes of adults, however, among patients with vision problems the mean age is 34.3 years old with 84% being eyes of adults. However, grouped by subject, the population is slightly different,

with a mean age of 19.2 years old for subjects with no vision problems, and 35.8 years old for subjects with any vision problem. In both cases, the distribution of ages amongst subjects (or eyes) with vision problems and those without are statistically significantly different ($p<0.05$ in both cases). This indicates (and is confirmed in Table 13) that in this study group, older subjects tended to have worse vision than younger subjects.

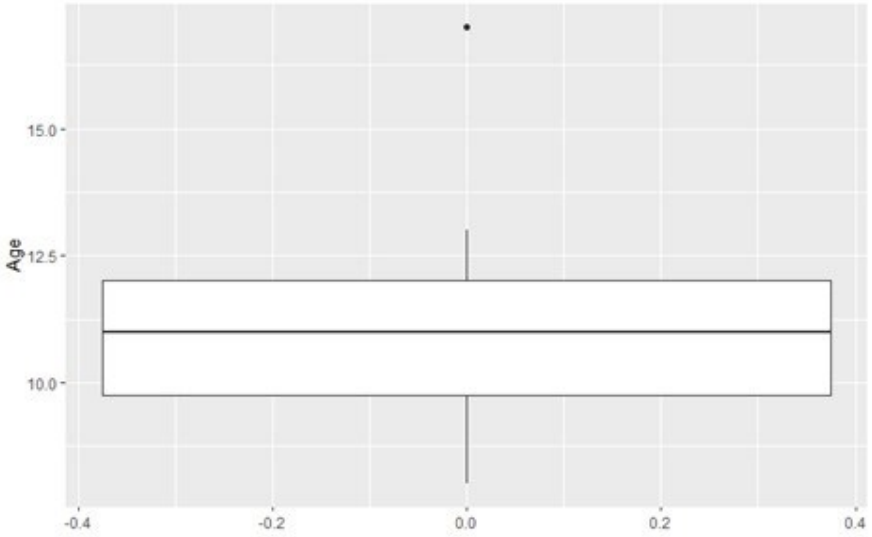

Figure 27: Boxplot Displaying the Ages of Children (less than 18y.o.)

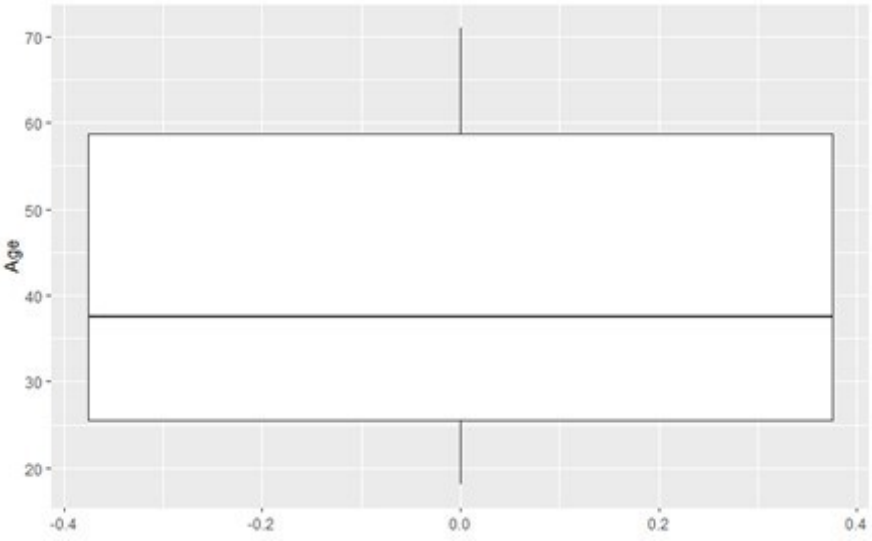

Figure 28: Boxplot of the Ages of Adults (greater than 18y.o.)

4.1.2 Population Visual Acuity

There were a wide range of vision acuities within this study. Table 13 shows the Snellen acuity scores for all study participants, except a single participant in the first trial who, due to time constraints, was unable to receive distance vision tests. The mean vision acuity is higher on average (representing worse vision) for adults than for children, which is expected. Additionally, visual acuity is lower on average (representing better vision) for both eyes than for either eye individually across all groups. Both of these are expected results, as vision tends to decline with age and people generally see better using both eyes than either eye individually.

Overall, there were 40 subjects in the sample (with usable data); of these, 23 had at least one vision problem, and 17 had no vision problems. Table 12 shows the mean and 95% confidence intervals (C.I.) for visual acuities of subjects by whether they had any vision problems. Acuity is the mean Snellen acuity denominator with a fixed numerator of 20. For example, the mean acuity of any subject with no vision problems was 20/19.7. However, the distributions for children without vision problems and adults without vision problems are extremely similar, which is the expected result, because healthy vision can only be healthy by having good acuity scores regardless of age. The differences between across age-group acuity average distributions for subjects with vision problems are also not statistically significant ($p > 0.05$), although the very different means for children and adults suggests that there are possibly differences between acuity for those groups.

Distance, Eye	**Mean Acuity ± SD**
Any Ages	
No Vision Problems	19.7 ± 2.0

Distance, Eye	Mean Acuity ± SD
Any Ages	
Any Vision Problem	81.6 ± 71.6
Adults Only	
No Vision Problems	19.5 ± 2.0
Any Vision Problem	87.8 ± 76.8
Children Only	
No Vision Problems	19.8 ± 2.2
Any Vision Problem	52.2 ± 29.2

Table 12: Vision Acuity Average Means and Standard Deviations by Subject Age Category

An unexpected result is that left eyes tend to have lower acuity than right eyes, as shown in Table 13. When this effect occurs in an individual, it is called anisometropia, and it is estimated to effect roughly 28% of U.S. adults. It is somewhat unexpected that the entire population has a bias in terms of acuity by eye, but given the very wide variation in acuity in this study's population, none of the across-distance, across-eye, nor across-age differences are statistically significant (p all >0.05 using a t-test), so it is difficult to draw statistically valid conclusions from these differences.

Distance, Eye	Snellen Mean (All Ages)	Snellen Mean (Adults)	Snellen Mean (Children)
Near, Both	20/34	20/40	20/25
Near, Right	20/45	20/56	20/28
Near, Left	20/50	20/65	20/25

Distance, Eye	Snellen Mean (All Ages)	Snellen Mean (Adults)	Snellen Mean (Children)
Distant, Both	20/52	20/72	20/19
Distant, Right	20/74	20/97	20/40
Distant, Left	20/77	20/109	20/29

Table 13: Snellen Visual Acuity Means by Age Category, Eye, and Distance

Table 14 shows the proportion of subjects with a vision problem in a certain eye at a certain distance by age category. The overall distribution of subjects with any vision problems is statistically significantly different between adults and children ($p < 0.001$). However, the across-eye proportions within age groups are not statistically significant, especially when taking into account multiple hypotheses testing ($p > 0.05$).

Distance, Eye	Vision Problem Proportion (All ages)	Vision Problem Proportion (Adults)	Vision Problem Proportion (Children)
Near, Both	0.32	0.46	0.12
Near, Right	0.42	0.58	0.19
Near, Left	0.4	0.54	0.19
Distant, Both	0.3	0.42	0.12
Distant, Right	0.35	0.46	0.19
Distant, Left	0.35	0.54	0.06

Table 14: Proportion of Vision Problems by Age and Test Category

4.1.3 Population by Test Modality

As discussed in Section 3.2, there were three different types of clinical tests performed on QuickCheck. The first type occurred at a mobile vision clinic event in Darrington, WA; the second was a series of days spent at the EYE SEE vision clinic in Bellevue, WA; the third was a series of tests performed on adult volunteers in Seattle and Bothell WA.

At the mobile clinic event, 19 subjects consented to testing; however, one of those subjects was not aiming to obtain correct vision test results (even on the comparison tests), so they were dropped from the data. This leaves a total of 18 subjects from the first clinical trial. Of these subjects 12 were adults (mean age 41.5 years old) and 6 were children (mean age 10.5 years old). For the second clinical trial (clinic visits) there were 5 adults (3 of which were exactly 18 years old) and 10 children, with a mean age of 16 years old. For the adult volunteers, the mean age was 49 years old. Table 15 shows the details of mean visual acuity by age group. Interestingly, the difference in acuity means among adults and children is statistically significantly different in the mobile clinic subjects but not clinic visit subjects ($p < 0.001$ vs $p > 0.05$ with t-tests). Additionally, the mean acuity of adult clinic visit subjects is statistically different than adult mobile clinic subjects ($p < 0.002$ with a t-test). This indicates that the adults that visited the mobile clinic are different than adult subjects from other trial types. Considering their much younger average age (26 vs 42 and 49 respectively) this is not unexpected.

Trial Type	Number of Subjects	Number of Adults	Mean Acuity* (Adults)	Number of Children	Mean Acuity* (Children)
Mobile Clinic	18	12	79	6	28
Clinic Visit	15	5	32	10	27
Volunteers	7	7	97	0	N/A

Table 15: Visual Acuity by Trial Type and Age Category

*Mean Snellen Denominator with Numerator of 20

An additional factor to consider is the purpose of clinic visitors as compared to mobile clinic subjects or volunteers. Clinic visitor subjects often had relatively good vision and were at a vision clinic not to get glasses, but to receive vision therapy for strabismus and/or similar conditions. Given the similarity between symptoms of strabismus and convergence insufficiency (at least, compared to the similarity between measures of strabismus and visual acuity), it was predicted that clinic visitors would have higher CISS scores on average. This pans out, as the average CISS score is higher at clinic visits than in mobile clinic subjects, as shown in Table 16. However, this difference in means is not statistically significant ($p > 0.05$).

Trial Type	95% Conf. Int. Lower	Mean CISS score	95% Conf. Int. Upper	Proportion with CI*	Number of Subjects
Mobile Clinic	15	17	19	0.67	18
Clinic Visit	17	20	23	0.56	15
Volunteers	10	12	14	0.43	7

Table 16: CISS Scores by Event Type

*Symptoms of Convergence Insufficiency are considered present with a score > 15

4.2 Test Policy Analysis

Section 4.2 covers the initial testing results using the methods outlined in Methods subsection 3.2.4; further analysis is compiled in sections 4.3 and 4.4 (covering error analysis and survey analysis respectively). Subsection 4.2.1 provides a brief introduction to the results in section 4.2, as well as some key caveats and limitations of those results. Sections 4.2.2 through 4.2.5 discuss the performance of different testing policies which could be used by QuickCheck. Section 4.2.6 considers QuickCheck's agreement with reference acuity measurement techniques using Cohen's Kappa. Section 4.2.7 goes over crossover between performance and other factors, Section 4.2.8 reviews how different optotypes were read by subjects, and finally Section 4.2.9 briefly summarizes the results of this section.

4.2.1 Clinical Results Introduction

The results from the clinical tests showed that the QuickCheck application can provide an accurate assessment of a subject's visual acuity when it comes to refractive errors at a distance. For both near and distance vision, the 'ground truth' threshold of "poor vision" was a Snellen score of 20/32; any vision worse than that (e.g., 20/33 and worse) was treated as "poor vision".

However, there were several key limitations to these results. Most importantly, the comparison mechanism for near vision problems (near vision cards) was less reliable than the distance vision mechanism (FrACT), meaning that there may have been uncaught cases of subjects with vision problems, or subjects with healthy vision being falsely reported as having poor vision. Importantly, the near vision cards have a limited number of possible acuities they can detect: 20/15, 20/20, 20/30, 20/40, 20/50, and 20/100, whereas FrACT can determine a subject's visual acuity at a much more granular level. In turn, this means that comparing the results for near

vision between QuickCheck and the ‘ground truth’ produced less reliable performance metrics than for distance vision.

In terms of filtering out subjects, there was a single subject whose goal was clearly not to obtain accurate test results, and was instead trying to obtain glasses, even if it meant falsifying their ability to complete QuickCheck and other tests. Due to this, one of the 41 subjects needed to be removed from the subject pool, so there are 40 subjects under consideration in this section.

Before statistical analysis began, an estimation of the 95% confidence intervals for the main metrics used to analyze QuickCheck was performed. Not only did this calculation give a sense of how much more data was needed in order to fully understand the performance of QuickCheck, but this was also used in order to get a sense for what scale of differences between testing policies would be meaningful.

Using the current estimate of the proportion of children with vision problems given by the CDC (0.07), and an estimation that QuickCheck would correctly identify a subject’s vision problems roughly 70% of the time (split evenly between error types), the 95% confidence intervals were estimated to be roughly ± 0.10 for specificity and ± 0.33 for sensitivity [12]. However, it was also thought that the study population could have a much higher rate of vision problems than standard, with the highest estimate being that roughly 35% of subjects would have vision problems. Again, using the estimate of accuracy for QuickCheck of 0.7 for both error types, the 95% confidence interval for sensitivity was estimated to be ± 0.18 for specificity and ± 0.24 for sensitivity.

However, it is also possible to calculate performance by eye rather than by person, which increases the relevant sample size, as each person is tested 6 times. Using these numbers, the low

prevalence estimate of vision problems the 95% confidence intervals were estimated to be ± 0.06 for specificity and ± 0.22 for sensitivity. Using the higher prevalence estimate for vision problems the 95% confidence intervals were estimated to be ± 0.07 for specificity and ± 0.10 for sensitivity. Of course, this estimation is not perfectly accurate, as each data point is not completely independent from another.

To determine which version of the QuickCheck tests was most accurate, each of the individual testing policies were evaluated separately.

4.2.2 Four Correct Policy

As discussed in Section 3.2.2, the Four Correct, Four Incorrect policy was a testing policy that gave a negative (no vision problems) result if the subject read four optotypes correctly in a row, and a positive (vision problems present) result if the subject had attempted to read 10 optotypes without reading four correctly in a row. This policy was based on the idea that, since there are four possible optotypes, a person who could not see the letters accurately would need to guess correctly 4 times over, for a total probability of 1 in 256 (0.3%) of getting four letters correctly in a row without any incorrect guesses. Of course, this basic reasoning falls short in at least two ways. First, imperfect (but not completely absent) vision may allow subjects a better than 1 in 4 chance of reading the correct optotypes. Second, the probability of reading 4 optotypes correctly in a row is not the same as the probability of getting a negative result, as a subject has more than one chance to read four optotypes correctly in a row. Therefore, it was important to determine the actual performance of this policy in context.

For this study, a true positive was when both QuickCheck and the comparison test determine a vision problem is present, a false positive was when QuickCheck finds vision problems, but the

comparison test does not, a true negative was when QuickCheck and the comparison test find no problems, and a false negative was when QuickCheck finds no vision problems but the comparison test does. For QuickCheck, a false negative is a worse type of error than a false positive. This is because a false positive is more likely to be corrected in future testing, but a false negative is likely to remain uncaught as the subject is unlikely to seek further testing for a vision problem they do not believe they have. In this case, the comparison tests differed depending on the distance checked. For distance vision (at 10 feet), the comparison test was the FrACT system, while for near vision (40cm/16in), the comparison was an HOVT card.

Table 17 presents the accuracy, sensitivity, and specificity of QuickCheck using the four-correct policy. The accuracy metric is calculated as the proportion of correctly classified subjects (true positives and true negatives) out of the total number of subjects. The sensitivity is the proportion of true positives to all cases where the subject actually had poor vision, as determined by the comparison test. In other words, the sensitivity represents the probability that QuickCheck would determine that a subject has poor vision provided that they actually had poor vision. Finally, the specificity represents the probability that QuickCheck can determine that a subject has no vision problems given that the subject does not have vision problems. The last column in Table 17 ("N") is the number of individuals who received both a QuickCheck and a comparison test.

One important piece of information is that the estimation of accuracy made in Section 4.1 (when the 95% confidence intervals for specificity and sensitivity were made) was an underestimation. Additionally, the assumption that QuickCheck would make false negative and false positive errors at roughly the same rate appears to have been incorrect as well, which further decreases the reliability of the initial confidence interval estimates. Information about confidence intervals

of performance is discussed in section 4.3 alongside a discussion of what types of errors occur and why.

As shown in Table 17, QuickCheck's performance when detecting near vision problems was worse than its ability to detect distance vision problems. Comparing near vision specificity and sensitivity, this is because of a higher false positive rate for near vision than for distance vision. That is, QuickCheck was likely to determine that a subject had poor near vision even when the subject had good near vision using the four-correct policy. However, while distance vision had a higher accuracy than near vision overall, the false negative rate was higher for distance vision than for near vision. While no type of error is good, for QuickCheck (as with many diagnostic tests) false negatives are worse than false positives, because it is worse for a user to believe their vision is good when it isn't (and therefore potentially not seek further treatment) than the reverse. This means that the performance of near vision represents a tradeoff against the performance of distance vision and given the importance of having a low false negative rate for a screening test, from a certain point of view it is near vision that has better performance, despite the lower accuracy.

Also notice that in distance vision, the performance of the application is worse with both eyes than with either individual eye for all metrics. There are a number of different plausible causes for this difference; for example, it could be that because subjects get more accurate test results as they continue to use the application (the order of tests is near both, near right, near left, then distance both, distance right, distance left). If that is the case, then it could also be an explanation for why distance vision performs better than for near vision, as near vision is tested before distance vision. However, timing alone is unlikely to explain the entire difference in false

negative rates between distance types, as specificity and sensitivity do not follow the same pattern as accuracy with regards to test timing.

Distance-Eye Category	Accuracy	Sensitivity	Specificity	N
Near Vision				
Both Eyes	0.78	0.92	0.71	40
Left Only	0.82	1.00	0.73	40
Right Only	0.72	0.86	0.65	40
Distance Vision				
Both Eyes	0.85	0.77	0.89	40
Left Only	0.92	0.94	0.92	40
Right Only	0.85	0.76	0.91	40
Average	0.82	0.88	0.8	40

Table 17: Performance Metrics of the Four Correct Testing Policy

Next, the shuffling test as described in section 3.2.4 was performed. This test was designed to examine how much variance there could be for a subject by randomly re-ordering the tests that are used to simulate a given testing policy being used. If a subject's proportion of shuffled positive results was close to 50%, that indicates that the test results were 'on the edge'; that is, there are many letter orderings where the subject may have received a different result than they did. Similarly, the 'matching rate' (that is, the proportion of shuffled results that are true positives or true negatives using a given testing policy) being close to 50% indicates that this subject's QuickCheck results were close to being correct or incorrect. Note though that just using

the mean positivity or matching rates is not sufficient to understand QuickCheck's performance. Instead, the proportion of uncertain cases (that is, the proportion of shuffled results which are close to 50%) is also important to know. Both proportions together build a clearer picture of QuickCheck's performance than either alone.

Table 18 shows the positivity rate, matching rate, and uncertainty proportions for both using the four positives policy (counting all proportions between 5% and 95% as being "uncertain"). In all cases, the proportion of uncertain cases was below 0.15, which indicates a strongly bimodal distribution of the proportion of test results around 0 and 1. This indicates that the QuickCheck test was usually confident in its predictions; that is, there are few cases where a random re-ordering of letters provided to the subject could have changed the outcome.

Looking at Table 18, the positivity rate means were much higher for near vision than for distance vision, with near vision's mean positivity rate being well above the actual rate of poor near vision in the study, which was roughly 45%. Distance vision's positivity rate was much closer to the actual rate of distance vision problems in the study population, which reinforces the higher false positivity rate for near than for distance vision tests. On the other hand, all distances and eyes had relatively low uncertainty proportions, indicating that QuickCheck tended to be fairly 'certain' of its decisions regarding which eyes had vision problems.

Distance-Eye Category	**Positivity Rate Mean**	**Positivity Uncertainty Proportion**	**Matching Rate Mean**	**Matching Uncertainty Proportion**
Near Vision				
Both Eyes	0.73	0.06	0.50	0.06

Distance-Eye Category	Positivity Rate Mean	Positivity Uncertainty Proportion	Matching Rate Mean	Matching Uncertainty Proportion
Left Only	0.78	0.00	0.66	0
Right Only	0.82	0.11	0.50	0.11
Distance Vision				
Both Eyes	0.38	0.17	0.64	0.17
Left Only	0.55	0.06	0.83	0.06
Right Only	0.48	0.17	0.81	0.17
Average	0.63	0.09	0.66	0.09

Table 18: Comparison of Shuffled and Actual Results for the Four Correct Test Policy (n=40)

While QuickCheck did store the time each test took, due to difficulties caused by delays during testing, using actual testing time as recorded by the application is not a good way of estimating the time a test actually took. A better estimate is to use the number of letters displayed to each subject, as that is the component of the QuickCheck application which is actually affected by changing the testing policy. Using the mean number of letters, the time each testing sequence would was estimated take by assuming that each letter would take about five seconds to be read by the subject. This estimation is based on the actual time letters took to read during clinical testing, but it should not be thought of as being highly accurate, as letters often took very different amounts of time to be read by different subjects.

Table 19 shows that the number of letters read was very different for near and distance vision tests. This is because near vision letters were more likely to be read incorrectly, as more near

vision tests concluded that subjects had poor vision. However, note that subjects who could not see any optotypes at all actually finished near vision tests faster, as they requested that the tester simply skip through the tests at speed and therefore took much less time per optotype seen. The average subject would have seen between 41 and 42 letters using the four correct policy, which is about 70% of the 60 letters that were actually shown to them. This represents an estimated time savings of about 93 seconds (just over a minute and a half) using the four correct policy as compared to the ten tests policy that was used in clinical testing.

Distance-Eye Category	**Mean Number of Letters**	**Mean Time Estimation (s)**
Near Vision		
Both Eyes	7.2	36
Left Only	7.41	37.05
Right Only	7.44	37.2
Distance Vision		
Both Eyes	6.12	30.6
Left Only	6.66	33.3
Right Only	6.59	32.95
Total	41.42	207.1

Table 19: Mean Number of Tests Per Eye Using the Four Correct Testing Policy (n=40)

The four-correct policy was the initial policy designed for QuickCheck; therefore, the rest of the policies will be examined using the four-correct policy as a baseline.

4.2.3: Two Correct Policy

Similar to the "four correct" policy described in the previous section, the "two correct" policy requires a certain number of correct or incorrect guesses in a row before the test concludes. As the name suggests, the "two correct" policy determined that a subject has good eyesight if they read two letters correctly in a row, and poor eyesight if they read two letters incorrectly in a row. This policy was expected to be faster but perform less well when compared with the four-correct policy. As shown in Table 20, the expectation that this policy would perform less well than the four correct policy was largely correct. The cells in Table 20 which contain higher metric scores than their corresponding cells in Table 17 have been highlighted in green to better show where the two policies differ, and cells where the two-correct policy performed worse are highlighted in orange. In other words, in all non-highlighted cells, the four-correct policy performs the same as the two-correct policy (to two decimal places of accuracy).

The two-correct policy has slightly worse sensitivity and slightly better sensitivity in the near vision-right only row, but these changes are minimal and are not significant. The policies differ much more in the distance vision tests, where the two correct policy performs worse than the four-correct policy in all cases except left eye specificity and right eye accuracy and specificity. However, it is important to note that no difference is large enough to be statistically significant, using the (admittedly quite limited) estimated 95% confidence interval for specificity and sensitivity calculated in Section 4.2.1 above. Although there are numerous problems with that initial error bound estimation (discussed in Sections 4.2.1 and 4.2.2), but as a standard to compare against it suggests caution when interpreting these results, as there really is not a strong indication that either policy performs better than the other beyond what could be caused by random chance.

Overall, though, Table 20 shows that the two-correct policy performs worse than the four-correct policy regarding specificity, better regarding sensitivity, and roughly the same regarding accuracy. The similar performance of the two testing was unexpected, as the two-correct testing policy was expected to be significantly more error-prone than the four-correct study. This discrepancy emphasizes the low sample size of the study, and the fact that most subjects are identified confidently by QuickCheck (as shown in Table 18 and Table 21). The change in sensitivity and specificity means that the two-correct policy is less likely to produce false positives, but more likely to produce false negatives than the four-testing policy. As discussed in section 4.1.1, if accuracy stays roughly the same, it is better for QuickCheck to have fewer false negatives and more false positives than the reverse. Therefore, the two-testing policy, while being similar to the four-testing policy in accuracy, produces more of the worse type of error, so it is inferior to the four-testing policy in terms of performance.

Distance-Eye Category	**Accuracy**	**Sensitivity**	**Specificity**	**N**
Near Vision				
Both Eyes	0.75	0.83	0.71	40
Left Only	0.82	0.93	0.77	40
Right Only	0.7	0.71	0.69	40
Distance Vision				
Both Eyes	0.85	0.69	0.93	40
Left Only	0.9	0.88	0.92	40
Right Only	0.9	0.76	1	40

Distance-Eye Category	Accuracy	Sensitivity	Specificity	N
Average	0.82	0.8	0.84	40

Table 20: Performance Metrics of the Two Correct Testing Policy

As in Table 18, Table 21 shows that the two correct testing policy produces a bimodal distribution with modes around 0 and 1. The uncertainty proportionals never exceeds 0.07 for any eye-distance pair, which means that there are very few cases where QuickCheck was even slightly likely to produce a different result given a re-ordered set of letters.

Again as in Table 18, Table 21 shows that the near vision tests are more likely to produce a false positive than a false negative, although the positivity proportions are somewhat lower for the two policy test, which reinforces the finding that the two correct policy has a lower false positive rate than the four correct policy.

Distance-Eye Category	Positivity Rate Mean	Positivity Uncertainty Proportion	Matching Rate Mean	Matching Uncertainty Proportion
Near Vision				
Both Eyes	0.3	0.06	0.64	0.06
Left Only	0.45	0.06	0.83	0.06
Right Only	0.35	0.11	0.85	0.11
Distance Vision				
Both Eyes	0.72	0.00	0.5	0.00

Distance-Eye Category	Positivity Rate Mean	Positivity Uncertainty Proportion	Matching Rate Mean	Matching Uncertainty Proportion
Left Only	0.72	0.11	0.61	0.11
Right Only	0.72	0.00	0.5	0.00
Average	0.54	0.06	0.66	0.06

Table 21: Comparison of Shuffled and Actual Results for the Two Correct Test Policy (n=40)

Table 22 shows the mean number of letters read by each test for each distance-eye pair. Note that again, as in Table 19, the number of letters read for near vision tests is higher than in distance vision tests. This is because more subjects needed to read all or nearly all of the letters for near vision tests, as more subjects were determined to have poor vision at close range. Compared to the four-correct policy, Table 22 shows that the two-correct policy has a slight decrease in mean number of letters shown per test for near vision, and a larger decrease for distance vision. Overall, the average subject would have seen between 40 and 41 letters, a decrease of roughly 7 letters from the four-testing policy. The fact that the two-correct policy did not roughly halve the number of letters seen, as one might initially suspect, is because subjects which would be unable to read any letters (or could only read a few letters) would have read roughly the same number of letters for the four- and two-correct policies. Considering that there is only an estimated reduction of about 45 seconds of runtime when moving from the four-correct to the two-correct policy, the increase in false negative rate using the two-correct policy means that, of the two, the four-correct policy is the better choice for a final QuickCheck policy.

Distance-Eye Category	Mean Number of Letters	Mean Time Estimation (s)
Near Vision		
Both Eyes	5.88	29.4
Left Only	6.05	30.25
Right Only	6.07	30.35
Distance Vision		
Both Eyes	4.54	22.7
Left Only	5.24	26.2
Right Only	4.73	23.65
Average	32.51	162.55

Table 22: Mean Number of Tests for the Two Correct Testing Policy (n=40)

4.2.4: Five Tests Policy

Per the name of this testing policy, the "five tests" policy requires that QuickCheck's subject be tested with 5 letters; the subject passed (negative result) if at least 3 of the 5 presented letters are read correctly and did not pass (positive result) otherwise. Note that this is different from the "four correct" policy in that fewer letters must be read correctly or incorrectly to end the test per eye. Table 23 shows the performance metrics of QuickCheck using the five tests policy. Cells which differ from their corresponding cells in Table 17 have been highlighted. If the five-tests policy improves on a metric score from the four-correct policy, the cell will be highlighted in green. If the five-correct policy performs worse, the cell will be highlighted in orange. Of the metrics where the two policies differ, the four-correct policy exceeded the five tests policy only

in the sensitivity of the near vision tests; the difference in both the right-only and left-only sensitivities was 0.07.

In all other cases where the two policies differed, the five tests policy outperformed the four-correct policy. However, the differences between the policies' performances were very small; the five tests policy outperformed in accuracy by 0.02, in sensitivity it underperformed on average by 0.03, and in specificity it overperformed on average by 0.03. Because these differences were small, and the sample size is low, we cannot say with any confidence that the five tests policy would continue to outperform in a similar way in a larger study population. Additionally, the five tests policy had the same general results as the two- and four- correct policies; that is, the five tests policy had a high false positive rate for near vision tests, a high false negative for distance vision tests, and tended to have a higher sensitivity than specificity.

Overall, then, it is difficult to say if the five-tests policy is the best performing test policy, as it decreases false positive errors and increases accuracy but increases the false negative rate as compared to the four-correct policy. Therefore, choosing the best policy will come down to other factors, such as how long the five-tests policy takes in comparison to the four-tests policy.

Distance-Eye Category	**Accuracy**	**Sensitivity**	**Specificity**	**N**
Near Vision				
Both Eyes	0.78	0.92	0.71	40
Left Only	0.8	0.93	0.73	40
Right Only	0.72	0.79	0.69	40
Distance Vision				

Distance-Eye Category	Accuracy	Sensitivity	Specificity	N
Both Eyes	0.88	0.77	0.93	40
Left Only	0.95	0.94	0.96	40
Right Only	0.9	0.76	1	40
Average	0.84	0.85	0.84	40

Table 23: Performance Metrics of the Five Tests Testing Policy

The shuffling results for the five tests policy (Table 24) were very similar to the shuffling results for the two- and four- correct policies. Again, the positivity rate was high for near vision tests, and the matching rate was higher for distance vision tests than for near vision tests. The uncertainty proportions are higher on average here than in the four-correct policy, which is expected as there are fewer tests using the five-test policy when subjects fail to read several letters. What is surprising is that the five tests policy had a higher uncertainty proportion than the two correct tests, which is unexpected because the five tests policy should show more optotypes when the subject has read several letters correctly initially. One explanation for this is that those who read even two letters correctly in a row are very likely to pass using the five tests policy, but the reverse is less true. However, the difference in uncertainty proportion is small, so it is difficult to say that, with a larger sample size, the five tests policy would remain more uncertain than the two tests policy.

Distance-Eye Category	Positivity Rate Mean	Positivity Uncertainty Proportion	Matching Rate Mean	Matching Uncertainty Proportion
Near Vision				
Both Eyes	0.34	0.11	0.65	0.11
Left Only	0.51	0.11	0.82	0.11
Right Only	0.43	0.17	0.84	0.17
Distance Vision				
Both Eyes	0.72	0.00	0.5	0.00
Left Only	0.76	0.06	0.65	0.06
Right Only	0.78	0.11	0.5	0.11
Average	0.59	0.09	0.66	0.09

Table 24: Comparison of Shuffled and Actual Results for the Five Tests Policy (n=40)

Unlike the two- and four- correct policies, the five tests policy has the same number of letters read for each test. With five letters per test, this results in 30 letters being read per subject in total, or an estimated 150 seconds. This is faster than the four-correct policy by an estimated 87.5 seconds (17.5 letters on average fewer). To further increase the speed of testing, it is possible to use a shortened five tests policy. This policy would produce identical outcomes to the original five tests policy, but it would require fewer letters read by ending a test early as soon as the subject reads 3 correct or incorrect letters. Table 25 shows the mean number of letters read and estimated time for each test type using this shortened five tests policy. The shortened version of the five-correct policy has a moderate impact on testing length as compared to the original (a

decrease of only 6.3 letters on average in total), but it represents the shortest version of the five-tests policy that does not result in a different determination for any tests. This makes the shortened five-tests policy not only the fastest policy (by 8.8 optotypes as compared to the next shortest, the two-correct policy), but also arguably the best performing.

Distance-Eye Category	**Mean Number of Letters**	**Mean Time Estimation (s)**
Near Vision		
Both Eyes	4.07	20.35
Left Only	4.05	20.25
Right Only	4.15	20.75
Distance Vision		
Both Eyes	3.71	18.55
Left Only	3.93	19.65
Right Only	3.8	19
Average	23.71	118.55

Table 25: Mean Number of Tests for the Shortened Five Tests Policy

As the five tests policy has similar performance to (and is significantly faster than) than the four-correct policy, the five-tests policy is likely to be a better policy in general for the QuickCheck application. Some divergence in results between these two policies was expected, as the four-correct policy is less likely to result in false negatives given the requirement that 4 optotypes be read correctly, and that those optotypes must be read correctly one after the other. However, the bimodal distribution of test results (demonstrated though the shuffling results of both policies)

means that the minimal difference in false negative rates between the two policies is unlikely to produce very different outcomes between the two tests.

4.2.5: Ten Tests Policy

Per the name of this testing policy, the "ten tests" policy requires that QuickCheck's subject be tested with 5 letters; the subject passes (negative result) if at least 6 of the 10 presented letters are read correctly and does not pass (positive result) otherwise. As the ratio of required correctly read letters to total letters read remains the same as in the "five tests" policy, the "ten tests" policy was not expected to significantly diverge from the results of that policy. Instead, the "ten tests" policy was intended to provide a more confident version of the five tests policy. That is, it was expected that the ten tests policy would be more likely to sort subjects who were on the edge of having vision problems. Additionally, the ten tests policy was intended to be a point of comparison for the four-correct policy, which ended at ten tests if no series of four letters have been read correctly.

As shown in Table 26, the results of the five tests and the four correct policies were very similar to the ten tests policies. All cells where the ten tests and four correct policies differ are highlighted in Table 26 for ease of reading; cells in orange represent metrics where the four-correct policy performed better, and cells in green are where the ten-tests policy performed better. Overall, the four-correct policy had a better performance in near vision, while the ten-tests policy had a better performance in distance vision. Additionally, all of the performance gains using the ten tests come from decreasing false negatives at a distance, while all gains from the four-correct policy come from decreasing false positives for near vision. Furthermore, the difference between the two policies is quite small; their accuracies are the same on average, the four-correct policy improves on average sensitivity by 0.04, and the ten-tests policy improves on

average specificity by 0.02, all of which is much less than even the most generous error bounds estimated.

The fact that the five-test policy performs better than the ten-test policy is unexpected. It had been anticipated that the ten-test policy would present a tradeoff of increased accuracy for increased time spent as compared to the five-test policy, particularly due to the proportion of optotypes which must be read correctly for a negative result (which is 2/3 for both policies). Given the low sample size of this trial, it is likely that these differences are simply due to random chance. However, if they are not due to chance, then it is an indication that longer tests do not necessarily perform better. In fact, the higher false negative rate of the ten-tests policy may indicate that subjects can learn to perform better on the tests the more optotypes they are shown in a way that decreases QuickCheck's ability to detect their vision problems, although there is little way to test that hypothesis with the currently available data.

Distance-Eye Category	**Accuracy**	**Sensitivity**	**Specificity**	**N**
Near Vision				
Both Eyes	0.75	0.83	0.71	40
Left Only	0.8	0.93	0.73	40
Right Only	0.7	0.79	0.65	40
Distance Vision				
Both Eyes	0.88	0.77	0.93	40
Left Only	0.92	0.94	0.92	40

Distance-Eye Category	Accuracy	Sensitivity	Specificity	N
Right Only	0.88	0.76	0.96	40
Average	0.82	0.84	0.82	40

Table 26: Performance Metrics for the Ten Tests Testing Policy

As with the five tests policy, the ten tests policy had a constant number of letters read per test, with a total of 60 letters read by all subjects during clinical testing. However, the number of letters subjects need to read can be reduced by ending the test early when a majority of letters (6 or more) have been read correctly (or incorrectly). Table 27 shows the mean number of letters read and estimated time for each test. Using this shortcutting strategy, about 12.6 optotypes (63 seconds) can be saved per subject as compared to the original ten-tests policy. However, as there is very little difference in performance between the five- and ten-tests policies, and the five tests policy takes less than half the time of the ten tests policy, the five tests policy is the superior choice.

Distance-Eye Category	Mean Number of Letters	Mean Time Estimation (s)
Near Vision		
Both Eyes	8.12	40.6
Left Only	8.17	40.85
Right Only	8.24	41.2
Distance Vision		

Distance-Eye Category	Mean Number of Letters	Mean Time Estimation (s)
Both Eyes	7.44	37.2
Left Only	7.76	38.8
Right Only	7.63	38.15
Total	47.36	236.8

Table 27: Mean Letters and Estimated Time Using the Shortened Ten Tests Policy (n=40)

4.2.6 Cohen Kappa Analysis

In order to verify the performance of each of the four testing policies discussed thus far, an analysis was performed using Cohen's Kappa (discussed in Section 2.4.1.2). See Table 28 for the kappa and p-value of each testing policy. Using the 0.60 threshold as a standard for measurements that are in agreement, it is clear that overall, each policy is in agreement with the reference acuity measurement methods. However, when considered by sub-group, there are several points of concern. First, agreement is lower for near vision than distance vision across all vision distances, with some agreements being very low (below 0.40). These concerning cases have been marked in red for clarity. The most worrying result is the poor agreement on near vision for adults, with no policy exceeding a kappa value of 0.40, indicating little agreement between QuickCheck and the reference measurement (near vision cards) among adults.

P-values presented are calculated using the null hypothesis of 'Cohen kappa = 0.0' and an alternative hypothesis of 'Cohen kappa > 0'. This means that all the p-values indicate is that there it is unlikely that the agreement between QuickCheck, near vision cards (near vision), and FrACT (distance vision) is completely due to chance. They do not indicate that it is unlikely that the strength of any given agreement is due to chance. Additionally, with concerns about multiple

hypothesis testing, it is still possible that at least one kappa value is non-zero due to random chance despite the relatively small p-values.

When it comes to choosing the best policy, the Five Tests policy has the best overall kappa for all cases, although note that it performs slightly worse than the four-correct policy for near vision. However, considering the lower false negative rate and faster testing time for the five tests policy as compared to the four-correct policy, it is still the stronger candidate for QuickCheck's future testing policy.

Distance-Eye Category	**Four Correct Kappa (p-value)**	**Two Correct Kappa (p-value)**	**Five Tests Kappa (p-value)**	**Ten Tests Kappa (p-value)**
Children				
Near Vision	0.62 ($p < 0.0001$)	0.26 ($p = 0.0075$)	0.56 ($p < 0.0001$)	0.33 ($p = 0.018$)
Distance Vision	0.85 ($p < 0.0001$)	0.73 ($p < 0.0001$)	0.92 ($p < 0.0001$)	0.85 ($p < 0.0001$)
Adults				
Near Vision	0.38 ($p < 0.0001$)	0.35 ($p = 0.0004$)	0.35 ($p = 0.0004$)	0.35 ($p = 0.0004$)
Distance Vision	0.64 ($p < 0.0001$)	0.70 ($p < 0.0001$)	0.72 ($p < 0.0001$)	0.70 ($p < 0.0001$)
All Ages				
Near Vision	0.55 ($p < 0.0001$)	0.50 ($p < 0.0001$)	0.53 ($p < 0.0001$)	0.50 ($p < 0.0001$)
Distance Vision	0.73 ($p < 0.0001$)	0.75 ($p < 0.0001$)	0.80 ($p < 0.0001$)	0.77 ($p < 0.0001$)
All cases	0.64 ($p < 0.0001$)	0.62 ($p < 0.0001$)	0.66 ($p < 0.0001$)	0.62 ($p < 0.0001$)

Table 28: Cohen's Kappa and P-values for Each Test Policy Across Age Category and Distance

4.2.7 Performance across Population Types

As the five-correct policy is the best performing of the policies, that policy will be used to examine the differences between QuickCheck's performance across age groups. Table 29 shows the performance of QuickCheck for adults, and Table 30 shows the performance of QuickCheck for children. There is a noticeable difference between the performance of the policy on adults and on children, in that QuickCheck performs noticeably better on children than on adults. The differences in overall accuracy, specificity, and sensitivity between children and adults are statistically significant ($p < 0.001$ for all three using t-tests), although given the small sample sizes, it is difficult to know if this is actually a meaningful finding. This improvement in accuracy with children is a good sign, as QuickCheck is primarily intended to be used on children. Additionally, the sensitivity is quite high on average in children, which indicates a low false negative rate. However, there are several cases where sensitivity is very low in children but given the small sample size for children in this thesis, it is difficult to determine if this result will continue in larger study sizes. For example, there are roughly 4 children who have vision problems in each eye-distance category, which means that measures of sensitivity for children will be highly imprecise. Therefore, it is difficult to draw any conclusions from these results except that there are some promising initial results in children.

Distance-Eye Category	**Accuracy (Adults)**	**Specificity (Adults)**	**Sensitivity (Adults)**	**N Adults**
Near Vision				
Both Eyes	0.67	0.43	1	24
Left Only	0.71	0.45	0.92	24

Distance-Eye Category	Accuracy (Adults)	Specificity (Adults)	Sensitivity (Adults)	N Adults
Right Only	0.62	0.38	0.91	24
Distance Vision				
Both Eyes	0.79	0.85	0.73	24
Left Only	0.92	0.91	0.92	24
Right Only	0.88	1.00	0.79	24
Average	0.67	0.43	0.79	24

Table 29: QuickCheck Performance using Five-Tests Policy for Adults Only

Distance-Eye Category	Accuracy (Children)	Specificity (Children)	Sensitivity (Children)	N Children
Near Vision				
Both Eyes	0.94	1.00	0.5	16
Left Only	0.94	0.93	1.00	16
Right Only	0.88	1.00	0.33	16
Distance Vision				
Both Eyes	1.00	1.00	1.00	16
Left Only	1.00	1.00	1.00	16
Right Only	0.94	1.00	0.67	16
Average	0.95	0.96	0.92	16

Table 30: Performance of QuickCheck's Five-Tests Policy on Children Only

There are several other population types that are of interest, but unfortunately most of those subgroups are too small for in-depth examination to be useful. For example, the best type of data would come from children who do not have eyeglasses as those would be the subjects most representative of QuickCheck's target subject-base. Unfortunately, there are relatively few subjects who do not have glasses (only 10 out of the sample) and of those, only 8 are children. Using all of the non-eyeglass owning subjects, the average accuracy across all distances and eye types is 0.87, the average specificity is 0.46, and the average sensitivity is 0.96; this means that the false negative rate is low (2 out of 7 eyes with vision problems were predicted not to have vision problems) but the false positive rate is high (5 out of 11). However, as shown by the very low denominators in those example rates, there really is not enough data to draw out more results than "preliminary results are promising."

Another subgroup that would be interesting to examine are very young children (between 5 and 10). Unfortunately, no child less than 7 years old was in the sample, and only 7 subjects are less than 10. Again, preliminary results are promising, with an accuracy of 0.90, specificity of 0.6, and sensitivity of 1.00, but there are just too few subjects in this age range to draw a conclusion from.

4.2.8 Optotype Letter Analysis

This section briefly covers the types of optotypes shown to subjects. One concern is that different optotypes are read differently by different subjects; for example, during testing there was the sense that perhaps Ts are harder to read than Hs, especially among younger subjects. Table 31 shows the proportion of optotypes that are read correctly; none of the differences in proportions are statistically significant, either across letters or across age. Figure 29 shows the overall proportion of optotypes read correctly for all age groups by letter and distance tested. This

indicates that, although different groups have different baseline vision acuities (e.g., fewer optotypes are read correctly at near distances as compared to distance vision), different optotypes are not read correctly at different rates. Additionally, the fact that the standard deviations are so large indicates a very wide spread of optotypes read correctly. This is to be expected, as there are a number of subjects who read every optotype correctly as well as a number of subjects who read every optotype incorrectly.

Distance-Eye Category	**H Proportion Correct ± SD**	**O Proportion Correct ± SD**	**V Proportion Correct ± SD**	**T Proportion Correct ± SD**
Children				
Near Vision	0.93 ± 0.26	0.90 ± 0.31	0.90 ± 0.31	0.80 ± 0.40
Distance Vision	0.85 ± 0.35	0.85 ± 0.35	0.90 ± 0.31	0.81 ± 0.39
Adults				
Near Vision	0.24 ± 0.43	0.22 ± 0.42	0.25 ± 0. 44	0.26 ± 0.44
Distance Vision	0.61 ± 0.49	0.59 ± 0.49	0.57 ± 0.50	0.63 ± 0.49
All Ages				
Near Vision	0.53 ± 0.50	0.51 ± 0.50	0.50 ± 0.50	0.53 ± 0.50
Distance Vision	0.71 ± 0.45	0.71 ± 0.46	0.71 ± 0.46	0.71 ± 0.45
All cases	0.63 ± 0.48	0.62 ± 0.49	0.60 ± 0.49	0.62 ± 0.49

Table 31: Proportion of Optotypes Read Correctly by Age and Distance

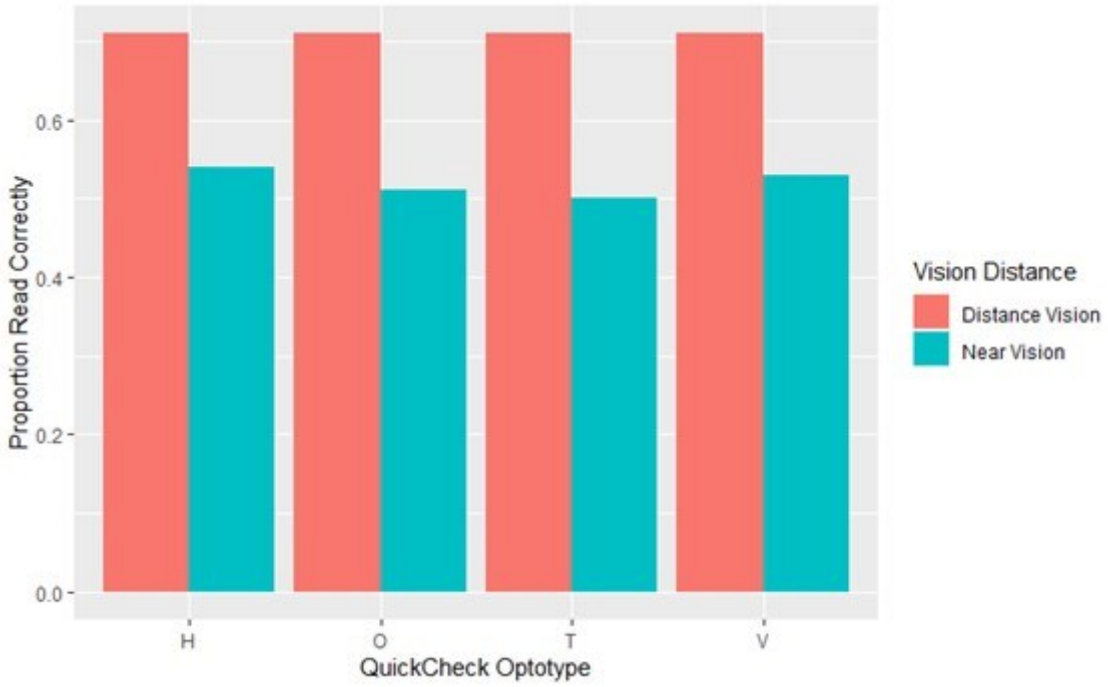

Figure 29: Proportion of Optotypes Read Correctly by Distance Tested and Optotype Letter

4.2.9 Summary of Testing Policy Results

All of the testing policies performed roughly the same, even the two-correct policy, which was unexpected, although overall the five-tests policy performed the best. All tests were expected to perform roughly the same (except for the two-correct policy); however, the very bimodal distribution of test results for all policies meant that there were very few subjects which were sorted differently by different testing policies. This means that the difference in testing policies would have been largely in the number of letters a subject would have to read.

The testing policy which seemed to strike the best balance between performance and speed was the shortened five test policy, which the shortest estimated time by far, and roughly the same performance as the four correct policy. Additionally, the five-tests policy performed very well with children, but the limited sample of children means that there are limitations to how well that performance can be expected to continue in the future. Furthermore, no policy performed

differently enough on average from the others to be certain that differences were not due to random chance, at least based on the initial confidence interval estimation for sensitivity and specificity.

Using the five-tests policy, QuickCheck achieved an average sensitivity of 0.85, which is very close to GoCheck Kids sensitivity of 0.81 (see section 2.3.3.3), although QuickCheck's average specificity was much lower than GoCheck Kids at 0.84 as compared to 0.91 [44] [48]. Similarly, QuickCheck's performance was comparable to the Spot Vision Screener (SVS), with an average sensitivity of 0.85 and specificity of 0.84 as compared to 0.85 specificity and 0.80 sensitivity for the SVS [48]. Of course, given the large confidence intervals for sensitivity specifically, it is currently very possible that QuickCheck performs worse (possibly much worse) than GoCheck Kids or the SVS.

However, even the best performing policy (five-tests) still had problems. Overall performance was worse for near vision than for distance vision, and distance vision had a higher false negative rate than false positive rate (which is bad for any screening tests). There are some hints that using smaller optotypes at a distance could decrease the false negative rate, although that would likely increase the false positive rate as well. It is also important to note that the comparison test used for near vision (a near vision card) was not as reliable as FrACT, which was used for distance vision. It is therefore possible that the distance vision tests are a more reliable estimation of QuickCheck's accuracy than the near vision test performance.

4.3 Error Analysis

Section 4.3 discusses the types and causes of errors made by QuickCheck. Given the findings of section 4.2, this section will focus on the errors made by the highest-performing rule, the three-

of-five rule. Section 4.3.1 provides an overview of the predictions made by that rule and section 4.3.2 considers the relationship between the false positives and false negative predictions made by QuickCheck.

In this section of the analysis, the unit of concern is usually not the subject, but rather a given test result for a specific eye at a specific distance. For example, a point on a scatter plot will represent an 'eye' (the result of testing a specific eye category [left eye only, right eye only, or both eyes] at a particular distance) rather than a whole subject.

4.3.1 Three-of-Five Error Overview

In general, there are two kinds of errors to be concerned about. First, there are false positives, which occur when QuickCheck 'detects' that a subject has Snellen visual acuity worse than 20/32, but the subject's actual visual acuity is better than 20/32. False negatives, on the other hand, occur when QuickCheck does not detect that a subject has vision worse than 20/32, but they do have vision problems. As QuickCheck is a screening tool, false negatives are a much worse outcome than false positives, because a false positive means that a student will get referred for further testing which can rule out vision problems, but a false negative means that a student with vision problems will not get them treated. Therefore, QuickCheck should focus primarily on reducing false negatives, even if that means increasing false positives to some degree.

Additionally, because each subject is tested multiple times for each eye and distance but would be referred as a whole person to an optometrist or ophthalmologist on any failed result, the worst case scenario is someone with a vision problem who is not detected at all. For example, it is better for a given subject to have their near vision problems detected and their distant vision problems not detected because they will still be referred to an eye exam if that occurs. Therefore,

there are actually two types of false negatives: non-referring false negatives (where a subject's vision problems are not detected at all) and referring false negatives (where at least one of a subject's vision problems are detected but some are not). These will be referred to as NRFNs and RFNs respectively throughout this section.

With that in mind, consider Figure 30, which shows the types of predictions that the three-of-five rule makes for each subject's eyes. Each point represents a single eye's vision (or both eyes measured at once). Each axis represents the Snellen acuity denominator of a subjects eyes with 20 as the numerator, with the horizontal axis representing near-vision acuity and the vertical axis representing distance-vision acuity. This means that moving up the graph represents having worse distance vision, and moving right represents having worse near vision. The black lines represent 20/32 vision at each distance, such that a point above the horizontal line represents a person whose vision is worse than 20/32 at 10 feet (and the same for the vertical line and near vision). Each point is colored by QuickCheck's prediction for that eye, either near vision problems only, distance vision problems only, both ("General vision problems"), or neither.

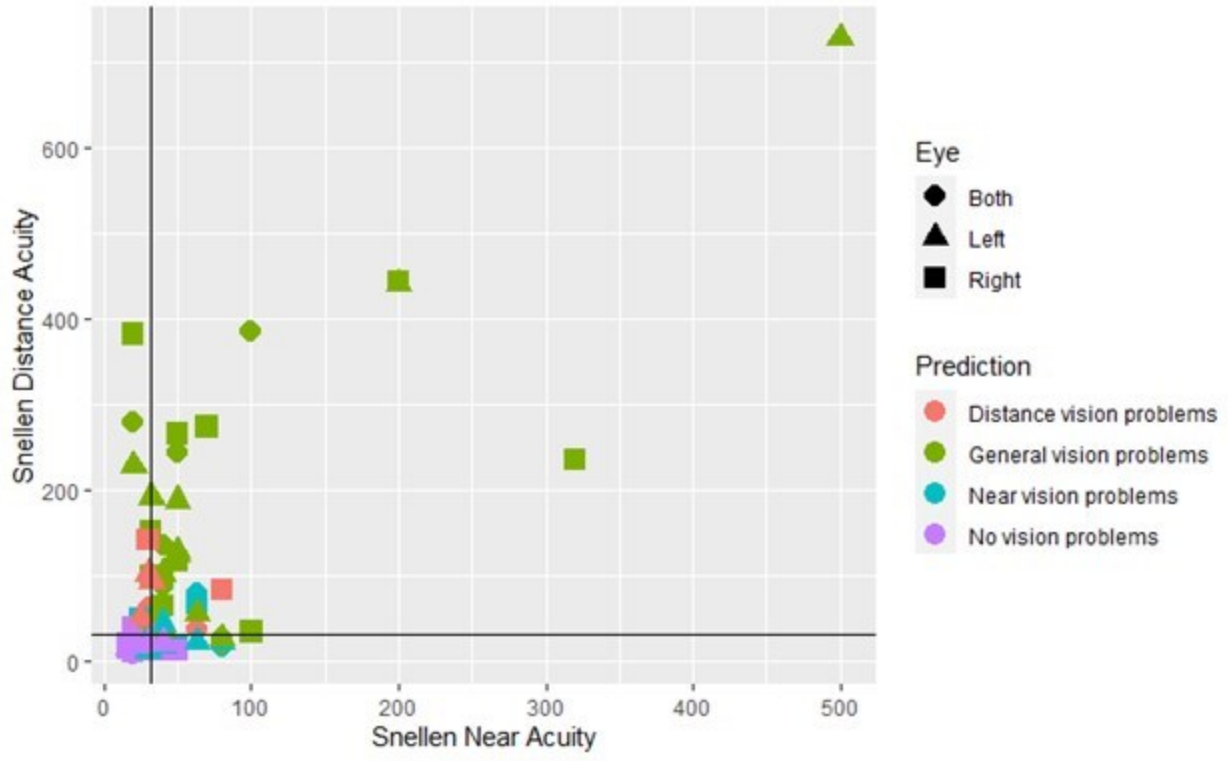

Figure 30: Scatter Plot of Snellen Near vs Distance Acuity Scores with Outliers

However, due to the presence of a number of outliers, it is somewhat difficult to read Figure 30. Therefore, Figure 31 shows the same data with all eyes which have visual acuity worse than 20/100 at 16in or 20/200 at 10 feet removed. For convenience, three of the 'quadrants' made by the lines at 20/32 have been labeled. Quadrant 1 contains eyes that have both near and distance vision problems, quadrant 2 contains eyes that have only distance vision problems, the unlabeled quadrant 3 contains eyes that have no vision problems, and quadrant 4 contains eyes that have only near vision problems. If QuickCheck using the three-of-five rule was perfectly accurate, all eyes in quadrant 1 would be green, all eyes in quadrant 2 would be red, all eyes in quadrant 4 would be blue, and all eyes in quadrant 3 would be purple.

From a brief examiniation it appears that QuickCheck is fairly good at distinguishing those who have no vision problems from those who have at least one vision problem, but it is less good at

distinguishing between the types of vision problems a subject has. This is shown in Figure 31 (and Figure 30) as a relative lack of purple points outside quadrant 3, but a variaty of different predictions in all other quadrants.

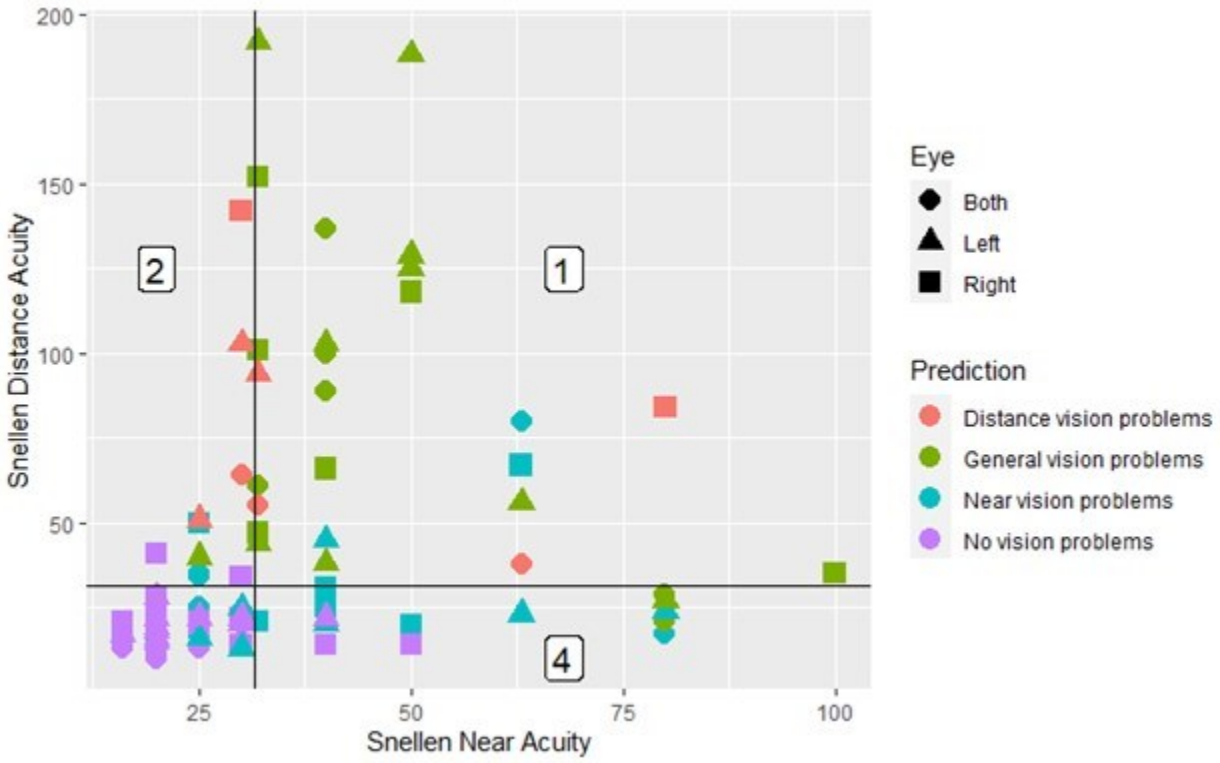

Figure 31: Snellen Acuity Scores for Distance vs Near Vision without Outliers

4.3.2 False Positives and False Negatives

Taking a look at the types of errors in more detail, Table 32 shows the accuracy, false positive rate, and false negative rate of the three-of-five rule's predictions. Note that the accuracies in this table will be different than those in Table 23, because rather than considering each eye at each distance independently, all eyes at each distance and all distances for each eye are grouped together. Margins of error in the form of 95% confidence intervals are provided for the performance metrics for all eyes and distances. Under this view, the False Positive and False Negative rates are on average very similar to each other. However, there appears to be a sort of

tradeoff: when one error rate is high, the other tends to be low, and visa-versa. Of course, there are some categories (left eye at any distance), for which both error rates are relatively high.

However, there is some variation in where the false negative rate is high; in particular, the false negative rate is lower for near distance and right eyes than for other distances and eyes. It is unclear why the right eye performs better than the left eye in terms of false negatives. There are a few possible causes, but none seem to jump out as the obvious source. It might be that the difference in this populations acuity between right and left eyes is to blame, but that is hard to understand because the left eyes are worse on average, which would make false negatives seem less likely. Order of tests might play a role, but because right eyes are tested between both eyes and left eyes only, it is unclear why the right eye would perform better than either the left or both eyes. One possible cause is that the distribution of acuity is more spread out in the left eye than the right eye. Additionally, if order is the cause, then that is a difficult problem to fix, because some test will need to come first and some test will need to go last.

Test Type	Accuracy	FNR	FPR	Pos. No.*
Both Eyes (any distance)	0.82	0.16	0.18	25
Left Eye (any distance)	0.88	0.07	0.16	30
Right Eye (any distance)	0.81	0.23	0.16	31
Distant (any eye)	0.91	0.17	0.04	46
Near (any eye)	0.77	0.12	0.29	40
All distances, eyes	0.84 ± 0.09	0.15 ± 0.08	0.17 ± 0.06	86

Table 32: Accuracy, False Positive Rate, and False Negative Rate for Three-of-Five Policy
*Number of eyes with a given vision problem

Unfortunately, the five tests policy does not have the desired attribute of having a very low false negative rate. One way that might work to decrease that false positive rate is to incorporate more tests, to give a potential eye time to reveal its vision problems. Unfortunately, the false negative rate remains relatively high for all testing policies, although the four-correct policy has a slightly lower false negative rate than the five tests policy, the difference is small on the scale of the initial error estimation performed. Additionally, calculating the 95% confidence intervals for all eyes at all distances shows that the four policies are not significantly different from one another. However, those confidence intervals should be thought of as the most generous (i.e. least conservative) confidence intervals, because they treat each eye observation as independent. However, even using the generous intervals, none of the four policies are distinguishable from the best policy.

Test Type	**Four-Correct**	**Two-Correct**	**Five-Tests**	**Ten-Tests**
Both Eyes (any distance)	0.16	0.24	0.16	0.20
Left Eye (any distance)	0.19	0.26	0.23	0.23
Right Eye (any distance)	0.03	0.10	0.07	0.07
Distant (any eye)	0.17	0.22	0.17	0.17
Near (any eye)	0.07	0.18	0.12	0.15
All distances, eyes	0.13 ± 0.06	0.20 ± 0.09	0.15 ± 0.08	0.16 ± 0.08

Table 33: False Negative Rates by Testing Policy

To understand why these errors occur, consider Table 34, which shows the mean and standard deviation Snellen acuity of each result type. The mean acuity of a True Positive is much higher than that of a False Negative. Similarly, the mean acuity of a True Negative is lower than that of

a False Positive. Unfortunately, many of these differences are not statistically significant at the 95% confidence level. On the one hand, this means that insufficient data was gathered to determine if these distributions really are different from one another. On the other hand, the raw difference in means gives reason to believe that better performance can be achieved by some improved version of QuickCheck.

Result Type	Acuity ± SD	Count
False Negative	50.7 ± 16.6	13
False Positive	26.4 ± 4.5	26
True Negative	20.2 ± 4.7	128
True Positive	128.0 ± 130.8	73

Table 34: Mean, Standard Deviation, and Number of Outcomes by Type using Five-Tests Policy

Therefore, the next step is to consider what factors QuickCheck can take into account in the future to achieve better results. However, examining QuickCheck's mistakes more deeply, it becomes clear that there is not an (obvious) pattern between acuity and mistake type, besides the fact that false negatives and true positives can only occur at Snellen acuities worse than 20/32 (and vice-versa for false positives and true negatives at acuities better than 20/32). Figure 32 shows that, while on average false positives and false negatives are closer to the dividing line (acuity of 20/32) than their true counterparts, there is no easy way to distinguish by true acuity alone which eyes QuickCheck can accurately identify as having or not having vision problems. Figure 34 shows a similar lack of identification of cause, and while there is some difference in false negative rate between eyes (with fewer false negatives for right eyes than other eye types), the cause of that difference is unclear from these figures alone.

However, Figure 33 does give a little insight into how false negatives occur less often for near as opposed to distance vision. The spread of Snellen acuity is smaller, and is more centrally grouped around the 20/32 threshold, and this difference is statistcally significant ($p<0.05$ using a t-test). Unfortunately, this is most likely due to the measurement tool used, as the near vision cards used had a much more limited range of acuity values they could test for compared to the FrACT system used for distance vision. It is therefore difficult to say whether this difference in false negative rate is caused by a difference in performance on QuickCheck's part, or if it is due to the different tools used to measure the Snellen acuity of an individual.

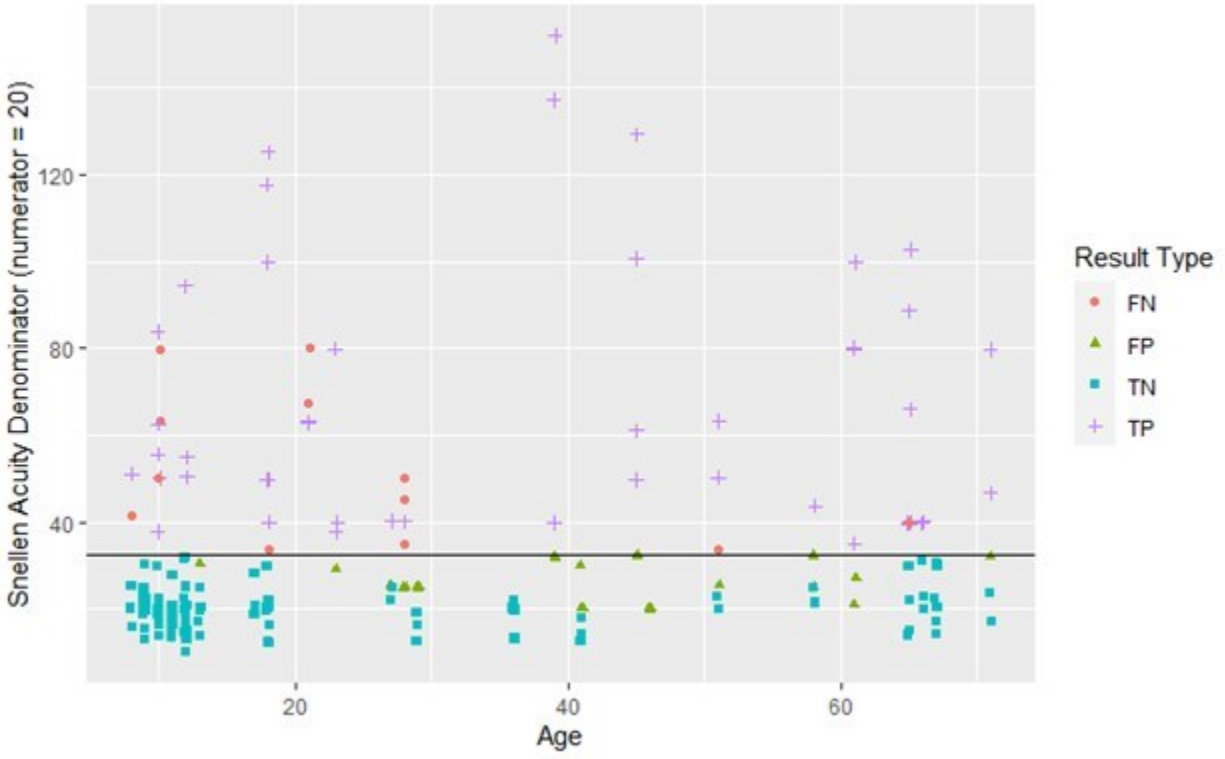

Figure 32: Snellen Acuity against Age by Result Type. Eyes with Acuity > 20/175 removed for clarity.

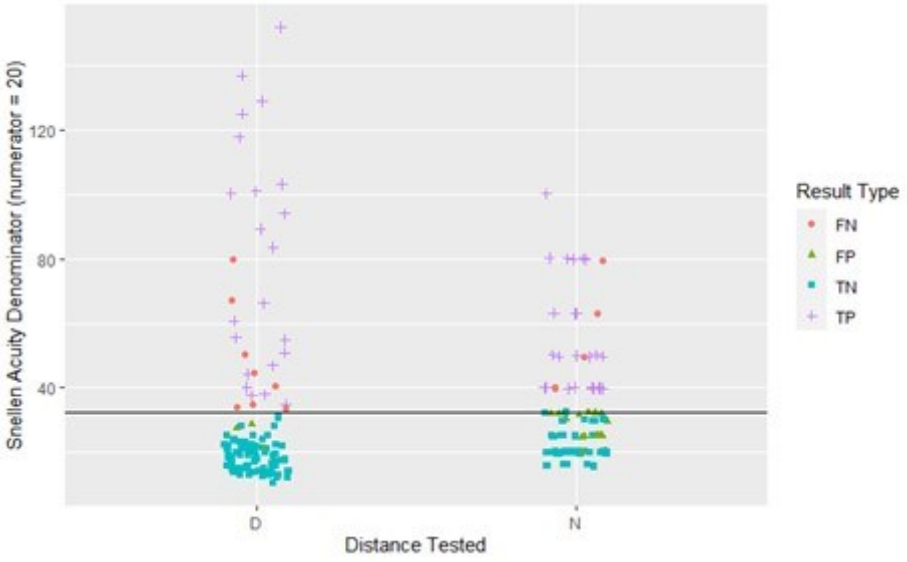

Figure 33: Snellen Acuity against Test Distance by Result Type. Eyes with Acuity > 20/175 removed..

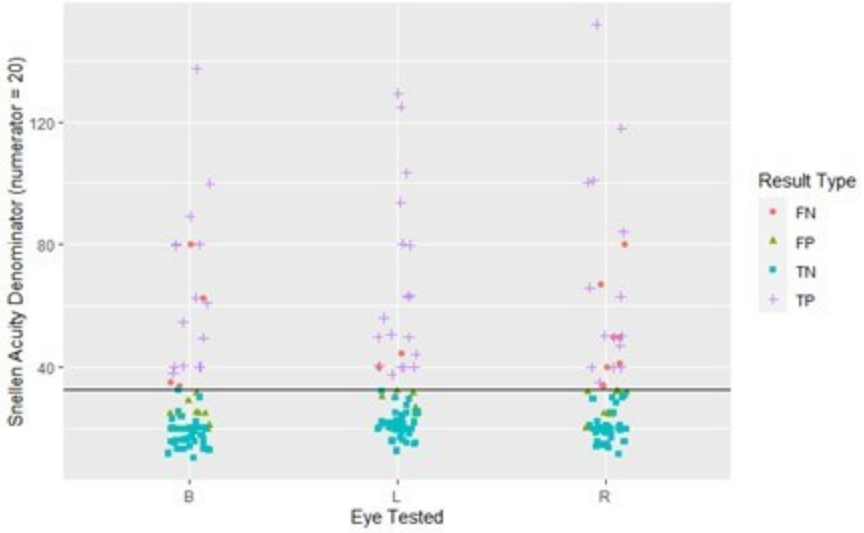

Figure 34: Snellen Acuity Denominator against Test Eye by Result Type. Eyes with Acuity > 20/175 removed for clarity.

Reversing the comparison, an alternate way of looking for patterns in errors is by considering the proportion of QuickCheck optotypes a given eye reads correctly in the context of their acuity. Figure 35 shows the proportion of QucikCheck optotypes read correctly by an ‘eye’ on the y-axis against the actual Snellen acuity denominator on the x-axis. The vertical line represents the

cutoff point (at 20/32) at which an eye's vision is considered to be poor enough to reffer to a vision exam under Washington state law, and the color and shape of a data point reresent's QuickCheck's determination of that eye's status using the three-of-five policy. False negatives are represented by red circles to the left of the virtical line at 20/32, while false positives are repersented by blue triangles to the right of the vertical line.

As one woud expect, as the proportion of optotypes read correctly decreases, more and more of the cases are determined to have vision problems. There are a few unexpected cases, where a subject with a low proportion read correctly is determined to have no vision problems (or vice-versa), but these cases occur when an individual has a particularly 'lucky' (or unlucky, as the case may be) run at the beginning of a given test. The correlation (Pearson correlation $r = -0.39$) between the proportion of optotypes read correctly and the acuity of the eye is not very strong, although it is in the expected direction (i.e., fewer optotypes read correctly indicates worse acuity).

As shown in Figure 35, most errors occur between 20% and 70% of optotypes read correctly; this is to be expected, as QuickCheck is more likely to make mistakes for subjects which read a roughly equal proportion of optotypes correctly and incorrectly. However, there are five false negatives made at above 80% optotypes read correctly. This indicates that finding a better testing policy is unlikely to correct all of QuickCheck's false negatives, because the a high proportion of optotypes are being read correctly by the subjects who then go on to have worse than 20/32 vision.

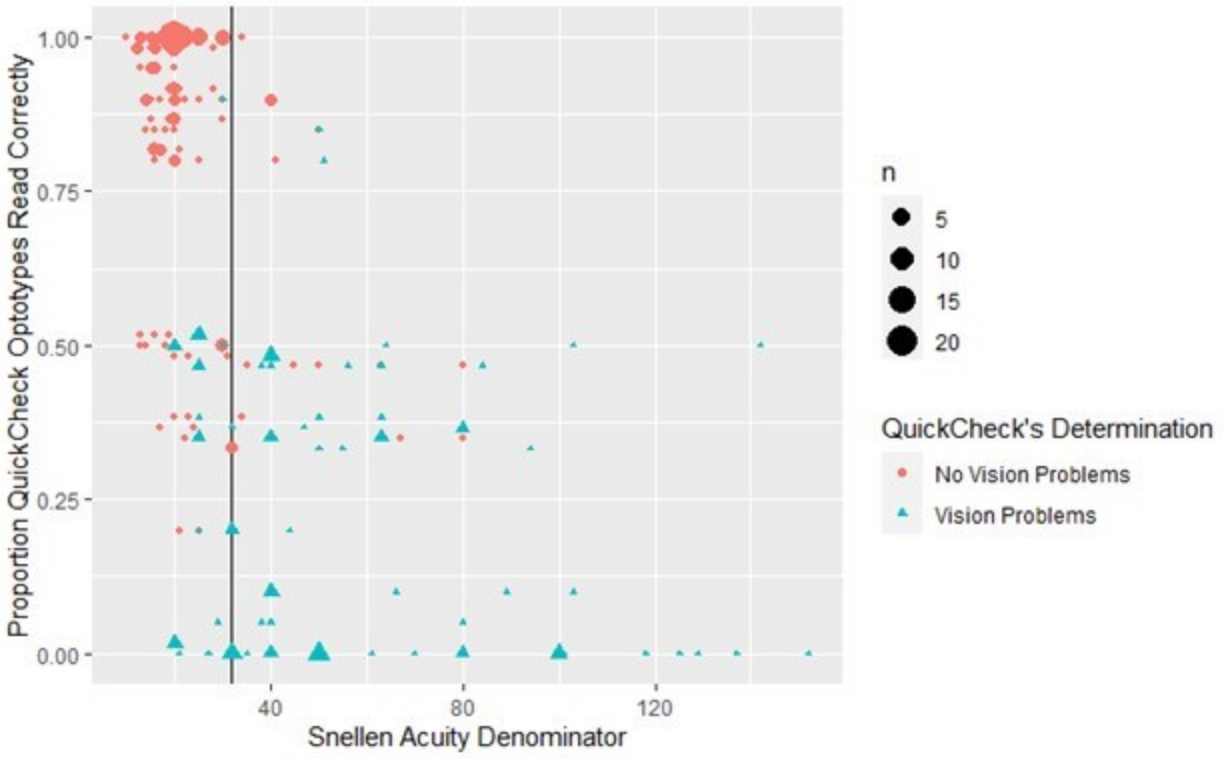

Figure 35: Snellen Acuity Denominator against Proportion of QC Optotypes read Correctly.

4.3.3 Thresholding Considerations

Considering both Figure 32 and Figure 35, it is clear that there isn't some easy way to extract all of the eyes with good vision from those who do not based only on the proportion of QuickCheck optotypes read. An additional piece of evidence in favor of this view is that there is no good alternate threshold of proportion of optotypes read correctly where QuickCheck performs significantly more accurately than it does at the current threshold of 50%. This is shown in the R-Operator curve (ROC) displayed in Figure 36, where the proportion threshold of 0.5 has roughly the same performance in terms of false positive proportion at 0.6 as it does at 0.5 and 0.4. Note that it is possible in theory to overcome this limitation by using a more complicated testing policy than just using the proportion of optotypes read correctly, but neither the four-correct nor

the two-correct policy perform substantially differently than either of the proportion-based policies, so if such a policy exists it was not found in this thesis.

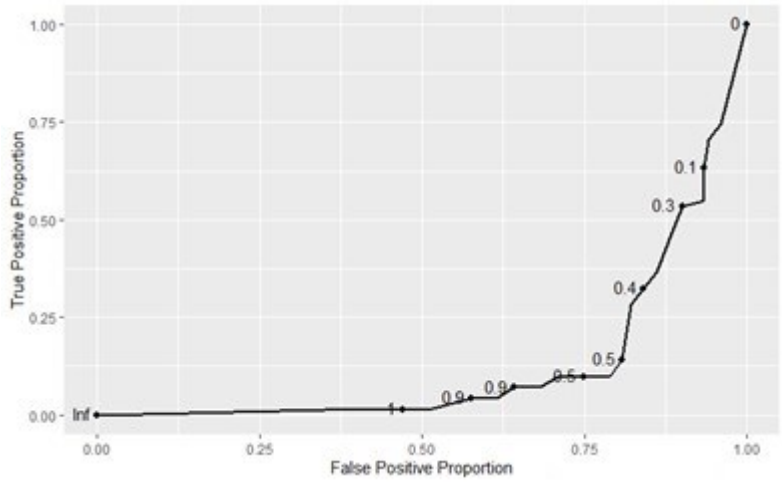

Figure 36: Receiver-Operator Curve for Proportion of Optotypes Read Correctly

Additionally, changing the threshold at which vision is considered "healthy" does not improve QuickCheck's results. As shown in Table 35, the accuracy, sensitivity, and specificity of QuickCheck does not significantly improve when the threshold (the Snellen acuity deviding "healthy" from "must reffer" vision) is changed. In particular, specificity (or the true negative rate) increases when the threshold is decreased to 20/35 or 20/40, while sensitivity decreases, indicating a rise in false negative rate. Additionally, while decreasing the threshold to 20/35 slightly lowers accuracy specifity, it does improve senstivity; this indicates a decrease in false negative rate that does not fully offset an increase in false positive rate. While changing the actual threshold is impossible, as it is set by state law, changing the size of the optotypes displayed by QuickCheck would have a similar effect. In other words, by decreasing the size of the optotypes is likely to decrease false negative rates, although it will also likely reduce

accuracy by increasing false positive rates. Due to a low sample size, none of the changes presented here are statistically significant at the 95% confidence level. This means that not only is there no clear way to improve overall performance by changes in optotype size, it is possible that all differences in performance are due to random chance.

Snellen Threshold	Accuracy	Sensitivity	Specificity
20/25	0.82 ± 0.05	0.72 ± 0.08	0.95 ± 0.07
20/32	0.86 ± 0.04	0.84 ± 0.07	0.87 ± 0.05
20/35	0.85 ± 0.05	0.87 ± 0.07	0.83 ± 0.06
20/40	0.84 ± 0.05	0.88 ± 0.07	0.82 ± 0.06

Table 35: Performance with 95% C.I.s of the Five-Tests Policy Using Various Acuity Referral Thresholds

However, as different distances have different false negative rates, each distance type's new threshold should also be examined. Table 36 shows the performance of the five-tests policy for near vision, and Table 37 shows the performance of the same policy for distance vision only, both across multiple referral acuity thresholds. Under this view, it is clear that, while there is a clear tradeoff between specificity and sensitivity for all distances, for distant vision lowering the acuity threshold produces a decrease in false negatives that is not offset fully by a corresponding increase in false positives (at least, up until about 20/35). However, this effect does not appear when only near vision is considered.

This is an indication that the distance vision optotype sizes are currently slightly too large, as they are most accurate at predicting a vision threshold of 20/35. However, given the large confidence intervals even for distance vision, it is possible that the differences in performance across threshold are due to random chance. However, there is a reasonable mechanism for this

decrease in false negatives to occur when shrinking optotypes, as decreasing optotype size would make the test more conservative. Therefore, it is likely that slightly shrinking the distance optotypes (but not the near vision optotypes) could improve QuickCheck's overall performance and decrease the false negative rate in the process.

However, decreasing the optotype size may be impossible due to state regulations stating that screenings must test for 20/32 vision at a distance. There are a few pieces of information to note which may alleviate concerns about violating that regulation. First, this decrease in size would need to be small; moving from a 20/32 to a 20/35 optotype reduces the distance optotype's size to roughly 96% of its original size. Additionally, the imprecision in the calibration process means that the difference between a 20/32 and a 20/35 optotype (especially at a distance) is small enough to be within the calibration error of QuickCheck. In other words, it is possible that QuickCheck is using distance optotypes that are too large despite the work put into correcting the calibration process (see section 3.1.6) for 20/32 vision at 10 feet due to a lack of precision in the calibration process. In that case, reducing the optotype size would not just be important for reducing false negatives, but needed to follow regulations.

Snellen Threshold	Accuracy	Sensitivity	Specificity
20/25	0.78 ± 0.07	0.71 ± 0.10	0.89 ± 0.09
20/32	0.81 ± 0.07	0.86 ± 0.10	0.77 ± 0.10
20/35	0.77 ± 0.08	0.88 ± 0.10	0.71 ± 0.10
20/40	0.77 ± 0.08	0.88 ± 0.10	0.71 ± 0.10

Table 36: Performance of the Five-Tests Policy for Near Vision Using Various Acuity Referral Thresholds

Snellen Threshold	Accuracy	Sensitivity	Specificity
20/25	0.88 ± 0.06	0.74 ± 0.12	0.98 ± 0.03
20/32	0.91 ± 0.05	0.83 ± 0.11	0.96 ± 0.05
20/35	0.92 ± 0.05	0.86 ± 0.10	0.96 ± 0.04
20/40	0.91 ± 0.05	0.88 ± 0.10	0.92 ± 0.06

Table 37: Performance of the Five-Tests Policy for Distance Vision Using Various Acuity Referral Thresholds

4.3.4 Non-Referring False Negatives

The worst type of error (non-referring false negatives, or NRFNs) occur when none of a subject's eye problems are detected. In fact, there are two types of non-NRFN results where one of a subjects vision problems is missed: a type where at least one other vision problem is detected correctly, and the other when all vision problems are missed but QuickCheck determines that a subject has a vision problem that they do not actually have. In other words, the second type of non-NRFN result occurs when there is a false positive for some eye at some distance, and some number of false negatives for other eyes and distances. If the second type of non-NRFN result is included, then every subject that has a vision problem of some sort is detected by QuickCheck as having a vision problem. Of course, as discussed early in this section, that does not mean that every vision problem is detected correctly. If the second type of non-NRFN results are excluded, then acurracy is still high, at 0.95, with 1.0 sensitivity and 0.89 specificity. This shows that although QuickCheck does not catch every eye problem that an individual has, it is still good at detecting that a subject has an eye problem if they have any.

Continuing this examination, Table 38 shows the estimated ability of QuickCheck using the five-tests policy to detect a subject's vision problems by the vision problem they have. Note that this is different than what is shown in Table 33, because that table displays performance metrics with

‘eyes’ as the unit, not the subject. It is also different than what is displaed in the by-policy performance metric tables (e.g., Table 17), because what is being considered is if QuickCheck can detect a given problem, not whether it can detect a problem using a specific test. The most important metric in Table 38 is the sensitivity by subject condition, which is to say, how likely it is that a given subject with a specific vision problem will be identified by QuickCheck.

Table 38 initially would have displayed sensitivity for subjects with only distance acuity deficits, only near acuity deficits, or any distance deficits, as those would have been a more precise context to consider the sensitivity. However, the number of subjects with specific conditions (e.g., near vision problems in a single eye but no other conditions) was too small for those metrics to be useful. For example, only a single subject had a near vision deficit in one eye and no other acuity problems. Therefore, Table 38 shows performance by non-exclusive acuity deficit category. For example, the “Right Eye Acuity Deficit” row’s specificity answers the question “how likely was QuickCheck to detect any problem in a subject’s right eye given that they have a problem in their right eye?”

As Table 38 shows, performance looked at through this lense is much improved when compared to the by-eye comparisons. Note that the high specificity values are somewhat meaningless, because a “false positive” has an unclear definition in this context. Confidence intervals are provided for accuracy and specificity at the 95% level for any acuity deficit. This improvement in performance is caused by the co-occurrence of multiple vision problems within a single subject. For example, if a subject has hyperopia in both eyes, then QuickCheck has not one but three chances to detect that problem (once each for both eyes, the left eye, and the right eye). This increased performance does not occur for subjects that have vision problems in a single eye

at a single range (when they can use the other eye to compensate when both eyes are tested). However, such subjects are relatively rare; for example, 23 subjects have vision problems overall, and 21 have a vision problem only in their left eye. This means that most people who have vision problems in one area also have vision problems in another. As expected from that perspective, Table 38 shows that QuickCheck is better at detecting problems at the condition level than the eye level. Because detecting even a single vision problem is enough to generate a referal to an eye exam, the final row (considering performance for any vision problem) is a good representation of QuickCheck's ability to determine if a subject should be reffered to an eye exam at all.

Subject Problem Type	Accuracy	Sensitivity	Specificity	Num. Pos.*
Right Eye Acuity Deficit	0.88	0.77	1.00	22
Left Eye Acuity Deficit	0.98	0.95	1.00	21
Near Vision Acuity Deficit	0.98	0.94	1.00	18
Distance Vision Acuity Deficit	0.92	0.85	1.00	20
Any Acuity Deficit	0.95 ± 0.07	0.91 ± 0.11	1.00	23

Table 38: Performance by Vision Problem Type Using Five-Tests Policy
*Number of subjects with the given problem

However, it is unclear from Table 38 alone how precise the measurement of sensitivity is. Therefore, Table 39 shows the 95% confidence interval for sensitivity for each problem type. The intervals are relatively wide, given the low sample sizes, although they on the narrower side of initial projections which estimated errors of between 0.33 and 0.10 (see Section 4.2.1). This is due to the fact that a higher proportion of the sample had vision problems compared to the initial estimates based on overall population data. Overall, the right eye and distance vision stand out as

areas where false negative rate is high (sensitivity is low). The right eye situation is somewhat puzzling, but as explained earlier in Section 4.1.2, this thesis' subjects had better vision in their right eyes than on their left on average, so that likely contributed to this discrepancy in error rates between eyes. Regarding distance vision, the average subject was more likely to have worse near acuity than distance acuity. Additionally, as discussed earlier in Section 4.3.2, there is some indication that the distance optotypes are slightly too large for the current acuity threshold used to determine if a subject has a vision deficit.

Problem Type	Lower Bound	Mean Sensitivity	Upper Bound	Num. Pos.*
Right Eye	0.60	0.77	0.95	22
Left Eye	0.86	0.95	1.00	21
Near Vision	0.84	0.94	1.00	18
Distance Vision	0.85	0.85	1.00	20
Any Deficit	0.80	0.91	1.00	23

Table 39: 95% Confidence Interval for Sensitivity using Five-Tests Policy by Problem Type

4.4 Survey Results

Two surveys were presented to the users of the QuickCheck application after the vision tests were completed. First was the CISS. The CISS (Convergence Insufficiency Symptoms Survey) is a tool used to test for a condition where the two eyes of an individual have difficulty converging on the same object in the visual field (most commonly when that object is close to the face). Untreaded, convergence insufficiency can progress into strabismus and amblyopia, which can prevent binocular fusion. The second survey presented was a single question survey asking when

the subject had last been to an eye doctor. Section 4.2.1 discusses the results of each question on the CISS, and section 4.2.2 discusses the results of the second (single question) survey.

4.4.1 CISS Results

Each Convergence Insufficiency Symptom Survey (CISS) response is given a score from 0 to 5, with 0 being “Always” and 5 being “Never”. A subject is considered to have a positive result (high likelihood of convergence insufficiency) if the subject receives a total score of 16 or higher (where the total score is the sum of all of the scores of each answer given).

Importantly, the CISS survey QuickCheck uses is the same as the survey used in traditional vision screenings. It is possible that the results of QuickCheck’s CISS were systematically different from those performed by an optometrist or ophthalmologist (especially if subjects had questions about the meaning of certain questions), but overall, those differences are more likely to be caused by the tester than the application. Therefore, there is no strong ground truth to compare the CISS results to, so this section outlines the results of the CISS survey but does not attempt to estimate the accuracy of QuickCheck’s CISS.

The mean and standard deviation of each question in the CISS is given in Table 40. Interestingly, most questions have a mean response between 1 and 2, with only questions 7, 8, 10, and 12 having a mean response between 0 (“Never”) and 1 (“Infrequently”). These questions all deal with specific symptoms of convergence insufficiency, including asking about double vision (question 7) or asking about feeling a pulling sensation around the eyes when reading (question 12). Therefore, it makes sense that these questions would have a lower average score than more general questions, which ask things like “do you feel sleepy when reading or doing close work” (question 4).

Question No.	Mean Score	Score Std. Dev.
1	1.7	1.1
2	1	1.3
3	1	1.1
4	1.6	1.4
5	1.4	1.3
6	1.6	1.4
7	0.8	1.1
8	0.5	1.1
9	1.4	1.7
10	0.8	1.1
11	1	1.1
12	0.6	1.2
13	1.1	1.5
14	1.4	1.3
15	1.6	1.2
Total Score	17.4	11.5

Table 40: CISS Questions Response Means and Standard Deviations

The CISS mean score was 17.4, with an average per-question score of 1.2 (between "Infrequently" and "Sometimes"). Of all subjects, 18 had a score greater than 15, indicating symptoms of convergence insufficiency, or 45% of all subjects. The convergence insufficiency rate was higher in adults than children (44% for children and 48% for adults) but the difference

between age groups was not statistically significant ($p>0.05$ using a t-test). On the other hand, considering raw CISS scores, as shown in Figure 37, the trend seems to reverse, in that younger people tend to have higher CISS score totals. This discrepancy seems to be due to a lack of natural closeting in the age and CISS data, such that natural groups of people with no symptoms and people with lots of symptoms are not present (or at least, not apparent with this sample).

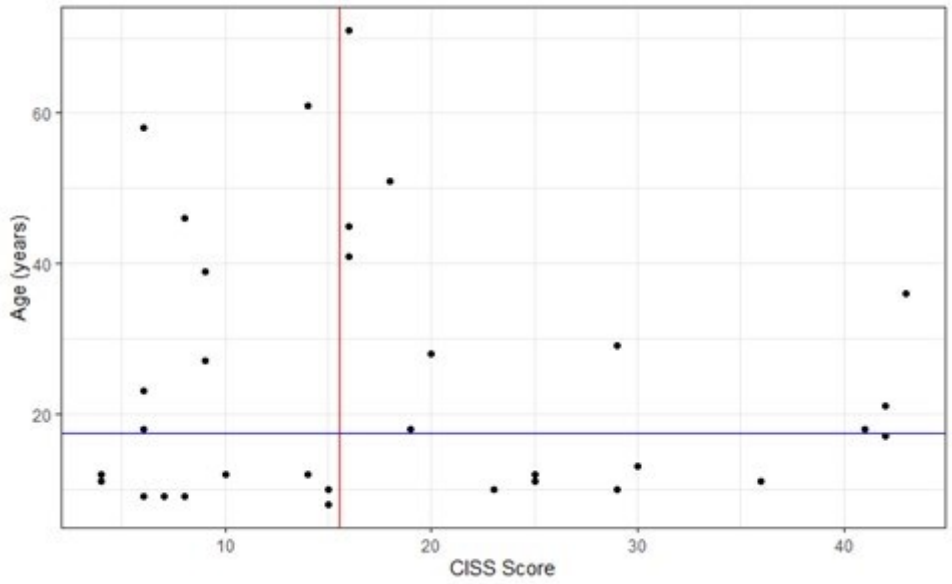

Figure 37: CISS score and Subject Age.

Points right of the vertical line have symptoms of CI. Points below the horizontal line are children.

The symptom survey reporting a prevalence rate of 52% is unexpected, as convergence insufficiency has an estimated prevalence between 4.2% to 17.6% in children (and slightly less that for adults) [64]. In fact, the difference between the estimated prevalence in children and the sample prevalence is statistically significant ($p < 0.006$ using a t-test). There are several possible reasons for this difference. First, the CISS is a screening tool, and it is expected to have a higher false positive rate than an exam. Additionally, many of the participants in the study were at a vision clinic specifically because of convergence insufficiency or similar problems. Therefore,

the fact that the CISS flagged a higher proportion of the study subjects than might be initially expected is not surprising.

Qualitatively, it was difficult for subjects to determine the answers to some of the CISS questions, especially if those subjects were children. Many subjects asked clarifying questions (such as "what is close work?"), and although most of these questions could be answered in the moment, there were some that the researcher was uncertain about, which likely degraded the quality of the results.

4.4.1.1 The CISS and QuickCheck Performance

One area of interest for this thesis is the relationship between the CISS results and a subject's visual acuity. While the CISS and visual acuity are not necessarily related physically or optically, it is conceivable that convergence insufficiency or a similar vision problem could impact a subject's ability to make accurate use of QuickCheck's near vision acuity test. Figure 38 shows the correlation between the percentage of QuickCheck optotypes read during each sub-test and the CISS score (N = Near vision, D = Distance vision, B = Both eyes, R = Right eye only, L = Left eye only). Unexpectedly, examining the correlation between a subject's CISS score and their QuickCheck results shows a higher correlation between distance vision test results and CISS scores. As all correlations are positive, increasing the CISS score is associated with an increase in visual acuity score (decrease in visual ability). However, no correlation between the CISS and any test percent has a magnitude higher than 0.45, so the CISS does not seem to be correlated with percent of optotypes read correctly.

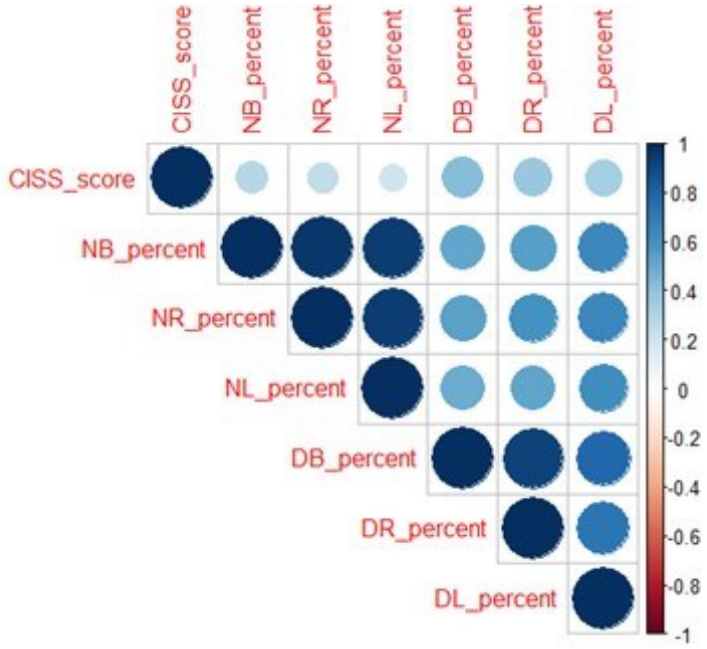

Figure 38: Pearson Correlation Heatmap between Percent of QuickCheck Optotypes Read Correctly per Distance-Eye Pair and CISS score.

Unexpectedly, the CISS is slightly negatively correlated with a subject's visual acuity (with an increase in CISS indicating an increase in visual acuity). This is likely to be a feature of the unique study population, because many of the subjects tested on at the EYE SEE clinic had good vision but had some eye problems with symptoms similar to convergence insufficiency, whether that be convergence insufficiency itself or a condition like strabismus.

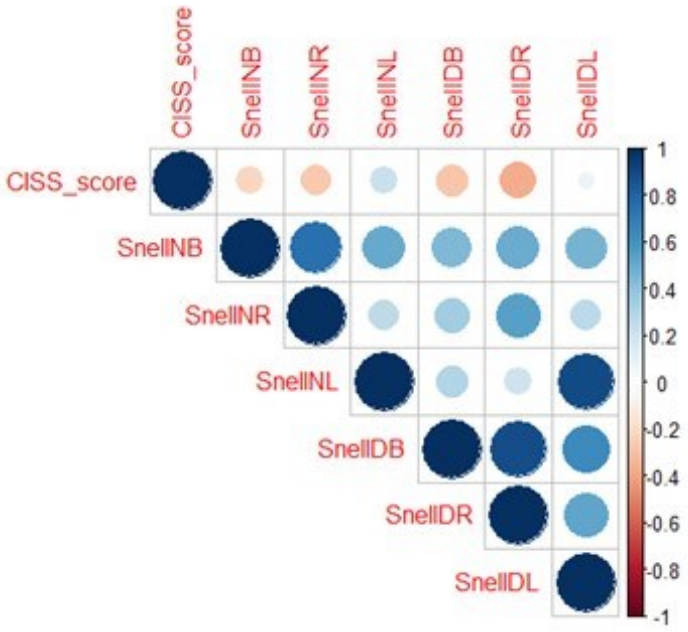

Figure 39: CISS Score correlations With Snellen Fraction Denominator (Numerator fixed at 20)

Looking at a plot of CISS score against the proportion of optotypes read correctly (Figure 40), it is clear that while there is some correlation between an increased CISS score and optotypes read, it is not a particularly strong relationship. Additionally, examining Figure 41 shows that there is little relationship between whether QuickCheck made an error and CISS score (Pearson correlation 0.01). This highlights the lack of association regarding CISS score and QuickCheck performance.

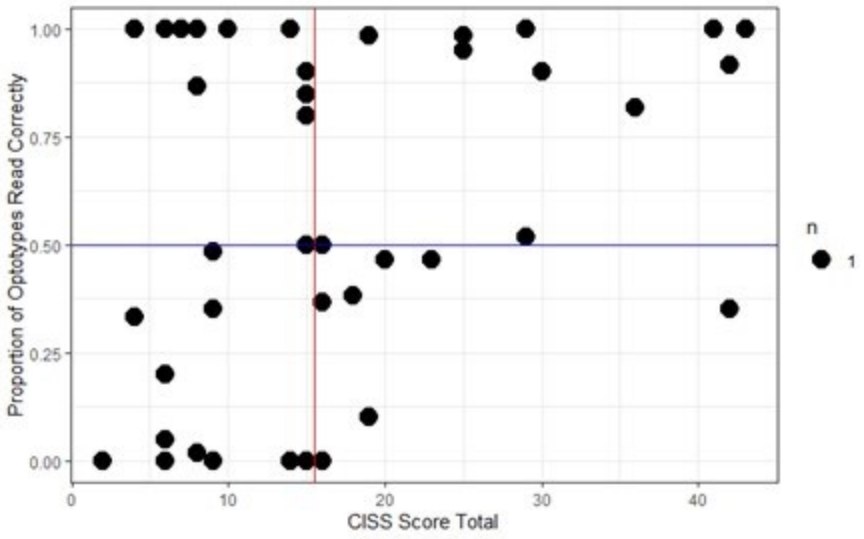

Figure 40: CISS Score against Proportion of QuickCheck Optotypes Read Correctly

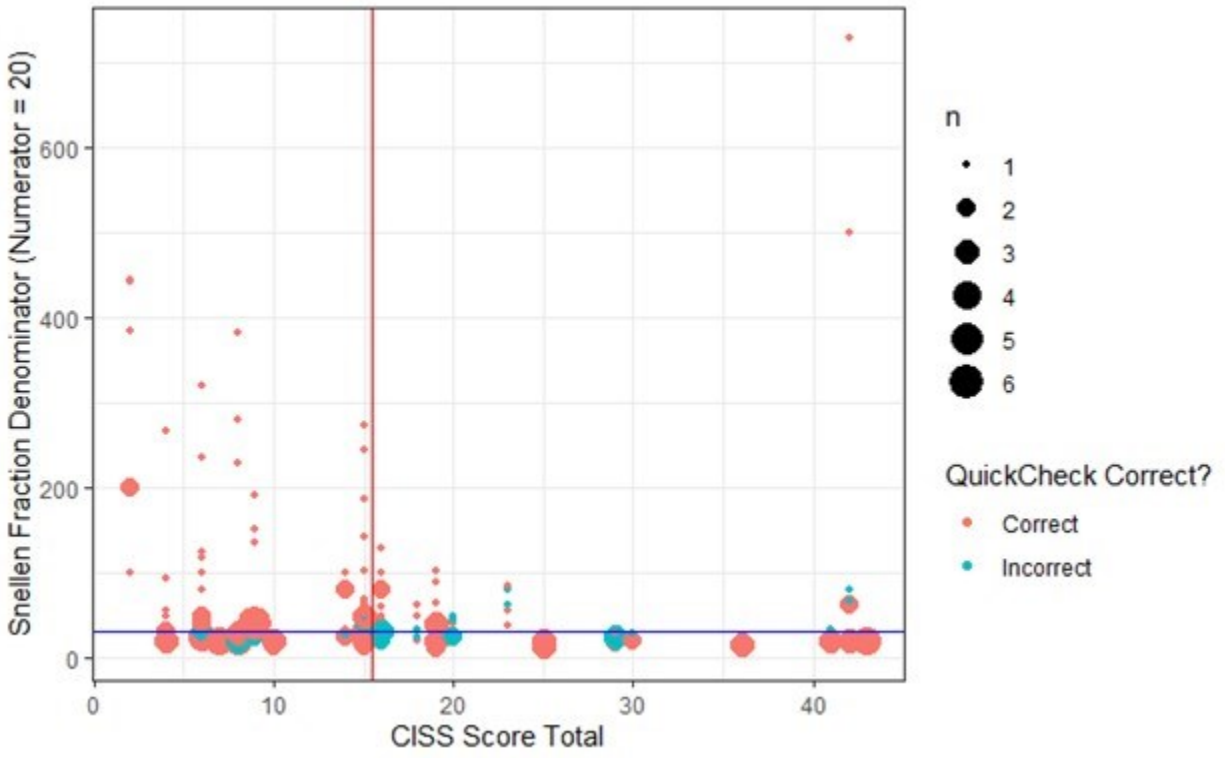

Figure 41: CISS Score and Acuity by QuickCheck's Correctness

4.4.1.2 Shortening the CISS

One of the key interests in QuickCheck's future use of the CISS is in reducing the amount of time taken to complete the survey. Several of the CISS's questions are very similar to each other (e.g., "Do your eyes feel sore when reading or doing close work?" and "Do your eyes feel uncomfortable when reading or doing close work?"), so it is reasonable that the survey could be shortened without significantly reducing the quality of its ability to detect convergence insufficiency. To determine if this is possible, several comparisons between the original CISS and shortened versions of the survey were performed.

One strategy for shortening the CISS without significantly reducing its diagnostic ability is to filter out questions which are not correlated strongly with the overall CISS score. First, a correlation matrix (see Figure 42) was made between each individual CISS question and the total CISS score, which is calculated by summing the score for each individual CISS question (see Appendix B for all CISS questions). All questions, as expected, showed a positive correlation with the overall score. This makes sense, as no answer to a question could negatively impact the CISS score.

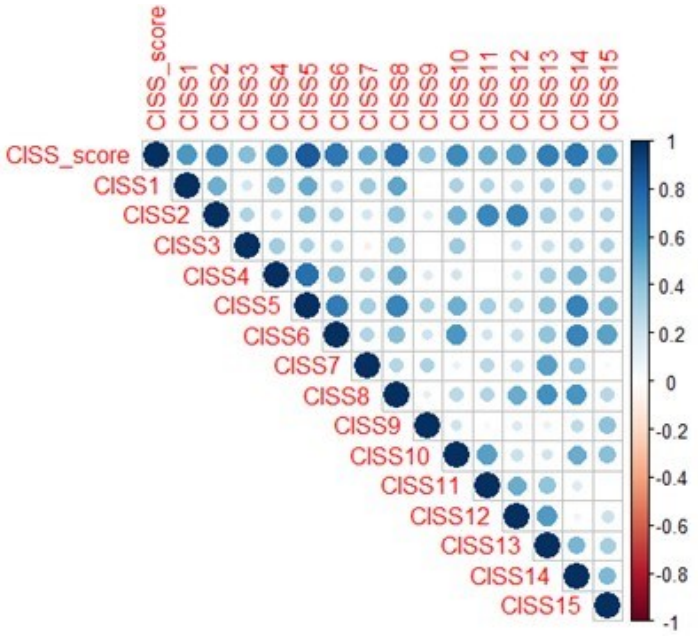

Figure 42: Pearson Correlations Between CISS questions and CISS score

Then, the new CISS score can be recalculated by summing all the remaining questions, and the CISS outcome can be recalculated using the standard that a respondent has symptoms of convergence insufficiency if their total score is greater than the number of questions in the survey. For example, questions with a Pearson correlation less than 0.5 removes questions 3, 9, and 11, and the new CISS will report symptoms of convergence insufficiency if the total score is greater than 12. The results of using the same filtering approach with increasing Pearson correlation thresholds are shown in Table 41. Removing three or four questions results in an accuracy of 97%. Reducing the survey by roughly half its length (threshold of r = 0.6) results in an accuracy of 88%, although an even further reduction maintains that accuracy while keeping only 5 (one-third) of the original questions.

Correlation Threshold	Questions Removed	Total Questions Remaining	Accuracy
0.5	3, 9, 11	12	0.97
0.56	3, 9, 11, 12	11	0.97
0.57	1, 3, 9, 11, 12	10	0.91
0.6	1, 3, 7, 11, 12, 15	8	0.88
0.7	1, 2, 3, 4, 7, 9, 10, 11, 12, 15	5	0.88

Table 41: Shortened CISS Versions and their Accuracies

While the sample size of this thesis is too small to definitively report that removing any given set of questions could allow the CISS to be performed more quickly without reducing screening utility, this preliminary work shows that it is very likely that the CISS contains some redundant questions that can be removed without impacting its ability to detect convergence insufficiency.

4.4.1.3 Qualitative Results of the CISS

In testing, the key points of difficulty surrounding the CISS were its length and the difficulty interpreting the answers received from subjects as one of the possible CISS answers. Additionally, for younger children (younger than 10), there was sometimes a challenge explaining what each question meant. The speed of the CISS was addressed in subsection 4.4.1.2; however, the difficulty interpreting answers and asking questions to young children are harder problems to address. Some success was found in informing subjects of the type of answers the CISS accepted before the survey began. However, especially in young children, each

question required a bit of back-and-forth discussing what their answer meant and how to convert it into one of the five frequency-based answers accepted by the CISS on QuickCheck.

The most common problem was having difficulty differentiating between “Sometimes” and “Infrequently”, as it was often hard to gauge what a response of “sometimes” meant; was that response really an “infrequently” rather than a “sometimes” or should the response be taken at face value? Additionally, some younger subjects did not know what “infrequently” meant. Whenever a subject would answer a question with “sometimes”, “not that often”, or similar middle-of-the-road responses, further digging was necessary to determine which CISS response their answer was best suited to be. Unfortunately, this is a somewhat unavoidable problem. Rephrasing some CISS questions might decrease the need for explanations, but ultimately the nuanced responses of a subject to a CISS question will still need to be condensed down into a one-word answer, so a true solution seems impossible. Of course, patience, careful explanations, and experience working with young children will all help alleviate some of the difficulties in interpreting a subject’s CISS responses, so in that sense the best solution to this problem is training and experience.

4.4.2 Final Question

The final question of the QuickCheck application is intended to help subjects know when they have gone too long between visits to an eye doctor. It asked, “When was the last time you had an eye exam with an eye doctor?” and the possible answers were “Less than one year ago”, “More than one year ago”, “More than two years ago”, and “Never”. Table 42 shows the response rate (proportion of subjects who selected a given answer) for each of the answers to the final question. Somewhat unexpectedly, the response rate does not smoothly decrease as the time

since the last doctor's visit increases. Instead, it seems like there are two groups of people: those who go to an eye doctor regularly, and those who rarely if ever go to see an eye doctor.

Answer	Response Rate
Less than one year ago	0.62
More than one year ago	0.08
More than two years ago	0.20
Never	0.10

Table 42: Final Question Answers and Response Rates

Considering the response rate by modality, it is clear that response rates are very different by the setting the subject was encountered in. All subjects who were seen at a clinic visit or were adult volunteers had been to an eye doctor within the last year. Therefore, Table 43 shows the response rates by answer for only those subjects seen at a mobile clinic event. From this perspective, the response rates become much smoother. Most mobile clinic subjects (66%) had last seen an eye doctor at least two years ago if at all, and only 17% had seen an eye doctor within the last year. It is difficult to draw conclusions about the mobile clinic population, as the baseline response rates to this question are unknown for the population as a whole. However, this finding does show that when it comes to eye health, the mobile clinic subjects and all other subjects were drawn from very different groups.

Answer	Response Rate
Less than one year ago	0.17

Answer	Response Rate
More than one year ago	0.17
More than two years ago	0.44
Never	0.22

Table 43: Response Rates for the Final Question (Mobile Clinic Subjects Only)

4.5 Qualitative Findings

This section covers the qualitative findings of clinical testing not covered in previous sections. Subsection 4.5.1 covers key problems that occurred during testing, as well as long- and short-term ways to solve or address those problems, and Section 4.5.2 covers advice for future researchers, as well as interesting and potentially fruitful research questions to consider as well as some study design possibilities that could be used to answer those questions.

4.5.1 Problems Encountered During Testing

In addition to the quantitative results discussed in Sections 4.1 and 4.2, qualitative findings, including problems encountered during testing, were also gathered during the course of research. The most important of these are highlighted in Table 44.

No.	Problem Encountered	During-Testing Fix	Long-Term Solution
1	The subject had difficulty reading letters.	Encourage subject to make a guess or describe the shape of the letters they see.	Implement a way to use LEA symbols instead of letters.

No.	Problem Encountered	During-Testing Fix	Long-Term Solution
2	The subject had goals other than receiving accurate test results.	Discuss the importance of accurate results with subjects.	Improve training of testers to ensure that they can detect these cases and educate subjects on the importance of accurate results. Improve study exclusion rules to prevent subjects who are 'wishy-washy' from being asked to test QuickCheck in the first place.
3	One parent consented to test, but child and/or another parent did not.	Don't test that child.	Improve study onboarding process for subjects and their parents, to give them time to discuss their feelings on testing and come to a consensus.
4	Subjects "gave up" and did not try to read all optotypes.	Encourage the subject to read all the optotypes they could.	Improve training of testers to ensure that they

No.	Problem Encountered	During-Testing Fix	Long-Term Solution
			can detect these cases and educate subjects on the importance of accurate results. Include a few 'easy' larger optotypes to boost confidence.
5	Difficult to keep phone at the correct distance from a subject's face.	Use a 16in long string to maintain the correct distance from face for near vision tests. In the future, an in-app way of keeping the phone at the correct distance.	Include a distance-measuring feature in QuickCheck that can determine if a subject is too far away or too close to the phone screen.
6	Subjects were wearing glasses/contacts.	Ask subjects to remove their glasses, note where subjects were wearing contacts.	Include a way of gathering multiple data points from a single subject so that with and without glasses performance can be compared.

No.	Problem Encountered	During-Testing Fix	Long-Term Solution
7	Subjects wanted to be "right" more than they want an accurate result and would attempt to read body language and/or guess many letters to do so, or became discouraged when they could not read letters.	Confirm a guess before moving on to the next optotype, do not look at the currently presented letter until a final guess has been made. Educate subjects about the importance of accurate test results and reassure them that failure is not possible.	
8	Near vision test optotypes were difficult to read for testers, particularly at an angle.	Ensure that the tester had up-to-date glasses and were seated in a position that would allow both them and the subject to see the phone screen.	
9	Reading from the screen was difficult for younger subjects if the screen was tilted	Ensure the screen is tilted correctly based on subject feedback.	
10	Subjects reported eye strain before or during testing	If the subject reports eye strain before testing, either	Re-work study design to space out eye-straining

No.	Problem Encountered	During-Testing Fix	Long-Term Solution
		do not test or wait until strain subsides. If the subject reports eye strain during testing, pause testing until strain subsides.	activities before and during test. For example, provide a way to perform the CISS between near and distance vision.
11	Trials could take a long time, especially if the subject had vision acuity around 20/32	Inform subject of timeline of studying so they don't feel like testing will continue forever	Implement a way to skip longer parts of the study (e.g., CISS) and/or find a shorter way to measure distance vision. Include multiple researchers so one can enter data as the other tests a subject.
12	Internet access could be unpredictable and the Android phone used to run QuickCheck had no data plan.	Write down all data as a backup, use hot-spot from a phone with cellular data.	Implemented before this thesis began: store data that cannot be sent to the database locally until a connection can be made with the database through the server.

No.	Problem Encountered	During-Testing Fix	Long-Term Solution
13	Tester did not have an Android phone to test QuickCheck on.	Obtain an android phone from UWB's computing and software system department phones.	Implement an Apple/iPhone version of QuickCheck.
14	Subjects (younger children) would try to grab the phone running QuickCheck and play with it.	Guard the phone running QuickCheck to prevent interference.	Enforce log-in for each trial of QuickCheck (implemented).

Table 44: Problems Encountered During Testing

A few long-term solutions pop out as occurring multiple times in Table 44 or being important to solving a problem that occurred many times during testing. These are:

1. Implementing a more flexible way to use QuickCheck (i.e., what is tested, what scenes are included, etc.).
2. More training for researchers, particularly when it comes to interacting with children. This includes increased preparation (having good glasses, bringing a 16" long string, etc.) as well as carefully set up testing spaces (e.g., so that both tester and subject could see the phone screen clearly).
3. Implementing LEA symbols so young children can use QuickCheck more easily.

The most common problem encountered during testing (difficulty reading near-vision optotypes) had little to do with the design of QuickCheck. However, other problems, such as difficulty keeping the phone at the correct distance from the subject and long testing times, were caused by

the design of QuickCheck. In particular, problems represent areas of improvement regarding the ease-of-use non-functional requirement. Implementing the suggestions in Table 44's 'Long-Term Solution' column would therefore significantly improve QuickCheck's usability.

On the other hand, very few of the problems encountered during testing had to do with QuickCheck's reliability as an application. Even limited internet access due to a lack of cellular data was remedied by QuickCheck's ability to save data locally for later upload when a Wi-Fi connection was possible. This also meant that if the tester forgot to Most reliability problems were caused by factors beyond the application's control (including the previously mentioned lack of internet), but the save-for-later feature was able to preserve data for times when QuickCheck could connect to the server. In this way, the application's reliability was tested and verified.

4.5.2 Future Research Questions

A summary of plausibly fruitful research questions is included in Table 45. Note that Table 45 does not include research questions which were addressed previously in Chapter 4, but which require more data to answer fully, including the key question of "is QuickCheck ready to be used as a clinical tool?"

No.	Research Question	Potential Study Design
1	Do (or can) subjects learn what characters (as optotypes) are shown by QuickCheck?	Ask subjects after they test QuickCheck what types of letters they think can appear in QuickCheck.
2	If subjects learn which optotypes QuickCheck uses, does that impact performance?	Ask some volunteers to perform QuickCheck without

No.	Research Question	Potential Study Design
		telling them that the only optotypes are H, O, V, and T; compare that to a different randomly assigned groups that were told about the types of optotypes.
3	Does the order of tests matter?	Randomize test order for each subject (e.g., randomly test near vision after distance vision).
4	How well does the CISS detect convergence insufficiency?	Measure convergence insufficiency for each subject and include that in the research data.
5	Do subjects who have glasses but do not wear them during testing perform differently than subjects without glasses at all?	Ask subjects with glasses to test QuickCheck both with and without glasses and compare the results to those generated from subjects who do not have any glasses.

No.	Research Question	Potential Study Design
6	Why does QuickCheck perform differently on different eyes?	Gather enough data that those with better right eyes than left eyes and vice-versa can be compared.
7	Is the near vision reference method used distorting results?	Compare near vision cards performance to other methods of testing near vision acuity.
8	How well does QuickCheck perform on kindergarten-age children?	Gather data from subjects who are in kindergarten. This will likely require implementing the ability to use LEA symbols instead of optotypes in QuickCheck.
9	Can a more complicated test policy perform better than the simple policies tested in this thesis?	Gather enough data to, for example, train a decision tree and compare the results of that model to the current testing policies.
10	Will QuickCheck continue to perform well with a larger, more representative sample?	Work with NVI to include consent in their standard

No.	Research Question	Potential Study Design
		permissions for in-school events, so that all children NVI provide screenings for can test QuickCheck

Table 45: Future Research Question Ideas and Potential Study Designs for Each

The most important question is naturally question 8. While some analysis of statistical significance was performed in Chapter 4, limited sample sizes mean that the true extent to which current results will continue to hold with more data is unknown. Another important question is number 7, as no subject within this study's sample was of kindergarten age but kindergarteners must be screened under Washington state law. Questions 1-7 deal with verifying the extent to which currently unmeasurable factors are affecting QuickCheck's performance (and to what degree), so answering those questions is important to understanding why certain patterns of results appear in this preliminary work.

4.6 Performance Summary and Suggested Changes to QuickCheck

4.6.1 Performance Summary

QuickCheck's ability to detect vision problems is better than initially suggested by the performance of the testing policies, because it is unlikely that a subject has only a single vision problem in a single eye at a specific range. Therefore, rather than the average per-eye performance of roughly 80% (plus or minus roughly 10 percentage points), QuickCheck instead correctly finds if a subject has a vision problem 95% of the time. The sensitivity is slightly

lower, so QuickCheck will only find a vision problem if a subject has one in roughly 91% of cases (again, plus or minus about 10 percentage points).

However, there are several key problems with QuickCheck that should be addressed in future development. Primary among these is the high false negative rate for distance vision tests and right eye tests. There is some evidence suggesting that reducing the optotype size of distance vision tests would bring the false negative rate more in line with the near vision results, although further study is needed to gather enough subjects to determine if that would be an effective fix. However, it is currently unclear why right eye tests have high false negative rates as compared to other testing categories. Other problems include ease-of-use challenges, such as difficulty holding the phone steadily 16in away from the subject's face, and long testing times. Suggestions for how to fix those issues are included in subsection 4.6.2 below.

4.6.2 Suggested Changes to QuickCheck

Given the difficulty QuickCheck had with near vision acuity measurements, and with a high false positive rate, there are a few changes that should be implemented to improve performance. Additionally, there are some situations that arose during real world testing which may be avoided or ameliorated by changing the QuickCheck algorithm. This section serves as a summary of solutions to the problems presented in Section 4.5.1.

4.6.2.1 Suggested Changes to Acuity Testing and Calibration

There are two categories of planned changes to acuity testing scenes and the calibration process. The first category of changes is those alterations intended to improve performance and useability. The second category is comprised of changes which will improve the scope of use of the application.

To improve performance, particularly in the near vision acuity tests, the following changes are under consideration.

1. Implement a larger optotype size for the first letter shown to the user in each test. This is intended to build the user's confidence so that if their vision is close to the healthy/unhealthy border (around 20/32 at 16 inches), they will attempt to read the remaining letters rather than give up if the task is somewhat difficult. Additionally, it may be possible to use this additional optotype to provide a more detailed analysis of the user's vision (e.g., if the user cannot read even the larger example letter, then they should get a vision test as soon as possible).
2. Implement a second calibration method in the calibration scene; for example, providing an image of an inch or 5cm or similar measurements. This is likely to be more accurate than using a quarter, because quarters come in various sizes. In particular, older quarters tend to have worn down edges and be smaller than newly minted quarters.
3. Develop a way to randomize the order in which tests occur, so that Distance vision could be tested before near vision and both eyes after left eyes at random. This would allow further research to control for test order when considering performance.
4. Develop an in-application method for ensuring that a subject is the correct distance from the phone. This is more important for near vision tests, where even small absolute changes in distance can represent large relative changes in how far away the phone screen is.
5. Add an option to use Lea symbols for young children who may not be able to easily read letters. Additionally, include an additional screen presenting each symbol so the tester can describe what each one is named to improve testing accuracy when symbols are used.

To expand the use cases of the application, the following changes are suggested:

1. Implement a way to test very young children or others who do not read English. For example, using LEA symbols (discussed in Section 2.3.3.1) would allow the range of possible subjects to be expanded. This may require an additional presentation scene which describes each LEA symbol and tells the subject how to indicate which symbol they see.
2. Implement ways to customize at test-time what scenes are shown to a subject, and how long each acuity test is.
3. Implement additional tests for other kinds of vision problems; for example, digital versions of color vision tests are straightforward to implement and would expand usability.

4.6.2.2 Changes to CISS Survey

To expand the usability of the CISS survey, providing an optional explanation regarding the meaning of each question may improve survey accuracy. Additionally, providing a pre-CISS survey information screen regarding how to ask the CISS questions and the types of responses users should provide (e.g., how often is "sometimes" compared to "infrequently") could improve accuracy and decrease user error. Additionally, to decrease testing time, it may be useful to remove the CISS survey from the clinical testing version, especially if its results continue to be largely unverifiable.

4.6.2.3 Other QuickCheck Development Suggestions

Currently, the QuickCheck application does not have a good way of managing data gathered through clinical trials or by school nurses in the future. The QuickCheck web portal can display information about subjects, but there is currently no usable authentication system to allow

different users to log in and see the details of only the patients they have screened. Additionally, for school nurses who want to keep track of their students by name, there is no way to link the current database to an external list of names for ease of reference. Additionally, the process of adding new research data to the database is manual and slow; adding a new feature to the server to allow faster data entry would be extremely useful for future research, especially as new features get added. Finally, a security 'audit' of some type to ensure compliance with HIPAA and FERPA regarding data privacy and security would ensure that QuickCheck follows all regulations as an in-school healthcare application.

4.6.2.4 Suggestions for Future Clinical Research

The most important suggestion for future clinical tests is to somehow include a consent form for QuickCheck in the standard consent form the Near Vision Institute uses for school visits. NVI sends the consent forms to schools well in advance of their arrival, and parents can then sign a consent form for their children to bring into school. Adding a QuickCheck consent form to the current consent form for schools would allow QuickCheck to be tested during school visits, which would greatly increase the rate at which data could be collected. Additionally, it may allow for QuickCheck to be tested on kindergarten-aged children, which there are currently none of in the study sample. Additionally, utilizing randomization of test order (e.g., whether near or distance vision tests are performed first) would allow future research to control for which order tests are performed when considering by-eye and by-distance results.

There are also factors which might have influenced QuickCheck's performance, but which could not be accounted for with the data gathered as part of this thesis. These factors include whether subjects learn the letter of the optotypes they are being shown, and whether knowing that information changes QuickCheck's performance. Testing performance with and without telling

subjects about what optotypes are used (or about whether crowding bars are used) could help shed light on these currently unstudied factors. Furthermore, while comparing QuickCheck against the Spot screening system was originally planned, that comparison could not be made as not all subjects were screened with Spot, so gathering data using Spot would improve the range of screening tools to compare QuickCheck against.

Finally, there is currently no way to assess the accuracy of QuickCheck's CISS, because there is no reference measurement confirming whether or not subjects have convergence insufficiency or a condition with similar symptoms. Therefore, including tests for convergence insufficiency other than the CISS in future research would allow this aspect of QuickCheck to be more thoroughly analyzed.

4.7 Criteria of Evaluation

As discussed in Section 1.4, this thesis' criteria of evaluation are:

1. The production of a clinically testable QuickCheck which meets the following non-functional requirements:
    a. Safety and Security – the application does not harm its users. There is no leaking of private data, and there being little expected deviation in vision condition prediction from traditional methods of vision screening.
    b. Testability – the application must produce falsifiable results associated with an identifiable individual who has also been tested with a traditional screening approach.
    c. Minimal Data Gathering – the application gathers the minimum amount of information required to make a good estimate of vision quality.

d. Reliability – the application must be reliable enough for several hours of use on several devices.
e. Ease of Use – the application is easy to use for the purpose of clinical verification.
f. Legality – all tests must follow legal rules around the use of medical information and medical exams. QuickCheck must comply with both HIPAA and FERPA.

2. An IRB application for testing the application is accepted.

3. A clinical test comparing QuickCheck to optotype-based vision acuity tests is performed.

4. Data gathered in the clinical test is used to determine what changes to make in the QuickCheck application and the next steps to take in future studies.

As demonstrated in Sections 3 and 4, each of these criteria have been met. Starting with the production of a clinically testable version of QuickCheck, see Section 3.1, which covers the development of QuickCheck performed as part of this thesis. The Non-Functional Requirements a, c, and f are met through the implementation of the Cognito authentication system, the new log in page, and the alterations made to the Student Information scene. NFR b, d, and e were largely developed before this thesis began, but were demonstrated throughout section 4, Of course, the improvement of useability in particular is still in progress, with challenges faced during testing highlighted in Section 4.5.1 and solutions presented in 4.6 largely covering how improvements to useability can still be improved.

Regarding components 2 and 3, these are discussed in the methods section (Chapter 3 Section 2). Both requirements have been met. Finally, requirement 4 has been met, as covered by Chapter 4.

Chapter 5: Conclusion

5.1 Implications of Results

During the clinical testing of QuickCheck, it was discovered that while the application was often able to detect a subject's vision problems if they had any (accuracy of 0.95 ± 0.07, sensitivity of 0.91 ± 0.11, see Table 38), performance by eye was worse (accuracy of 0.84 ± 0.09). Furthermore, overall accuracy was worse for near vision than for distance vision, although near vision tests had lower false negative rates. Simulating different testing policies indicated that switching the policy would not have significantly changed this finding. Additionally, nearly all subjects tested with QuickCheck (on nearly all tests at each distance and eye) were very 'confident', that is, there were very few subjects which were close to having poor vision but did not, and vice-versa. This means that any changes made to the application would likely need to change the test themselves, rather than the analysis QuickCheck performs on test results to determine if a person has a vision problem.

While the test results did not vary much by testing policy, the estimated time of testing did. The longest testing policy (full ten tests) was estimated to take about 300 seconds for acuity testing alone, ignoring other components of the application such as CISS. However, using the shortened five tests policy (a variation on the five tests policy which ends a test early if the subject cannot pass or will always pass), that time could have been reduced to about 130 seconds without significant changes in performance. This is in line with the high 'confidence' of QuickCheck's results.

In the future, continued testing will be needed to address the small sample size of this study. In particular, including a QuickCheck consent in the Near Vision Institute's standard in-school consent form would greatly expand the range of subjects that can be tested and would align the pool of subjects more closely with what QuickCheck is trying to test. Additionally, implementing a way to use LEA symbols for young (kindergarten and first grade-aged) children would improve useability. Also, using the five-test policy would decrease testing time while proving the best possible performance, including the lowest overall false negative rate. Finally, determining a way to decrease QuickCheck's false negative rate (such as slightly decreasing distance optotype sizes) would is likely to QuickCheck's performance as a screening tool, even if it results in a slight decrease in overall accuracy.

5.2 Limitations

There are several key limitations to this thesis. The most important of these limitations is that the study population is very small (18 subjects in total). Furthermore, less than half of those subjects were children, the target population for QuickCheck. This limitation means that it is very difficult to know how well the results of this thesis' clinical testing reflect the performance of QuickCheck. Additional key limitations include:

- The limited type of subjects. Because testing occurred at a single location, the subjects of the trial were atypical in terms of socio-economic status.
- Those the application was tested on may have had goals other than getting the best test results, for example:
    - Getting glasses even if they are unneeded.

    - Playing a ‘game’ with the tester, such as seeing how long it takes the tester to notice they are purposefully missing certain letters.
- Application testers may have ‘given up’ on tests which they might pass if they find it difficult or boring.
- High number of letters presented alongside their low variety mean that people may be able to improve their test results by simple memorization.
- The calibration technique QuickCheck used makes determining ideal optotype size difficult, as there is a wide variation in the size of U.S. quarters.
- QuickCheck could only find refractive errors and convergence insufficiency; other vision problems, such as stereoscopic errors, are not currently tested for.

5.3 Next Steps

To overcome the limitations outlined in sections 4 and 5, further studies on QuickCheck are needed. The most important limitation to overcome is the small study population, further testing will necessarily increase the population size. Ordinarily, significant changes to the application would not be made before further studies. However, the current data strongly suggests that QuickCheck is not good enough at determining which subjects do not have near vision problems, so changes to the near vision acuity tests should be made to attempt to improve performance there. Additionally, further studies should attempt to focus on testing subjects more in line with QuickCheck’s target population, grade-school children.

In addition to the limitations of the clinical test, the QuickCheck application has some key limitations to overcome. First, the poor false negative rates, especially for distance vision, must be reduced. This will likely involve determining a better way to size optotypes as well as a way

for the application to verify that it is at the correct distance from the subject. Additionally, the application should pivot to a testing policy which takes less time than the ten-tests policy; currently, the shortened five test policy seems like the best policy to use in terms of performance and speed. Finally, additional testing types, such as testing stereopsis, should be added, and a way to toggle the type of test the application uses should also be incorporated to allow testers to tailor QuickCheck to their subjects in real time..

5.4 Lessons Learned

As I completed this thesis, I have learned many important lessons about application development, clinical testing, working with others, and maintaining organizational knowledge. In particular, the importance of documentation, open lines of communication, and keeping past work around has been a through-line of my work. Because QuickCheck (both the front and backend) has been worked on by so many people, there are many different places to look for documentation, and there were (and still are) pieces of the application which are no longer used (or used well). However, due to a lack of documentation organization, these components were sometimes difficult to quickly (or safely) adjust or remove. Additionally, this work has emphasized for me the importance of practical, on-the-ground experience in evaluating an application. In that sense, testing QuickCheck not only provided important performance data, but also helped me to understand the context in which the applications will actually be used.

Bibliography

[1] "Vision Screening Significant Legistlative Rule Analysis," Significant Legislative Rules Analysis, Sep. 2016.

[2] "WAC 246-760-020:" https://app.leg.wa.gov/wac/default.aspx?cite=246-760-020 (accessed Sep. 05, 2022).

[3] D. Ethan and C. E. Basch, "Promoting Healthy Vision in Students: Progress and Challenges in Policy, Programs, and Research," *J. Sch. Health*, vol. 78, no. 8, pp. 411–416, 2008, doi: 10.1111/j.1746-1561.2008.00323.x.

[4] S. Goldstand, K. C. Koslowe, and S. Parush, "Vision, Visual-Information Processing, and Academic Performance Among Seventh-Grade Schoolchildren: A More Significant Relationship Than We Thought?," *Am. J. Occup. Ther.*, vol. 59, no. 4, pp. 377–389, Jul. 2005, doi: 10.5014/ajot.59.4.377.

[5] E. L. Irving, A. M. Sivak, and M. M. Spafford, "'I can see fine': patient knowledge of eye care," *Ophthalmic Physiol. Opt.*, vol. 38, no. 4, pp. 422–431, 2018, doi: 10.1111/opo.12566.

[6] CDC, "Vision Loss | Kids' Quest | NCBDDD | CDC," *Centers for Disease Control and Prevention*, Oct. 03, 2019. https://www.cdc.gov/ncbddd/kids/vision.html (accessed Jan. 16, 2023).

[7] "What is the Difference Between a Vision Screening and an Eye Examination? - National Center," Oct. 13, 2020. https://nationalcenter.preventblindness.org/what-is-the-difference-between-a-vision-screening-and-an-eye-examination/ (accessed Jan. 16, 2023).

[8] "File:Snellen chart.svg," *Wikipedia*. Feb. 03, 2011. Accessed: Feb. 08, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=File:Snellen_chart.svg&oldid=411714235#filelinks

[9] "Photoscreening - American Association for Pediatric Ophthalmology and Strabismus." https://aapos.org/glossary/photoscreening (accessed Jan. 30, 2023).

[10] E. Borsting, M. W. Rouse, and P. N. De Land, “Prospective comparison of convergence insufficiency and normal binocular children on CIRS symptom surveys. Convergence Insufficiency and Reading Study (CIRS) group,” *Optom. Vis. Sci. Off. Publ. Am. Acad. Optom.*, vol. 76, no. 4, pp. 221–228, Apr. 1999, doi: 10.1097/00006324-199904000-00025.

[11] “Validity of the Convergence Insufficiency Symptom Survey: A Confirmatory Study,” *Optom. Vis. Sci. Off. Publ. Am. Acad. Optom.*, vol. 86, no. 4, pp. 357–363, Apr. 2009, doi: 10.1097/OPX.0b013e3181989252.

[12] “Fast Facts of Common Eye Disorders | CDC,” Jun. 09, 2020. https://www.cdc.gov/visionhealth/basics/ced/fastfacts.htm (accessed Sep. 02, 2022).

[13] “Prevalence and Impact of Vision Disorders in U.S. Children - Prevent Blindness Wisconsin,” Mar. 25, 2016. https://preventblindness.org/common-childrens-vision-problems-prevalence/, https://wisconsin.preventblindness.org/prevalence-and-impact-of-vision-disorders-in-u-s-children/ (accessed Jan. 16, 2023).

[14] M. Scheiman, J. Gwiazda, and T. Li, “Non-surgical interventions for convergence insufficiency,” *Cochrane Database Syst. Rev.*, no. 3, p. CD006768, Mar. 2011, doi: 10.1002/14651858.CD006768.pub2.

[15] “Vision-related learning problems.” https://www.aoa.org/healthy-eyes/eye-and-vision-conditions/vision-related-learning-problems?sso=y (accessed Oct. 24, 2022).

[16] S. M. Handler, W. M. Fierson, and A. A. of O. the Section on Ophthalmology and Council on Children with Disabilities American Association for Pediatric Ophthalmology and Strabismus, and American Association of Certified Orthoptists, “Learning Disabilities, Dyslexia, and Vision,” *Pediatrics*, vol. 127, no. 3, pp. e818–e856, Mar. 2011, doi: 10.1542/peds.2010-3670.

[17] E. Borsting, M. Rouse, and R. Chu, “Measuring ADHD behaviors in children with symptomatic accommodative dysfunction or convergence insufficiency: a preliminary study,” *Optom. - J. Am. Optom. Assoc.*, vol. 76, no. 10, pp. 588–592, Oct. 2005, doi: 10.1016/j.optm.2005.07.007.

[18] A. Grzybowski, P. Kanclerz, K. Tsubota, C. Lanca, and S.-M. Saw, “A review on the epidemiology of myopia in school children worldwide,” *BMC Ophthalmol.*, vol. 20, no. 1, p. 27, Jan. 2020, doi: 10.1186/s12886-019-1220-0.

[19] P. K. Hrynchak, A. Mittelstaedt, C. M. Machan, C. Bunn, and E. L. Irving, “Increase in Myopia Prevalence in Clinic-Based Populations Across a Century,” *Optom. Vis. Sci.*, vol. 90, no. 11, p. 1331, Nov. 2013, doi: 10.1097/OPX.0000000000000069.

[20] R. Varma, K. Tarczy-Hornoch, and X. Jiang, “Visual Impairment in Preschool Children in the United States: Demographic and Geographic Variations From 2015 to 2060,” *JAMA Ophthalmol.*, vol. 135, no. 6, pp. 610–616, Jun. 2017, doi: 10.1001/jamaophthalmol.2017.1021.

[21] H. Snellen, *Probebuchstaben zur Bestimmung der Sehschärfe*. 1862.

[22] “Craig Blackwell M.D. Opthamology » Calc Size of 20/20 Letter,” Jan. 05, 2012. https://web.archive.org/web/20120105104152/http://www.blackwelleyesight.com/eye-math/2020-letter/ (accessed Feb. 08, 2023).

[23] Fvasconcellos, *English: ETDRS Chart R, one of the three logMAR chart models designed by Ferris, Kassoff, Bresnick, and Bailey for use in the Early Treatment Diabetic Retinopathy Study (ETDRS).* 2021. Accessed: Feb. 08, 2023. [Online]. Available: https://commons.wikimedia.org/wiki/File:ETDRS_Chart_R.svg

[24] “Near Vision LEA Card with 16 inch (40cm) Cord.” https://www.amconlabs.com/product/4461/Near-Vision-LEA-Card-with-16-inch-(40cm)-Cord/ (accessed Mar. 14, 2023).

[25] “Symbol Near Vision Card-66071,” *School Nurse Supply®*. https://www.schoolnursesupplyinc.com/Symbol-Near-Vision-Card_p_3603.html (accessed Mar. 04, 2023).

[26] “Near Vision Card , Near: Bernell Corporation.” https://www.bernell.com/product/ODNVC/Near (accessed Mar. 14, 2023).

[27] H. Zoltán, “Color Blind Test,” *Colorlite | Color Blind Glasses | Color Blind Test*. https://www.colorlitelens.com/color-blind-test.html (accessed Mar. 15, 2023).

[28] M. Bach, “FrACT – Homepage.” https://michaelbach.de/fract/ (accessed Feb. 08, 2023).

[29] I. Krumholtz, “Educating the educators: increasing grade-school teachers’ ability to detect vision problems,” *Optom. - J. Am. Optom. Assoc.*, vol. 75, no. 7, pp. 445–451, Jul. 2004, doi: 10.1016/S1529-1839(04)70159-1.

[30] S. J. H. Lalor, M. A. Formankiewicz, and S. J. Waugh, “Crowding and visual acuity measured in adults using paediatric test letters, pictures and symbols,” *Vision Res.*, vol. 121, pp. 31–38, Apr. 2016, doi: 10.1016/j.visres.2016.01.007.

[31] “Welch Allyn Spot Vision Screener.” https://www.hillrom.com/en/products/spot-vision-screener/ (accessed Sep. 05, 2022).

[32] H. Gaiser, B. Moore, G. Srinivasan, N. Solaka, and R. He, “Detection of Amblyogenic Refractive Error Using the Spot Vision Screener in Children,” *Optom. Vis. Sci.*, vol. 97, no. 5, p. 324, May 2020, doi: 10.1097/OPX.0000000000001505.

[33] G. Srinivasan, D. Russo, C. Taylor, A. Guarino, P. Tattersall, and B. Moore, “Validity of the Spot Vision Screener in detecting vision disorders in children 6 months to 36 months of age,” *J. Am. Assoc. Pediatr. Ophthalmol. Strabismus*, vol. 23, no. 5, p. 278.e1-278.e6, Oct. 2019, doi: 10.1016/j.jaapos.2019.06.008.

[34] “Welch Allyn Spot Vision Screener,” *MFI Medical*. https://mfimedical.com/products/welch-allyn-spot-vision-screener (accessed Oct. 24, 2022).

[35] “Welch Allyn Spot Vision Screener,” *MDMaxx*. https://mdmaxx.com/products/welch-allyn-vs100s-b-spot-vision-screener (accessed Oct. 24, 2022).

[36] “Welch Allyn VS100S-B-WelchAllyn SPOT VISION SCREENER,W/CASE,PLUGB/US,” *MedicalDeviceDepot.com*. https://www.MedicalDeviceDepot.com/product-p/vs100s-b-welchallyn.htm (accessed Oct. 24, 2022).

[37] P. Glewwe, A. Park, and M. Zhao, “A better vision for development: Eyeglasses and academic performance in rural primary schools in China,” *J. Dev. Econ.*, vol. 122, pp. 170–182, Sep. 2016, doi: 10.1016/j.jdeveco.2016.05.007.

[38] M. Bach, “The Freiburg Visual Acuity Test-Variability unchanged by post-hoc re-analysis,” *Graefes Arch. Clin. Exp. Ophthalmol.*, vol. 245, no. 7, pp. 965–971, Jul. 2006, doi: 10.1007/s00417-006-0474-4.

[39] M. Bach, “The Freiburg Visual Acuity Test—Automatic Measurement of Visual Acuity,” *Optom. Vis. Sci.*, vol. 73, no. 1, p. 49, Jan. 1996.

[40] L. N. Ayton *et al.*, “Harmonization of Outcomes and Vision Endpoints in Vision Restoration Trials: Recommendations from the International HOVER Taskforce,” *Transl. Vis. Sci. Technol.*, vol. 9, no. 8, p. 25, Jul. 2020, doi: 10.1167/tvst.9.8.25.

[41] “Subretinal electronic chips allow blind patients to read letters and combine them to words | Proceedings of the Royal Society B: Biological Sciences.” https://royalsocietypublishing.org/doi/10.1098/rspb.2010.1747 (accessed Mar. 15, 2023).

[42] D. Palanker, Y. Le Mer, S. Mohand-Said, and J. A. Sahel, “Simultaneous perception of prosthetic and natural vision in AMD patients,” *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Jan. 2022, doi: 10.1038/s41467-022-28125-x.

[43] “Virtual Vision Test | Prescription Renewal,” *Warby Parker*. https://www.warbyparker.com (accessed Dec. 17, 2022).

[44] “GoCheck Kids - Clinical.” https://www.gocheckkids.com/clinical/ (accessed Jan. 17, 2023).

[45] “GoCheck Kids - Pediatric Vision Screening Solution.” https://www.gocheckkids.com (accessed Jan. 17, 2023).

[46] R. W. Arnold, J. W. O’Neil, K. L. Cooper, D. I. Silbert, and S. P. Donahue, “Evaluation of a smartphone photoscreening app to detect refractive amblyopia risk factors in children aged 1-6 years,” *Clin. Ophthalmol. Auckl. NZ*, vol. 12, pp. 1533–1537, 2018, doi: 10.2147/OPTH.S171935.

[47] M. M. W. Peterseim *et al.*, “Effectiveness of the GoCheck Kids Vision Screener in Detecting Amblyopia Risk Factors,” *Am. J. Ophthalmol.*, vol. 187, pp. 87–91, Mar. 2018, doi: 10.1016/j.ajo.2017.12.020.

[48] “GoCheck Kids - Compare Leading Pediatric Vision Screeners.” https://www.gocheckkids.com/compare-screeners (accessed Feb. 08, 2023).

[49] “GoCheck Kids - Workflow.” https://www.gocheckkids.com/workflow/ (accessed Feb. 08, 2023).

[50] “Start Simple with MyPlate | MyPlate.” https://www.myplate.gov/resources/tools/startsimple-myplate-app (accessed Jan. 17, 2023).

[51] “Virtual Mason,” *App Store*. https://apps.apple.com/us/app/virtual-mason/id1460283255 (accessed Jan. 17, 2023).

[52] A. H. Dahlmann-Noor *et al.*, “Vision screening in children by Plusoptix Vision Screener compared with gold-standard orthoptic assessment,” *Br. J. Ophthalmol.*, vol. 93, no. 3, pp. 342–345, Mar. 2009, doi: 10.1136/bjo.2008.138115.

[53] “Plusoptix: Vision Screening - Plusoptix Vision Screener.” https://www.plusoptix.com/en-us/products/vision-screener/vision-screening (accessed Jan. 17, 2023).

[54] S. Gupta, D. Chavan, and T. De, “Validation of the Smartphone-Based Snellen Visual Acuity Chart for Vision Screening,” vol. 11, pp. 17–28, Mar. 2023.

[55] J. M. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” *Lancet Lond. Engl.*, vol. 1, no. 8476, pp. 307–310, Feb. 1986.

[56] P. Ranganathan, C. S. Pramesh, and R. Aggarwal, “Common pitfalls in statistical analysis: Measures of agreement,” *Perspect. Clin. Res.*, vol. 8, no. 4, pp. 187–191, 2017, doi: 10.4103/picr.PICR_123_17.

[57] “A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960.” https://journals.sagepub.com/doi/10.1177/001316446002000104 (accessed May 20, 2023).

[58] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012.

[59] “Chapter 246-760 WAC:” https://apps.leg.wa.gov/wac/default.aspx?cite=246-760&full=true (accessed May 08, 2023).

[60] O. for C. Rights (OCR), “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule,” *HHS.gov*, Sep. 07, 2012. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed May 08, 2023).

[61] “Family Educational Rights and Privacy Act (FERPA),” Aug. 25, 2021. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html (accessed Oct. 12, 2022).

[62] “Health Information & Privacy: FERPA and HIPAA | CDC,” Oct. 11, 2022. https://www.cdc.gov/phlp/publications/topic/healthinformationprivacy.html (accessed Oct. 12, 2022).

[63] K. Kalyani, “User-Centered Design and Usability Study of Android-Based Vision Screening App QuickCheck,” M.S. thesis, University of Washington - Bothell, 2019.

[64] “Underlying neurological mechanisms associated with symptomatic convergence insufficiency | Scientific Reports.” https://www.nature.com/articles/s41598-021-86171-9 (accessed Feb. 27, 2023).

Appendix A: QuickCheck Functional Requirements

The functional requirements in Tables A.A.1, A.A.2, and A.A.3 are taken from the initial QuickCheck specification documents written by prior EYE research group members, including Dr. William Erdly, Dr. Alan Pearson, Kai Yang, Aleksander Dimitrov, Jerahmy Bindon, Max Nguyen, Michael Lee, Charlie Cox, Ahmad Yosif, Sesario Imanputra, and Huantong Ji. The documentation these tables are taken from was not produced for this document.

Req-ID	Requirement	Measure
FR-UDC-01	The application must notify the user on the limitations of the screening application's results	Documentation on the current limitation specified to the user when notified
FR-UDC-02	The application must notify the user on its' capability to store and disclose collected information	Documentation on the current information specified to the user regarding information storage and disclosure
FR-UDC-03	The application must save and store user's consent of storing and disclosing their information	Documentation of ER diagram that explicitly describes how and where user consent is stored
FR-UDC-04	The application must only record the following user information: consent of data collection and disclosure, grade, birthday, zip code, prescription eyeglasses.	Documentation of ER diagram and any other diagrams that describe QuickCheck™ schema
FR-UDC-05	The application must validate input fields by providing users pre-selected options	Documentation on what pre-selected options are implemented within the app
FR-UDC-06	The application must collect both anonymous data and user associated data	Documentation on the difference in activities between both anonymous data and user associated data

Table A.A.1: User Data Collection Requirements

Req-ID	Requirement	Measure
FR-VST-01	The application shall use a specific number of letter presentations per test such that the results effectively describe the user's vision	Evaluation by Dr. Pearson on the current number of letter presentation per test
FR-VST-02	The application must design the following screening tests: NOU, NOD, NOS, DOU, DOD, DOS*	Evaluation on behalf of Dr. Pearson on the current designs of screening tests
FR-VST-03	The application must ask the user to disclose any ongoing symptoms of visual impairment	Documentation on the current information stored in databases
FR-VST-04	The application must meet the Washington state standard when designing each test	Evaluation on behalf of Dr. Pearson on the current designs of the screening tests
FR-VST-05	The application shall record each response to the database, not only the average results	Documentation on the business logic of QuickCheck™.
FR-VST-06	The application must allow users to receive their test results through a specified email	Documentation describing the functionality and output of the email server/function
FR-VST-08	The application must show a tutorial preceding a screening test	Documentation that describes the process of designing tutorials and the current version of the tutorials
FR-VST-9	The application must not use any design elements that would influence the results of a user's test	Documentation that describes the process of designing tests and the current versions of the tests
FR-VST-10	The application must clearly explain concerns discovered regarding a user's vision and next steps that can be taken	Documentation that describes a stage in the application where results are given to the user
FR-VST-11	The application must not prompt users of adolescence complex questions	Documentation on activities for adolescent users

Table A.A.2: Visual Screening Test Requirements

Req-ID	Requirement	Measure
FR-CAL-01	The application must prompt the user to scale objects within the app on their first experience	Documentation on the current design process of calibration
FR-CAL-02	The application must enable users to re-scale objects in the menu screen	Documentation on the current design process of calibration
FR-CAL-03	The application must use a physical coin as a reference point to the correct size of an app's object	Documentation on the current design process of calibration
FR-CAL-04	The application must use a slider to resize the referenced object during calibration	Documentation on the current design process of calibration

Table A.A.3: Calibration Requirements

*These are shorthand for test types; N indicates a near vision test, D indicates a distance vision test, OU represents a test for both eyes, OD represents a right-eye only test, and OS represents a test for the left eye only. For example, NOD is a test for near vision using only the right eye.

Appendix B: CISS Questions

The CISS consists of 15 questions, each with 5 possible answers, which are always “Never”, “Infrequently”, “Sometimes”, “Fairly Often”, and “Always”. The questions are supplied in Table A.B.1 along with the order they are presented by default.

No.	Question Text
1	Do your eyes feel tired when reading or doing close work?
2	Do your eyes feel uncomfortable when reading or doing close work?
3	Do you have headaches when reading or doing close work?
4	Do you feel sleepy when reading or doing class work?
5	Do you lose concentration when reading or doing class work?
6	Do you have trouble remembering what you have read?
7	Do you have double vision when reading or doing close work?
8	Do you see the words jump, swim, or appear to float on the page when reading or doing close work?
9	Do you feel like you read slowly?
10	Do your eyes ever hurt when reading or doing close work?
11	Do your eyes feel sore when reading or doing close work?
12	Do you feel a “pulling” feeling around your eyes when reading or doing close work?

13	Do you notice words blurring or coming in and out of focus when reading or doing close work?
14	Do you lose your place while reading or doing close work?
15	Do you have to re-read the same line of words when reading?

Table A.B.1: CISS Questions