Preliminary assessment of three protocols for the screening of amblyopia through Monte Carlo simulation

Keely Shand
Department of Biomedical Engineering
University of Strathclyde
Glasgow, United Kingdom
keely.shand@strath.ac.uk

Atirut Boribalburephan
Looloo Technology
Bangkok, Thailand
atirutbor1@gmail.com

Mario E. Giardini
Department of Biomedical Engineering
University of Strathclyde
Glasgow, United Kingdom
mario.giardini@strath.ac.uk

Abstract— **Amblyopia is a neurodevelopmental disorder that causes irreversible vision loss in one eye. In order to be effective, treatment needs to start as early as possible, ideally in early infancy. The screening for amblyopia assesses the disparity of the child's visual acuity (VA). In pre-verbal children, this screening is performed using a preferential looking test. These testing protocols rely on a subjective estimate by the test operator of the child's attention, and this estimate is subject to errors. Little is understood of the quantitative impact of error rates on screening performance. In this paper, a Monte Carlo simulation to compare the clinical performance of three preferential-looking test protocols for paediatric VA screening tests for amblyopia is described. The inter-protocol differences of "Cardiff Acuity Test" (CAT), "Keeler Acuity Cards for Infants" (KACI) and "Teller Acuity Cards" (TAC) are assessed by iteratively executing through a simulation loop of an examiner testing a test subject using the three set protocols. The measured VA from each protocol and the actual VA have been compared using Bland-Altman statistics. It was determined that CAT and KACI both have a systematic bias, whereby they measure the VA at a greater logMAR value than the smallest testing resolution. KACI bias is greater due to the greater step size used.**

Keywords*— *Amblyopia Screening, Monte Carlo Simulation, Preferential-Looking test.

I. Introduction

Amblyopia, also known as 'lazy eye', is a visual development disorder that occurs in early childhood. It originates from one of the eyes seeing better than the other. As the child grows, the eye with better vision overcompensates for the affected eye. This weakens the visual processing pathway to the affected eye, further reducing the vision until it is functionally lost [1]. It is the highest cause of childhood vision loss, with up to 5 % of children being affected by the condition [2]. If detected early, treatment is simple, consisting mostly of patching the eye with better vision, to allow the child to learn to use the poorer eye. However, due to reduction of neuroplasticity with age, if not treated within early childhood, the vision loss becomes irreversible. Yet, amblyopia goes easily unnoticed by children and their families [3]. Therefore, childhood screening, particularly before the age of seven, is of paramount importance [2].

A screening infrastructure is set in place that assesses the disparity of the infant's visual acuity (VA, the capacity to distinguish fine detail) between the eyes [3]. The gold standard for testing someone's VA is through an optotype acuity test, such as the ETDR letter chart [4] or, for pre-scholar children, a picture-based test [5]. However, these charts are not suitable for children too young to follow instructions or to describe what they see [4]. Therefore, in pre-verbal children, the screening of the infant's VA can be performed through a preferential-looking test [6]. This family of tests is based on the assumption that, if the examiner shows the infant a visual stimulus, the infant will instinctively direct its attention towards the stimulus [6]. By detecting the diversion of attention, it is therefore possible to detect whether the stimulus has been seen. There are numerous versions of this test [5], [7], the most common being "Cardiff Acuity Test" (CAT) [8], "Keeler Acuity Cards for Infants" (KACI) [9], and "Teller Acuity Cards" (TAC) [10]. These major protocols use the same form of stimuli, consisting of homogeneous grey cards containing a black and white grating with the same average illuminance of the card's grey background. The examiner presents these cards to the infant who, if they see the grating, will divert their attention towards it, briefly looking at it. The examiner detects this attention diversion by performing a subjective evaluation of the infant's looking direction. A sequence of these cards is presented to the infant, with decreasing grating spacing. When the grating spacing becomes too small for the infant to see, the grating will appear grey and become indistinguishable from the homogeneous grey background, and the infant's attention will no longer be diverted by the grating. The smallest grating thickness that the infant can see determines their VA [6].

In the test, the examiner needs to interpret the infant's response to these targets, which is used to determine if the stimulus has been seen or not [4]. This interpretation is then used to inform the examiner which grating spacing to show the infant next. This staircasing protocol allows the examiner to home in the infant's VA quickly [11]. Each of CAT, KACI, and TAC have distinct staircasing methods, with some instructing multiple shows of the same target [8], [12], and/or instructing to reverse the direction of the staircasing once the infant has not seen a target [8], [9]. This can affect the final measured VA given by each protocol.

Interpreting the infant's response to targets is difficult as an infant's attention span is short, and as infants will prefer to look at the examiner, rather than at the targets. So, the examiner needs to be fast, all while being hidden, e.g., by a board or by the card itself and viewing the child through an aperture to avoid drawing the infant's attention away from the test [12]. The intrinsic subjectivity in the detection of the diversion of attention and difficulty in the implementation of the testing protocol can result in testing errors. Also, the

This research was funded by the University of Strathclyde SRSS REA programme and IDCP Scotland Ltd.

targets physically degrade over time, which affects the accuracy of the test [13].

To address this, there has been several attempts to digitize the test [9], [14] and, to aid the examiners' interpretation of the test subject's attention, automation of the test by tracking the infant's gaze has been proposed [15], [16], [17], [18], [19]. Yet, gaze tracking is in itself a complex task, especially given that an infant cannot be instructed to perform the complex calibration required by most eye trackers, which compounds with the fact that, in itself, gaze direction is effectively a proxy for attention diversion. Therefore, this inference is subject to errors, whether performed by an examiner or by a gaze-based algorithm. In order to design a clinically viable tool, this opens the need to evaluate the performance of a diverse family of protocols, in the presence of examiner errors, to determine the test performance in terms of suitable metrics (e.g., uncertainties in the measurement of absolute VA and inter-eye VA difference, test duration, sensitivity to test subject engagement) as a function of the examiner error rate.

The performance of preferential looking protocols has been assessed in research [9], [20], [21], [22]. However, these studies rarely directly compare protocols to each other, and the study groups are small. Yet, given the deterministic nature of the tests and the well-defined protocol sequence, this assessment lends itself to simulation, at least to determine inter-test differences. In this paper, therefore, a Monte Carlo simulation to compare the clinical performance of three preferential-looking test protocols for paediatric VA screening tests for amblyopia is described.

II. Monte Carlo Model

Monte Carlo (MC) Simulation is a statistical modelling strategy that can be used to predict outcomes of a family of processes where the process-to-process variability can be parameterized [23]. In this paper, the MC model was designed to simulate 3 major preferential-looking VA test protocols. The computational experiment was completed by iteratively executing simulations of an examiner testing a test subject. The test subject is described by their (monocular) VA and the examiner by their error rate.

Each simulation consisted of a screening of the VA of the left eye. The screening involved the following steps.

1. The examiner shows the patient a target.
2. The examiner interprets whether the patient has seen the target or not.
3. The examiner shows the next target which is determined by referring to the protocol used.
4. The examiner repeats the previous steps until the protocol dictates that the criteria for stopping the test has been met. The measured VA is recorded.

Although the protocols follow the same basic show-interpret-staircase structure, the staircasing strategies are different, as can be seen by the protocol flowcharts in Fig. 1,2 and 3. For example, in the TAC protocol each target is shown three times before decreasing the coarseness of the grating, whereas it is twice and once in CAT and KACI respectively. The protocols were modelled based on the instructions provided by the commercial implementation of these tests

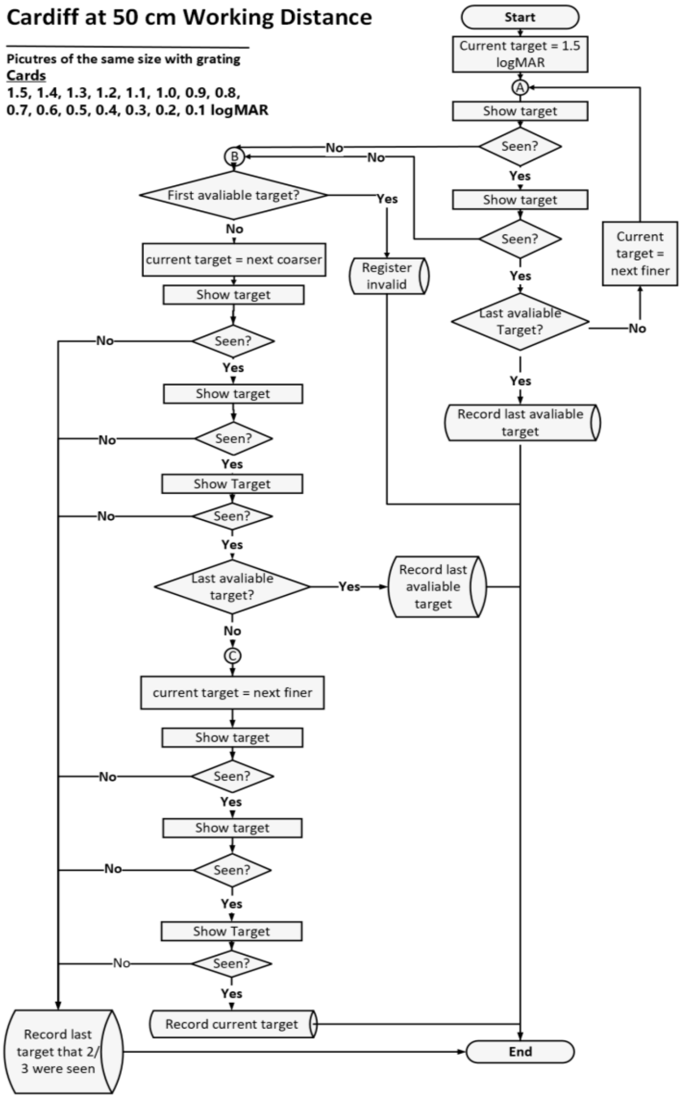
Cardiff at 50 cm Working Distance
Picutres of the same size with grating
Cards
1.5, 1.4, 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1 logMAR
Start
Current target = 1.5 logMAR
A
Show target
Seen?
No
No
Yes
Show target
Seen?
Yes
Last avaliable Target?
No
Current target = next finer
Yes
Record last avaliable target
B
First avaliable target?
Yes
Register invalid
No
current target = next coarser
Show target
Seen?
No
Yes
Show target
Seen?
No
Yes
Show Target
Seen?
No
Yes
Last avaliable target?
Yes
Record last avaliable target
No
C
current target = next finer
Show target
Seen?
No
Yes
Show target
Seen?
No
Yes
Show Target
Seen?
No
Yes
Record current target
Record last target that 2/3 were seen
End

Fig. 1. Flowchart of the CAT protocol.

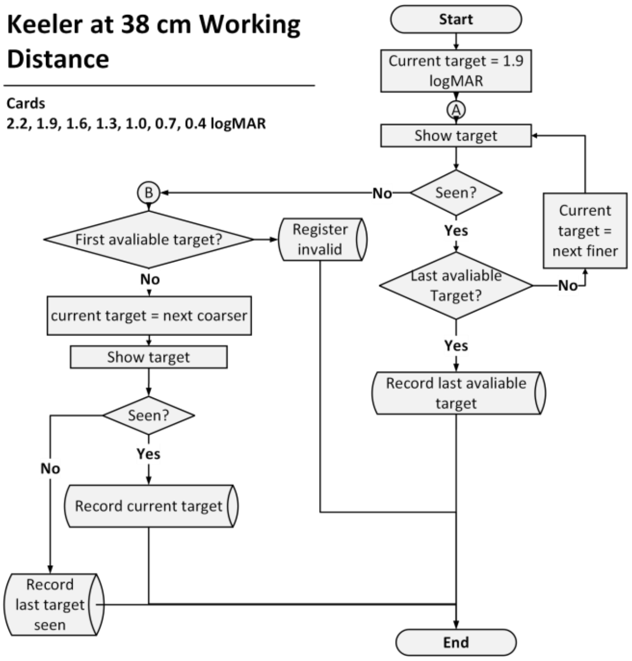
Keeler at 38 cm Working Distance
Cards
2.2, 1.9, 1.6, 1.3, 1.0, 0.7, 0.4 logMAR
Start
Current target = 1.9 logMAR
A
Show target
Seen?
No
Yes
Last avaliable Target?
No
Current target = next finer
Yes
Record last avaliable target
B
First avaliable target?
Register invalid
No
current target = next coarser
Show target
Seen?
No
Yes
Record current target
Record last target seen
End

Fig. 2. Flowchart of the KACI protocol.

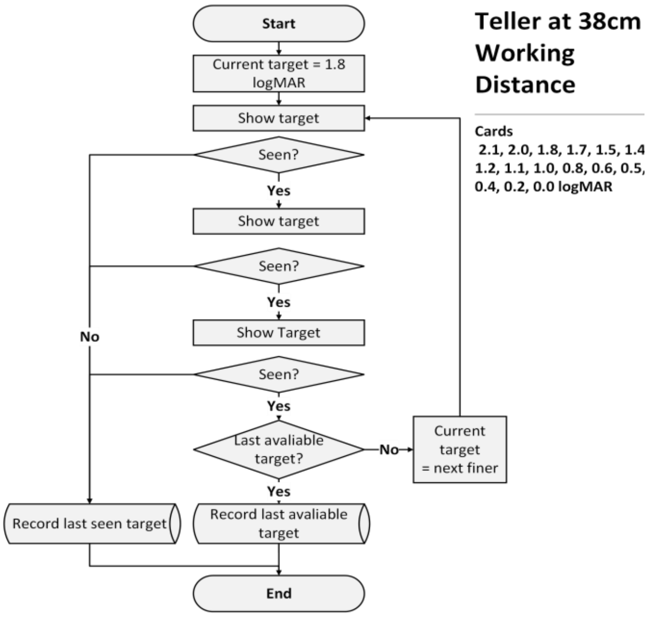
Teller at 38cm Working Distance
Cards
2.1, 2.0, 1.8, 1.7, 1.5, 1.4 1.2, 1.1, 1.0, 0.8, 0.6, 0.5, 0.4, 0.2, 0.0 logMAR
Start
Current target = 1.8 logMAR
Show target
Seen?
Yes
Show target
Seen?
Yes
Show Target
Seen?
Yes
No
Last avaliable target?
No
Current target = next finer
Yes
Record last seen target
Record last avaliable target
End

Fig. 3. Flowchart of the TAC protocol.

[8], [9], [10]. As some of these protocols can be performed at different distances, with nuances reflected in the staircasing, we set the testing distances to the most common ones used, namely 38 cm for TAC and KACI, and 50 cm for CAT. Each protocol also has its own set of targets which are listed in Fig. 1,2 and 3, both TAC and CAT have 15 targets, ranging from 1.5 to 0.1 logMAR and 2.1 to 0.0 logMAR respectively, whereas KACI has a set of 7 targets ranging from 2.2 to 0.4 logMAR. These protocols and their unique staircasing were integrated into the simulation.

III. Methods

In this paper we are using this MC model to compare the clinical performance of the three major protocols. To do this, a simulation loop was created. Each loop created a test subject and an examiner. The test subject's VA was chosen randomly, from a pool of uniform distribution to spread the data points amongst the full range analyzed by the comparative statistics, ranging from [0.0, 2.2] logMAR. This range was decided based on the VA range that can nominally be determined with the three protocols. The examiner's competency was set to 95 % [24], with 100 % meaning that the examiner would interpret the test subject's response correctly every time. In our case, the interpretation is correct in 95 % of cases, and the other 5 % of the interpretations were then randomly decided between seen (45 %), unseen (45 %) and disengaged (10 %). We note that, in the current absence of any quantitative literature, in further studies this assumption may need to be revised. The test subject is then tested on each of the protocols, CAT, KACI, and TAC. At the end of the simulation loop, the test subject's VA, the measured VA both of which are measured in terms of logMAR from each of the protocols and whether each of these measurements were valid (True/False) were recorded.

This simulation loop was run 2300 times. This number was defined in order to allow for 100 test subjects for each acuity step within the test, 23 (0.1 increments between 0.0 to 2.2, which corresponds to the combined extrema and best resolution of the protocols). This value was set on by running preliminary tests on the simulation, analysing the variation seen in the VA dataset produced. The mean and standard deviation of test subject dataset was determined to have negligible change from 230 to 2300 iterations. With the simulation runtime being under 30 s on an Intel i7-10870H processor running at 2.20 GHz, the number of iterations was opportunistically set to 2300 to provide a safe margin to the simple statistical considerations in the study.

The collected data is then screened to remove the measurements invalidated by test subject disengagement or the actual VA being outside measurable range for the specific protocol being simulated. The raw dataset is screened for measurement validity for each protocol separately, producing three separate datasets for each protocol. The number of valid tests that remained after each screening is recorded at each screening step and can be seen in Table 1.

The VA measurements produced by each of the protocols (logMAR) was then compared to the actual VA (logMAR). This comparison was done by plotting the mean of the protocol's measured VA and the actual VA against the difference of the measurement and the actual VA, producing a Bland-Altman (BA) plot [25]. These plots can be seen in Fig. 4 for CAT, Fig. 5 for KACI, and Fig. 6 for TAC.

IV. Results

In Table 1, the size of the dataset after the removal of invalid measurements is reported.

The comparison of the protocol's measurement to the actual VA can be seen in Fig. 4 for the CAT, Fig. 5 for the KACI, and Fig. 6 for TAC.

In Fig. 4, the mean of the differences of CAT against actual VA is 0.18 logMAR. Their limits of agreement (LoA) are 0.84 logMAR and -0.48 logMAR. Most points are in the positive difference semiplane, and their distribution creates a triangular shape, where the greatest positive difference is for a mean of 0.79 logMAR, difference of 1.44 logMAR, with mean from 0.05 logMAR to 1.50 logMAR. The data points with a negative difference show no distinguishable pattern in their distribution. They range from the means of 0.30 logMAR to 1.80 logMAR and the greatest negative difference is -1.75 logMAR.

In Fig. 5, the mean of differences between the measurements of KACI and the actual VA is 0.21 logMAR. Their LoA are 0.77 logMAR and -0.34 logMAR. Again, most of the data points lie in the positive difference semiplane. Their distribution shows the lowest differences are near the extremities of the range seen at approximately 0.20 logMAR to 2.20 logMAR. The greatest positive difference is 2.05 logMAR and can be seen at a mean of 1.20 logMAR. A

TABLE 1. The size of the dataset at each stage of the removal of invalid measurements for each comparison.

Protocol	**Size of Dataset (test results)**	
	Before validity screening	***After validify screening***
CAT	2300	1413
KACI		2239
TAC		2181

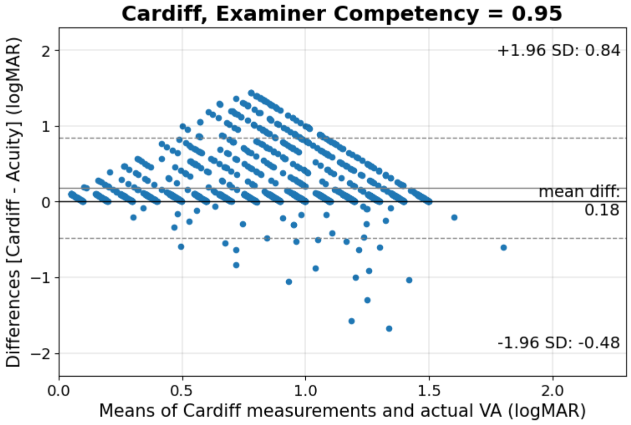

Fig. 4. CAT measurements against actual visual acuity in a Bland Altman Plot, with a mean difference of 0.18 logMAR and limit of agreement 0.84 to -0.48 logMAR.

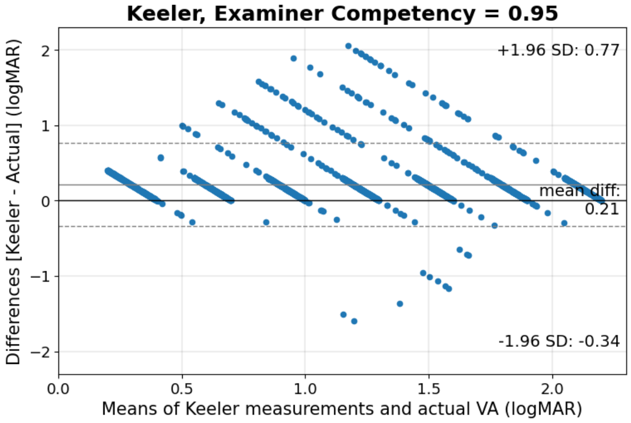

Fig. 5. KACI measurements against the actual visual acuity in a Bland Altman Plot, with a mean difference of 0.21 logMAR and limit of agreement 0.77 to -0.34 logMAR.

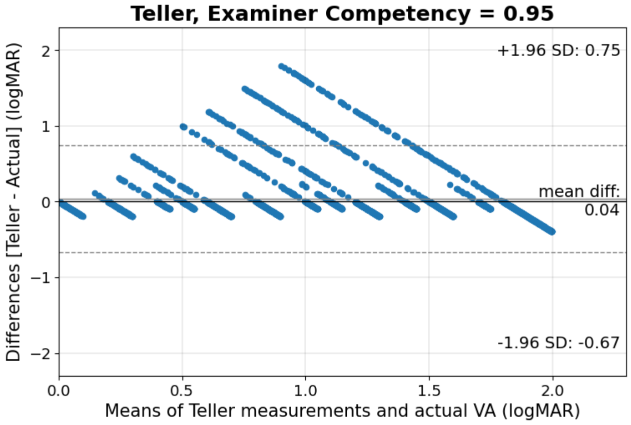

Fig. 6 TAC measurements against the actual visual acuity in a Bland Altman Plot, with a mean difference of 0.17 logMAR and limit of agreement 1.00 to -0.66 logMAR

pattern can be seen in the datapoints with a negative difference. The points create a trend where from the mean of 1.16 logMAR, the difference reduces from -1.50 logMAR to the negative LoA, as the mean increases to 1.67 logMAR.

In Fig. 6, the mean of differences between the measurement produced by TAC and actual VA is 0.04 logMAR. The LoA are 0.75 logMAR to -0.67 logMAR. Again, the majority of the datapoints lie in the positive difference semiplane. The distribution of all points ranges from the means of 0.00 logMAR to 2.00 logMAR. This range is smaller than the logMAR range of the targets provided from the protocol. The greatest difference is 1.80 logMAR which is seen at the mean of 0.90 logMAR. The few points with a negative difference all lie within LoA, the greatest negative difference is -0.40 logMAR which also is the highest mean at 2.00 logMAR.

V. Discussion

As seen in Table 1, the greatest number of invalid measurements was found in the CAT protocol. As the examiner’s compliance was set to an “almost perfect” compliance of 95 %, then these invalid measurements were due to the shorter VA range within CAT’s targets. Any test subject above 1.50 logMAR would have produced an invalid measurement. Since the ranges on the other two protocols reach a higher logMAR value, the number of invalid measurements was drastically smaller.

In all three BA plots (Fig. 4 – 6) the pattern seen within the distributions can be determined to be related to the targets available. Each of the repeating slants represents a target, as can be determined by cross referencing them to the slant’s x-intercepts. The number of slants is equal to the number of targets provided in the CAT and KACI, 15 and 7 respectively. However, only 13 slants can be seen in the BA plot for TAC, the two targets that are not represented are the 2.00 logMAR and the 2.10 logMAR, which are the highest value targets. The reason they are represented is that within the protocol, they are never used. The protocol instructs the test should start on the third highest target and there is no reversal staircasing. The question is then, why have them? The assumption is that, even though it is not described in the formal instruction documents provided with the commercial implementation of TAC, if the test subjects does not see the third greatest target (1.80 logMAR), then these will be used, albeit opportunistically and with no explicit protocol provided. If the test subject does not see these targets, then the examiner is instructed to drop the testing distance down to the next one listed, in this case that would be 19 cm [12]. This situation is not common as at 2.10 logMAR as, were this to occur, the VA would be so low as to be easily noticed by the infant and/or family, and the test subject would therefore in any case be considered for the classification of “severely sight impaired” [26].

We note that CAT and KACI have a systematic bias, whereby the mean of differences has a value greater than the smallest resolution of 0.1 logMAR that is used in the protocols. Since both of their mean of differences are positive, with CAT being 0.18 logMAR and KACI being 0.21 logMAR, the bias will systematically overestimate the logMAR value. This means that these protocols estimate the test subject’s vision to be worse than it is. The main difference between these two protocols and TAC is that the CAT and KACI protocols include an element of reversal of the staircasing direction (moving from narrow to wide gratings). Additionally, in CAT this reverse staircasing follows rules that are asymmetrical with respect to those followed when narrowing the grating spacing. When screening for amblyopia, indeed we wish to measure an inter-eye VA difference, rather than the VA of a single eye, as in this simulation. However, assuming one eye is emmetropic, “perfect vision”, at the testing distance, or close to being so, from Fig. 4-6 all tests report a very small bias for a VA close to 0.00 logMAR and, indeed, the single-eye VA bias is indicative of the inter-eye difference bias. Whether a large bias corresponds to a worse or better test ultimately depends on the reason for testing. If, indeed, testing is directed towards screening, avoiding false negatives takes priority over providing accurate results and, therefore, tests with higher bias may arguably be preferrable.

The KACI protocol has the greatest bias amongst those examined. While we do not have a statistical model for this, we note that this is reasonable due to two aspects. The first is that the step size used in KACI is 0.3 logMAR, unlike CAT which uses targets with a step size of 0.1 logMAR, and therefore, when a target reversal occurs, the test target up for KACI has a greater increase. The other aspect is that unlike KACI, CAT not only presents reversals in the staircasing, but also these reversals appear multiple times during the protocol.

An important limitation in the study so far is represented by modelling the examiner as having competency close to perfect, and assuming arbitrary values for the effect of examiner errors. Fatigue of the test subjects has not been accounted for. The behaviour described for the three test protocols may change as these issues are considered, and our future work will therefore need to address them.

VI. Conclusion

To conclude, a Monte Carlo simulation to compare the clinical performance of three preferential-looking test protocols for paediatric VA screening tests for amblyopia was created. From this simulation, it was determined that CAT and KACI protocols both have a systematic bias, whereby they measure the VA at a greater logMAR value than the smallest testing resolution. KACI has a greater bias which is due to the larger step size used in the protocol and it only uses one reversal in its staircasing. Future work will address the relative behaviour between protocols in the presence of varying examiner errors.

Acknowledgment

With thanks to Dr. Craig Robertson, University of Strathclyde, and Prof. Sam Philip, NHS Grampian for the support in this paper.

References

[1] "Strabismus and amblyopia in children (squint and lazy eye)," RNIB. Accessed: Nov. 20, 2023. [Online]. Available: https://www.rnib.org.ukhttps://www.rnib.org.uk/your-eyes/eye-conditions-az/strabismus-and-amblyopia-in-children-squint-and-lazy-eye/

[2] J. M. Holmes and M. P. Clarke, "Amblyopia," *The Lancet*, vol. 367, no. 9519, pp. 1343–1351, Apr. 2006, doi: 10.1016/S0140-6736(06)68581-4.

[3] J. M. Jefferis, A. J. Connor, and M. P. Clarke, "Amblyopia," *BMJ*, vol. 351, no. nov12 1, pp. h5811–h5811, Nov. 2015, doi: 10.1136/bmj.h5811.

[4] B. James and L. Benjamin, "The Assessment of vision," in *Ophthalmology: Investigation and Examination Techniques*, Butterworth Heinemann, 2007. doi: 10.1016/B978-0-7506-7586-4.X5001-6.

[5] "Kay Picture Test Single Crowded Book- Paediatric Vision Test," Kay Pictures. Accessed: Aug. 25, 2022. [Online]. Available: https://kaypictures.co.uk/product/kay-picture-test-single-crowded-book/

[6] J. Atkinson, O. Braddick, and E. Pimm-Smith, "Preferential looking for monocular and binocular acuity testing of infants.," *Br. J. Ophthalmol.*, vol. 66, no. 4, pp. 264–268, Apr. 1982.

[7] "Lea-Test Ltd." Accessed: Aug. 25, 2022. [Online]. Available: http://www.lea-test.fi/index.html?start=en/vistests/instruct/leagrati/leagrati.html

[8] T. O. Adoh and J. M. Woodhouse, "The Cardiff acuity test used for measuring visual acuity development in toddlers," *Vision Res.*, vol. 34, no. 4, pp. 555–560, Feb. 1994, doi: 10.1016/0042-6989(94)90168-6.

[9] I. Livingstone *et al.*, "Testing Pediatric Acuity With an iPad: Validation of 'Peekaboo Vision' in Malawi and the UK," *Transl. Vis. Sci. Technol.*, vol. 8, no. 1, p. 8, Jan. 2019, doi: 10.1167/tvst.8.1.8.

[10] G. E. Quinn, J. A. Berlin, and M. James, "The Teller Acuity Card Procedure: Three Testers in a Clinical Setting," *Ophthalmology*, vol. 100, no. 4, pp. 488–494, Apr. 1993, doi: 10.1016/S0161-6420(93)31617-9.

[11] "Vision for Doing: Appendix 3." Accessed: Nov. 28, 2023. [Online]. Available: https://www.ssc.education.ed.ac.uk/resources/vi&multi/vfdh/vfdpt3aIII.html

[12] "Teller Acuity Cards TAC II reference and instruction manual." Precision Vision Inc., 2014.

[13] G. Vivas-Mateos, I. A. T. Livingstone, R. Hamilton, A. Cheema, and M. E. Giardini, "Too Many Shades of Grey: Photometrically and Spectrally Mismatched Targets and Backgrounds in Printed Acuity Tests for Infants and Young Children," *Transl. Vis. Sci. Technol.*, vol. 9, no. 12, p. 12, Nov. 2020, doi: 10.1167/tvst.9.12.12.

[14] Y.-Y. Qin *et al.*, "A computerized resolution visual acuity test in preschool and school age children," *Int. J. Ophthalmol.*, vol. 13, no. 2, pp. 284–291, Feb. 2020, doi: 10.18240/ijo.2020.02.13.

[15] N. Vrabič, B. Juroš, and M. Tekavčič Pompe, "Automated Visual Acuity Evaluation Based on Preferential Looking Technique and Controlled with Remote Eye Tracking," *Ophthalmic Res.*, vol. 64, no. 3, pp. 389–397, Oct. 2020, doi: 10.1159/000512395.

[16] S. Deepika, A. S. Nivetha, R. Harchana, and M. S. S. Devi, "Development of eye gaze tracking system for strabismus diagnosis," *AIP Conf. Proc.*, vol. 2405, no. 1, p. 020021, Apr. 2022, doi: 10.1063/5.0072823.

[17] V. Sturm, D. Cassel, and M. Eizenman, "Objective Estimation of Visual Acuity with Preferential Looking," *Invest. Ophthalmol. Vis. Sci.*, vol. 52, no. 2, pp. 708–713, Feb. 2011, doi: 10.1167/iovs.09-4911.

[18] P. R. Jones, S. Kalwarowsky, J. Atkinson, O. J. Braddick, and M. Nardini, "Automated Measurement of Resolution Acuity in Infants Using Remote Eye-Tracking," *Invest. Ophthalmol. Vis. Sci.*, vol. 55, no. 12, pp. 8102–8110, Dec. 2014, doi: 10.1167/iovs.14-15108.

[19] A. Hathibelagal, M. Eizenman, E. Irving, and S. Leat, "Visual fixation as an objective measure of visual acuity in infants," *Invest. Ophthalmol. Vis. Sci.*, vol. 54, no. 15, p. 1305, Jun. 2013.

[20] J. R. Drover, L. M. Wyatt, D. R. Stager, and E. E. Birch, "The Teller Acuity Cards Are Effective in Detecting Amblyopia," *Optom. Vis. Sci. Off. Publ. Am. Acad. Optom.*, vol. 86, no. 6, pp. 755–759, Jun. 2009, doi: 10.1097/OPX.0b013e3181a523a3.

[21] S. Painter, R. Hamilton, and I. Livingstone, "Diagnostic Accuracy of Online Visual Acuity Testing of Paediatric Patients," *Br. Ir. Orthopt. J.*, vol. 19, pp. 35–43, Apr. 2023, doi: 10.22599/bioj.292.

[22] "Validation of the Acuity Card Procedure for Assessment of Infants with Ocular Disorders - ScienceDirect." Accessed: Dec. 07, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0161642087333986?via%3Dihub

[23] N. Metropolis and S. Ulam, "The Monte Carlo Method," *J. Am. Stat. Assoc.*, vol. 44, no. 247, pp. 335–341, Sep. 1949, doi: 10.1080/01621459.1949.10483310.

[24] A. Boribalburephan, "Gaze Direction Classification for Digital Visual Acuity Tests," University of Strathclyde, Glasgow, Scotland, MAy 5th 20222.

[25] S. Eksborg, "Evaluation of method-comparison data," *Clin. Chem.*, vol. 27, no. 7, pp. 1311–1312, Jul. 1981.

[26] "The criteria for certification," RNIB. Accessed: Dec. 06, 2023. [Online]. Available: https://www.rnib.org.uk/your-eyes/navigating-sight-loss/registering-as-sight-impaired/the-criteria-for-certification/