Inter-Rater and Intra-Rater Reliability in the Interpretation of MTI Photoscreener Photographs of Native American Preschool Children

KATHLEEN M. MOHAN, MA, JOSEPH M. MILLER, MD, VELMA DOBSON, PhD, ERIN M. HARVEY, MA, and DUANE L. SHERRILL, PhD

Department of Ophthalmology (KMM, JMM, VD, EMH), The Optical Sciences Center (JMM), Department of Psychology (VD), and The Respiratory Sciences Center (DLS), University of Arizona, Tucson, Arizona

ABSTRACT: Purpose. To evaluate inter- and intra-rater reliability for the interpretation of MTI Photoscreener photographs taken in a population of Native American preschool children with a high prevalence of astigmatism. Methods. Photographs of 369 children were rated by 11 nonexpert and 3 expert raters. Photographs for each child were scored as pass, refer, or retake. Nonexpert raters scored photos on two separate occasions, permitting analysis of intra-rater reliability. Results. Analyses of pass/refer responses only: inter-rater reliability was moderate to substantial among nonexpert raters and substantial among expert raters. Intra-rater reliability among nonexperts was substantial. Analyses of all responses (pass, refer, and retake): inter-rater reliability for pass and refer scores was moderate among nonexperts and substantial among expert raters; for retake scores inter-rater reliability was slight for nonexperts and moderate for experts. Intra-rater reliability among nonexperts was substantial for pass and refer scores and moderate for retake scores. Conclusions. In this population with a high prevalence of astigmatism, whether MTI photoscreening results are interpretable is much more variable among and within raters than whether an interpretable photograph should be scored as pass or refer. The level of agreement among raters in the current study was influenced by the experience of the raters. In addition, nonexpert raters were more likely to deem a photograph uninterpretable than expert raters. (Optom Vis Sci 2000;77:473–482)

Key Words: photoscreening, children, inter-rater reliability, intra-rater reliability
TABLE 1
Inter-rater Reliability for Pass/Refer Results

<table>
<thead>
<tr>
<th>Study</th>
<th>Population</th>
<th>Prevalence of Amblyogenic Factors</th>
<th>Training and Experience</th>
<th>Flash Eccentricity</th>
<th>k Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freedman &amp; Preston</td>
<td>Pediatric ophthalmology patients</td>
<td>63%</td>
<td>Extensive</td>
<td>0.5 mm</td>
<td>0.73</td>
</tr>
<tr>
<td>Freedman &amp; Griffin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hatch et al.</td>
<td>Children enrolled in Massachusetts migrant summer education program</td>
<td>2.5–5%</td>
<td>Self-study</td>
<td>1.5 mm</td>
<td>0.38</td>
</tr>
<tr>
<td>Lewis &amp; Marsh-Tootle</td>
<td>African American children enrolled in Head Start</td>
<td>2.5–5%</td>
<td>3-hr seminar</td>
<td>1.5 mm</td>
<td>0.55</td>
</tr>
<tr>
<td>Present study—Nonexpert raters</td>
<td>Native American children enrolled in Head Start</td>
<td>44%</td>
<td>1-Day seminar</td>
<td>1.5 mm</td>
<td>0.60 Session I 0.62 Session II</td>
</tr>
<tr>
<td>Present study—Expert raters</td>
<td>Native American children enrolled in Head Start</td>
<td>44%</td>
<td>Extensive</td>
<td>1.5 mm</td>
<td>0.77</td>
</tr>
</tbody>
</table>

* Cited in Lewis & Marsh-Tootle.4

METHODS
Subjects

Subjects were 369 children who were enrolled in the Tohono O’Odham Head Start program during the 1997–1998 or the 1998–1999 school year. All were between the ages of 3 years, 6 months and 4 years, 11 months on September 1 of the year in which they were tested. All children were participants in the vision screening program sponsored jointly by the University of Arizona and the Tohono O’Odham Nation. An additional 147 children participated in the vision screening program but their data were not included in the present analyses because they had been tested during the previous school year (n = 93), were younger (n = 15) or older (n = 3) than the specified age range, were classified by Head Start as having “special needs” (n = 21), had ocular abnormalities other than refractive error (nystagmus, n = 1; iris coloboma, n = 1; lenticonus, n = 1), did not complete the vision screening protocol (n = 3), or could not be photographed because of MTI Photoscreener malfunction (n = 9). The Institutional Review Board of the University of Arizona approved the study, and written informed consent was obtained from the parents or legal guardians of all children before testing.

Equipment and Procedure

The MTI Photoscreener produces a composite photograph based on two images of the subject’s eyes. Both images are taken with the flash located 1.5 mm off-axis from the camera’s lens; however, for one image the flash is positioned on the vertical meridian, and for the other image the flash is positioned on the horizontal meridian. Interpretation of the composite photographs is based on pupil visibility and size, clarity of the fundus reflex, the size of bright pupillary crescents in the images of the pupils, and the location of corneal reflexes, both of which are produced by the off-axis flash, and any external signs of disease. A photograph is rated as pass (indicating that no abnormalities are apparent), refer...
Each child’s eyes were photographed with the MTI Photoscreener at the instrument’s working distance of 1 m, using black and white Polaroid film (ASA 3200). Samples of the photographs are shown in Fig. 1. All photographs were taken in a darkened room to maximize pupil size. If a photograph was judged by the tester (i.e., the person taking the photograph) to be uninterpretable (e.g., if the pupils were not visible in both images or the child was looking away from the camera), another photograph was taken. Although the study protocol suggested that the maximum number of photographs per child should be three, testers were permitted to take more photographs if they felt that these photographs would provide interpretable data for the child. One photograph was taken for 238 children (63%), two photographs were taken for 97 children (26%), three photographs were taken for 30 children (8%), four photographs were taken for 2 children (0.5%), and five photographs were taken for 2 children (0.5%).

Testers (i.e., those taking the photographs) were members of the AANAC team whose experience with photorefraction ranged from minimal to extensive. A single tester, who had no prior experience with photorefraction, took 68.3% of the photographs. A second tester, who had used photorefractors other than the MTI Photoscreener, took 17.3% of the photographs, and the rest of the photos (14.4%) were taken by other AANAC team members with no prior photorefraction experience. Testers learned to use the MTI Photoscreener by studying its manual and practicing with the instrument.

The MTI photographs were first scored by 11 nonexpert raters: a pediatric ophthalmologist (rater D), two research faculty members of ophthalmology (raters H and I), three research associates (raters C, F, and J), a certified ophthalmic technician (rater E), an optical sciences doctoral student (rater G), a registered nurse (rater A), an administrative assistant (rater K), and a research assistant (rater B). One of the research faculty members (rater H) had previous experience scoring isotropic photorefraction, and five raters (raters D, G, H, I, and J) had participated in a pilot project in which they scored 77 MTI photographs according to the criteria specified in the MTI Photoscreener manual. None of the other raters had any prior experience with photorefraction. Before scoring the photographs, the nonexpert raters participated in an 8-hour training session, provided by Prevent Blindness America (PBA), on the interpretation of MTI Photoscreener photographs. Based on all photographs taken of a child, each rater scored the child’s photoscreening results as pass, refer, or retake according to the criteria recommended by PBA (Table 2). According to these criteria, for a photograph to be interpretable (i.e., scored as a pass or refer), both pupils must be visible in both the top and bottom image in the photograph, both images must be in focus, pupil diameter must be between 4 and 8 mm, and the subject’s eyes must be properly fixated on the camera. There are two exceptions to this rule: first, if one of the two images in the photograph is interpretable and shows clear signs of a refractive error or an ocular abnormality, the photograph may be scored as refer rather than retake; second, if more than one photograph is available for a child, a top image that is interpretable in one photograph may be combined with a bottom image in another photograph to yield an interpretable result for the child.

To collect data on intra-rater and inter-rater reliability, the photographs were scored twice by each of the 11 nonexpert raters. Photographs taken in the fall of the 1997–1998 school year were scored in each of two sessions that took place in the summer of 1998, approximately 6 weeks (session I) and 12 weeks (session II) after training. The photographs taken in the fall of the 1998–1999 school year were scored in each of two sessions during December of 1998, with session II occurring 1 week after session I. Before each scoring session, the nonexpert raters were asked to review the PBA instruction manual on the interpretation of MTI photographs. After completing the second scoring of the first year’s photographs, the nonexpert raters were shown slides of and had the opportunity to briefly discuss 16 of the MTI photographs for which the score given by the majority of raters in the first session disagreed with the results of the ophthalmologic examination. The nonexpert raters received no other feedback or training.

The same MTI photographs were later scored by three expert...
TABLE 2.
Photoscreening failure criteria
(Pupils must be 4–8 mm in size)

<table>
<thead>
<tr>
<th>Nonexpert Raters</th>
<th>Expert Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strabismus</td>
<td>- Asymmetry in the corneal reflexes</td>
</tr>
<tr>
<td>Media opacity</td>
<td>- Media opacity &gt; 1.0 mm</td>
</tr>
<tr>
<td>Spherical refractive error and astigmatism (all definitions refer to the same eye as photographed in the two different pictures).</td>
<td></td>
</tr>
<tr>
<td>Myopia = crescent at the bottom of the pupil in one picture and at the right side of the pupil in the other picture. Astigmatism = (1) a crescent in one picture but not in the other picture of the same eye, (2) a crescent larger in one picture than in the other, or (3) mixed crescents, i.e., a crescent at the top of the pupil in one picture and at the right side of the pupil in the other picture or a bottom/left mix.</td>
<td>4 mm pupil diameter:</td>
</tr>
<tr>
<td>- Myopic or hyperopic crescent &gt; 2.0 mm</td>
<td>- Myopic or hyperopic crescent &gt; 2.5 mm</td>
</tr>
<tr>
<td>- Astigmatism (1) &amp; (2): crescent &gt; 2.0 mm</td>
<td>- Any astigmatism &gt; 1.5 mm</td>
</tr>
<tr>
<td>- Astigmatism (3): sum of crescents &gt; 2.0 mm</td>
<td>- Any astigmatism &gt; 2.0 mm</td>
</tr>
</tbody>
</table>

Anisometropia (all definitions refer to different eyes photographed in the same picture).
(1) Crescent in one eye but not in other eye. (2) Crescent larger in one eye than in the other eye. (3) Opposing crescents, i.e., crescent at top of pupil in one eye and bottom of pupil in other eye, or at left of pupil in one eye and right of pupil in other eye.

- Anisometropia (1) & (2): crescent > 2.0 mm
- Anisometropia (3): sum of crescents > 2.0 mm

4 mm pupil diameter:
- Anisometropia (1): crescent ≥ 2.0 mm
- Anisometropia (2) & (3): if ≥ 1.0 mm crescent in fellow eye then need 1.0 mm difference between eyes
- 6 mm pupil diameter:
- Anisometropia (1): crescent ≥ 3.5 mm
- Anisometropia (2) & (3): if ≤ 2.5 mm crescent in fellow eye then need 2.0 mm difference between eyes. If ≥ 3.0 mm crescent in fellow eye then need 1.0 mm difference between eyes.
- 8 mm pupil diameter
- Anisometropia (1): crescent ≥ 4.0 mm
- Anisometropia (2) & (3): if ≤ 3.5 mm crescent in fellow eye then need 2.0 mm difference between eyes. If ≥ 4.0 mm crescent in fellow eye then need 1.0 mm difference between eyes.

Raters from the Fundus Reading Center (Department of Ophthalmology and Visual Sciences, Vanderbilt University Medical Center, Nashville, TN) with extensive experience in the interpretation of MTI Photoscreener results. The Fundus Reading Center raters independently scored the photographs as pass, refer, or retake based on criteria developed in their reading center. The criteria used by the expert raters and those used by the nonexpert raters are shown in Table 2.

Data Analysis
Using the κ-coefficient,14–16 data collected from the nonexpert raters were submitted to analyses of inter-rater and intra-rater reliability; data from the expert raters were analyzed for inter-rater reliability only. A comparison of the overall nonexpert inter-rater and intra-rater reliabilities for year 1 vs. year 2 session I and year 1 vs. year 2 session II showed no significant differences between the two years. Therefore, data from session I for the first and second years of photoscreening were combined, as were data from session II for both years.

The κ-coefficient yields a measure of the level of agreement among multiple raters, or for one rater with multiple scorings, beyond what is expected by chance. A κ-value below 0.00 indicates poor agreement between raters (0.00 represents the level of agreement expected by chance), κ-values between 0.00 and 0.20 indicate slight
agreement, \( \kappa \)-values between 0.21 and 0.40 indicate fair agreement, \( \kappa \)-values between 0.41 and 0.60 indicate moderate agreement, \( \kappa \)-values between 0.61 and 0.80 indicate substantial agreement, and \( \kappa \)-values above 0.81 indicate almost perfect agreement.

\( \kappa \)-Coefficients for multiple raters were calculated in two ways. First, as in previous studies,\(^3,4,8,9\) agreement was calculated only for subjects whose data were scored as pass or refer. This analysis yielded a single \( \kappa \)-coefficient reflecting the combined level of agreement among raters. Second, agreement was calculated for all possible scores (pass, refer, and retake). The second analysis yielded separate \( \kappa \)-coefficients for pass, refer, and retake scores and a combined \( \kappa \)-coefficient, which is the weighted average of the separate \( \kappa \)-coefficients. The \( \kappa \)-coefficients for each type of score are calculated by combining the other two types of scores into one category (i.e., pass/not pass, refer/not refer, and retake/not retake). The separate \( \kappa \)-coefficients allowed us to examine the level of agreement among raters for each specific type of response. To get a clearer picture of the variability among raters, we also looked at the level of agreement between each inter-rater pair for each session, (e.g., for session I, we calculated how well rater A agreed with rater B, how well rater A agreed with rater C, and so on). We then calculated a weighted mean \( \kappa \)-coefficient for each rater and recorded the range of \( \kappa \)-coefficients for that rater based on his/her pairing with each of the other raters.

Intra-rater reliabilities were also calculated for both pass/refer and pass/refer/retake analyses. For intra-rater reliability, each rater’s scores from session I were paired with his/her scores for the same subject in session II. Weighted means of the individual \( \kappa \)-coefficients were calculated to obtain overall \( \kappa \)-coefficients for each type of analysis.

**RESULTS**

**Inter-rater Reliability**

**Pass/Refer Data Analysis: Nonexpert Raters.** The overall \( \kappa \)-coefficients (combined across years) were 0.60 for session I and 0.62 for session II. These values indicate that overall there was moderate to substantial agreement among raters when photographs scored as retake were excluded from the analysis.

The weighted means and range of the pairwise inter-rater \( \kappa \)-coefficients (based on a given rater’s pairing with each of the other raters) for each of the 11 nonexpert raters for sessions I and II are shown in Fig. 2. Weighted means were used because the number of subjects that are included in the \( \kappa \)-analysis varied across rater pairs (i.e., to be included in the calculation of the \( \kappa \)-coefficient a subject’s photograph had to be deemed interpretable by both raters in the pair). For individual pairs of raters, the pairwise \( \kappa \)-coefficients ranged from 0.29 to 0.79 [median = 0.61; semi-interquartile range (Q) = 0.09] for session I and from 0.35 to 0.83 (median = 0.63; Q = 0.11) for session II. Some pairs of raters showed only fair levels of agreement with each other and others showed almost perfect agreement.

**Pass/Refer Data Analysis: Expert Raters.** The overall \( \kappa \)-coefficient was 0.77, indicating substantial agreement among the three expert raters when retake scores were omitted from the analysis. For the individual pairs of raters, the pairwise \( \kappa \)-coefficients were 0.76, 0.77, and 0.79. The expert raters all showed substantial agreement with each other.

**Pass/Refer/Retake Data Analysis: Nonexpert Raters.** For session I, the combined \( \kappa \)-coefficient was \( \kappa = 0.39 \), and the separate \( \kappa \)-coefficients for each type of score were \( \kappa = 0.45 \) for pass scores, \( \kappa = 0.47 \) for refer scores, and \( \kappa = 0.15 \) for retake scores. For session II, the combined \( \kappa \)-coefficient was \( \kappa = 0.41 \), and the separate \( \kappa \)-coefficients were \( \kappa = 0.47 \) for pass scores, \( \kappa = 0.51 \) for refer scores, and \( \kappa = 0.16 \) for retake scores. These values indicate that there was only fair to moderate agreement between raters when all possible scores were included in the analysis. The separate \( \kappa \)-coefficients for each type of score show that when a photograph was deemed interpretable (i.e., pass or refer), there was moderate agreement among raters. However, there was only slight agreement among raters as to whether or not a photograph was interpretable.

**FIGURE 2.** Inter-rater reliability weighted mean \( \kappa \)-values for 11 nonexpert raters, based on photographs scored as pass or refer. Bars indicate the range of pairwise \( \kappa \)-values for each rater.
The weighted means and ranges of the pairwise inter-rater \(k\)-coefficients (based on a given rater’s pairing with each of the other raters) for each of the 11 nonexpert raters, for each type of score (pass, refer, or retake), and the level of agreement between pairs of raters for retake scores was much lower than that found for either pass or refer scores.

**Pass/Refer/Retake Data Analysis: Expert Raters.** The combined \(k\)-coefficient was 0.68, and the separate \(k\)-coefficients were \(k = 0.70\) for pass scores, \(k = 0.73\) for refer scores, and \(k = 0.44\) for retake scores. These values indicate that when retake responses were included in the analyses, there was substantial agreement among the expert raters for pass and refer scores, and a moderate level of agreement for retake scores.

**Intra-rater Reliability (Nonexpert Raters Only)**

**Pass/Refer Data Analysis.** The results of the Pass/Refer data analysis are shown in Table 3. The weighted mean of the intra-rater \(k\)-coefficients for the Pass/Refer analysis was 0.75 (SD = 0.08), indicating that, in general, when retakes were ignored, raters showed substantial internal consistency in their scoring of MTI photographs between sessions I and II. The individual \(k\)-values ranged from 0.59 to 0.88. Three of the raters showed almost perfect agreement for their pass and refer scores between the two sessions; seven showed substantial agreement and one showed moderate agreement.

**Pass/Refer/Retake Data Analysis.** As shown in Fig. 4, there was considerable variability in the distribution of pass, refer, and retake scores for each of the raters in each of the two sessions. The results of the Pass/Refer/Retake data analysis are shown in Table 3. The weighted means of the intra-rater reliability \(k\)-coefficients were 0.67 (range, 0.53–0.80; SD = 0.07) for pass scores, 0.66 (range, 0.22–0.83; SD = 0.15) for refer scores, and 0.46 (range, 0.15–0.64; SD = 0.16) for retake scores. Overall, raters showed substantial internal consistency between sessions for pass and refer scores and moderate consistencty for retake scores. For pass scores, nine raters showed substantial agreement with themselves and two showed moderate agreement with themselves across sessions. For refer scores, one rater showed agreement in the almost perfect range, seven showed substantial agreement, and one showed fair agreement across sessions. For retake scores, two raters showed substantial agreement, five showed moderate agreement, three showed fair agreement, and one showed slight agreement across sessions.

**DISCUSSION**

The purpose of developing vision screening tools is to permit rapid, cost-effective identification of persons who require more
extensive assessment by an eye care professional. To be cost-effective, a screening tool must be valid (i.e., it must accurately distinguish children who require more extensive assessment from those who do not), and it must be reliable (i.e., it must give the same result on different occasions and with different raters). The present study examined the reliability of interpretation of MTI Photoscreener photographs, both among (i.e., inter-rater reliability) and within (i.e., intra-rater reliability) raters, for a population of children with a high prevalence of astigmatism. The level of reliability, as measured with the \( \kappa \)-coefficient, was higher when the analysis included only MTI photographs scored as interpretable (i.e., pass or refer) than when the analysis also included MTI photographs scored as retake. In addition, the level of reliability was higher for raters with extensive experience in scoring MTI photographs than for raters with less extensive experience. This occurred despite the fact that the set of photoscreening failure criteria used by the expert raters (Table 2) might be expected to introduce a bias against higher inter-rater reliability for the expert raters. That is, the set of criteria used by the expert raters required more measurements and decisions than that used by the nonexpert raters; as the number of measurements and decisions that have to be made increases, the opportunity for error and disagreement also increases.

The majority of studies that have examined the reliability of interpretation of MTI photographs only assessed inter-rater reliability, and all of the previous studies only assessed reliability for photographs that were scored as pass or refer. A comparison of previously reported inter-rater reliability for pass/refer results with those of the present study is shown in Table 1.

Freedman and Preston reported the percentage of agreement between two raters for the interpretation of photographs of 202 patients from a private pediatric ophthalmology practice whose ages ranged from 5 months to 23 years. The prescreening probability of amblyogenic factors in this subject group was 63%. Photographs were taken with the Eyecor camera (patented by H. L. Freedman), which is an early prototype of the MTI Photoscreener. One difference between the Eyecor and the MTI Photoscreener is that the Eyecor has a flash eccentricity of 0.5 mm compared to the 1.5 mm flash eccentricity of the MTI Photoscreener. The distribution of failures of the ophthalmologic examination given to each child were 5% media opacities, 13% anisometropia, 14% myopia, 19% astigmatism, 24% hyperopia, 26% amblyopia, and 32% strabismus (several children exhibited more than one amblyogenic factor). Each rater scored the photographs independently. Although the authors did not report the professions, training, or previous experience of the raters either with the Eyecor camera or in the interpretation of photoscreening results, Lewis and Marsh-Tootle reported that the raters in the Freedman and Preston study had prior experience with the device in a clinical setting where they had access to both screening and diagnostic results, and more opportunity to discuss photographic interpretations prior to attempting their study. The results indicated that the two raters agreed with each other 94% of the time. No photographs were scored as retake and their agreement calculation did not control for chance. However, a personal communication cited in a subsequent report indicated that Freedman and Griffin later corrected for chance and obtained a \( \kappa = 0.73 \), indicating substantial agreement between the two raters. Hatch et al. evaluated the inter-rater reliability for the interpretation of MTI photographs of 71 children between the ages of 2 years, 9 months, and 10 years, 9 months, who were from migrant worker families (ethnicities listed in order of occurrence were Portuguese, Asian, Italian, Haitian, other Hispanic, white, and Native American). Based on data collected in prior studies, the expected prevalence of amblyogenic factors in this population was 3 to 5%. In contrast to Freedman and Preston, the expected prevalence of specific amblyogenic factors was not reported. The MTI results were interpreted by 12 raters (an optometrist, six optometry students, and five medical students) who were masked to the results of an ophthalmologic examination given to the children and also to the scores of the other raters. Before scoring the photographs, the raters studied the instructions for the MTI Photoscreener for 1 to 3 h, practiced with the photorefractor on adults,
and practiced interpretation of MTI photographs using criteria recommended by MTI. None of the raters had any previous experience in interpreting MTI Photoscreener results. Readable photographs were scored as either pass or refer and the inter-rater reliability analysis yielded a \( \kappa \)-coefficient of 0.38, indicating only fair agreement among the raters in their interpretation of MTI Photoscreener results.

Lewis and Marsh-Tootle\(^4\) reported inter-rater reliability for the interpretation of MTI photographs taken of 54 African-American children enrolled in a Head Start program. The children were between the ages of 3 and 5 years and had no known risk factors for ocular abnormalities. The expected prevalence of amblyogenic factors in a population with no known risk factors is approximately 2.5 to 5\%.\(^{17, 18}\) The expected prevalence of specific amblyogenic factors was not reported. Five masked raters (a pediatrician, a nurse, and three optometry students) interpreted the MTI results. Before rating the photographs, the raters were given MTI training manuals and attended a three-hour-training seminar conducted by a consultant from MTI. None of the raters had any previous experience in interpreting MTI Photoscreener results. The inter-rater reliability analysis (for photographs that were deemed interpretable) yielded a \( \kappa \)-coefficient of 0.55, indicating a moderate level of agreement between the raters in their interpretation of MTI Photoscreener results.

The nonexpert inter-rater reliability for pass/refer responses in the present study is higher than that found by Hatch and colleagues,\(^3\) similar to that reported by Lewis and Marsh-Tootle,\(^4\) and somewhat lower than that found by Freedman and Preston.\(^8, a\) The expert raters in the current study had an inter-rater reliability similar to that of Freedman and Preston.\(^8, a\)

Several factors may have contributed to the differences in inter-rater reliability found across studies. One factor is the prevalence of amblyogenic factors in the population screened. Among the studies summarized in Table 1, those with the highest prevalence of amblyogenic factors show the largest \( \kappa \)-coefficients. Other studies have suggested that the \( \kappa \) statistic may vary considerably with the true prevalence of a disorder in a population such that when the true prevalence of a disorder is very low (less than about 5\%) or very high (greater than about 95\%), the \( \kappa \) statistic may be much lower than when the true prevalence is in an intermediate range,\(^4, 15, 16\) (see Thompson and Walter\(^15\) and Kraemer and Bloch\(^16\) for the mathematical basis and discussion of the effect of true prevalence on the \( \kappa \) statistic).

A second factor that may have led to variability in the level of agreement across studies is the type of amblyogenic factor(s) present in the subject populations. It is possible that some types of amblyogenic factors are difficult to detect reliably. For example, Simons et al.,\(^9\) who looked at the inter-rater reliability for specific diagnoses in three pairs of raters, found that for two of the inter-rater pairs, agreement was greatest for the diagnosis of myopia with the next highest level of agreement for the diagnosis of astigmatism. The third inter-rater pair agreed most often on the diagnosis of hyperopia and the next highest level of agreement was for myopia. Although only three inter-rater pairs were examined in that study, the results may suggest that some amblyogenic factors are easier to recognize consistently. Data collected in the first year of the current study showed that 22% of the 250 subjects photoscreened in the first year of testing had at least 2 D of astigmatism.\(^9\) It may be that astigmatism is more consistently recognized than other types of amblyogenic factors and therefore led to higher levels of inter-rater agreement in the current study relative to studies that contain higher prevalences of other types of amblyogenic factors.

Inter-rater reliability may also be affected by the magnitude of the amblyogenic factor(s) present in a population. For example, if the magnitude of astigmatism present in the subjects of the current study resulted in bright crescents in the MTI photographs that were close in size to the criteria set for referral, one might expect more variability among raters due to ambiguity as to whether the referral criteria were met or not. On the other hand, if the magnitude of astigmatism tended to be either very high or very low, one would expect high agreement among raters because the bright crescents from the astigmatism would be clearly above or below the criteria set for referral. To see if there was a relationship between magnitude of amblyogenic factor and level of agreement among raters, we categorized the data from the current study into six categories according to magnitude of astigmatism and then calculated inter-rater \( \kappa \)-coefficients for each category for both nonexpert and expert raters. The results are shown in Fig. 5. Agreement among raters tended to be lowest for the category of 0.00 to 0.74 D of astigmatism for both nonexpert and expert raters. When retakes were excluded from the analysis, the expert raters showed the highest agreement for the categories of 0.75 to 1.49 D, and 3.75 D or greater of astigmatism. No other trends were evident in the data, suggesting that, at least for this population, the magnitude of the primary amblyogenic factor (i.e., astigmatism) did not have a significant impact on inter-rater reliability.

A fourth factor that may have contributed to differences in inter-rater reliability is the level of training and experience of the raters. Among the studies summarized in Table 1, higher \( \kappa \)-coefficients are associated with more extensive training and experience. It may be that more training and experience, especially if it includes the opportunity for feedback and discussion, leads to greater agreement among raters as to how specific criteria should be applied. For
example, although all the nonexpert raters in the current study were given the same criteria on bright crescent size for pass vs. refer scores, several of the nonexpert raters stated that it was difficult to decide where a bright crescent ended because some of the bright crescents seen in the photographs were very bright near the pupillary border and then gradually faded into the darkness of the pupil. The training by Prevent Blindness America was not specific as to what should be included in the bright crescent measurement and what should not. Therefore, different raters may have had different ideas about where a crescent ended, thereby lowering inter-rater agreement. The fact that no effect of experience (i.e., no difference in inter-rater or intra-rater reliability between year 1 and year 2) was found for the nonexpert raters in the current study does not rule out the possibility that experience plays a significant role in inter-rater and intra-rater agreement. However, it does suggest that the amount and type of experience the nonexpert raters received in this study was not enough to significantly improve reliability.

A fifth factor that may contribute to differences in inter-rater reliability among studies is flash eccentricity. In off-axis photorefraction, there is a negative correlation between flash eccentricity and crescent size: as the eccentricity of the flash from the central axis of the camera decreases, the size of the crescent in the image of the pupil that indicates refractive error increases. Because it is easier to detect a large crescent than a small one, this may affect the sensitivity and specificity of the interpretation of MTI photographs, especially if referrals are to be made for relatively small refractive errors. Thus, flash eccentricity may indirectly affect the level of inter-rater reliability through its influence on sensitivity and specificity (i.e., as the accuracy of individual raters goes up, their agreement will also go up). This could explain, in part, the high reliability found by Freedman and Preston, implying that flash eccentricity did not have a large impact on the level of inter-rater agreement in the current study.

A final factor that may contribute to differences in inter-rater reliability across studies is iris color. Lewis and Marsh-Tootle reported difficulties in detection of the pupillary border and the red reflex in MTI photographs of children with darkly pigmented irides. This factor may have reduced inter-rater reliability in the studies of Lewis and Marsh-Tootle and Hatch et al., and in the current study, because all included a substantial proportion of children with dark irides. However, the fact that the experts in the current study had a high level of inter-rater agreement, despite the fact that only two of the subjects in the current study had light-colored irides, suggests that if iris color is contributing to variation in inter-rater reliability, its influence can be attenuated with training and experience.

A unique aspect of the present study is that, in addition to examining reliability of interpretation for MTI photographs rated as pass or refer, we also conducted an analysis that included photographs scored as retake. This analysis provided data on raters’ reliability for scoring MTI photographs as pass vs. not pass (not pass = refer + retake), refer vs. not refer (not refer = pass + retake), and retake vs. not retake (not retake = pass + refer). The results indicate that when all three possible responses are included in the analysis, both the inter-rater reliability (Fig. 3) and the intra-rater reliability (Table 3) for pass and refer scores are lower than that found when only pass and refer scores are analyzed (Fig. 2 and Table 3). This is probably caused by the marked lack of agreement among raters concerning whether a photograph should be scored as retake. For nonexpert raters, the inter-rater reliability χ-coefficients for retake scores indicated an overall level of agreement (k = 0.16) that was only slightly above chance. For expert raters, the χ-coefficient for retake scores was better (k = 0.38), but still indicated only fair agreement. Tong et al. also found poor agreement among raters in deciding whether or not a photograph was interpretable. In that study, the raters agreed fully on the interpretability of only 27 out of 100 photographs.

Determination of whether a photograph is interpretable is based on several explicitly defined criteria, including visibility of the pupils, clarity of the picture, and on-axis fixation. Tong et al. speculated that the interpretability of a photograph might be influenced by the age of the subject, with photographs of younger subjects being less likely to be interpretable, and by the experience of the person taking the photograph, with greater experience leading to more interpretable photographs. However, they found that whether a photograph was deemed interpretable correlated with neither the age of the subject nor the point at which the photograph was taken (i.e., at the beginning, when the photographer had little experience, or at the end of the study, when the photographer had more experience). A regression analysis on the uninterpretable photographs in the current study yielded similar results, in that a photograph was equally likely to be scored as a retake whether it was taken near the beginning of the study or near the end.

Miller et al. recently investigated raters’ ability to judge on-axis fixation. Twelve trained raters examined 50 MTI photographs of 10 adults, 9 of whom had refractive error. Each adult was photographed five times; once while looking at the correct fixation point on the MTI Photoscreener and then while looking 5 cm (2.9°) and 10 cm (5.7°) to the left and to the right of the correct fixation point. The 50 photographs were presented to each rater in random order and the rater was required to judge whether or not the subject was looking at the fixation point. More than 50% of raters judged that the subject was correctly looking at the MTI Photoscreener fixation point when they were actually looking 5 cm (2.9°) to the left or right, and more than 25% of the raters made an incorrect judgment when the subject was looking 10 cm (5.7°) to the left or right. Miller et al. also found that the accuracy of the subject’s fixation affected the size of the bright pupillary crescents that appear in the photographs when refractive error is present, with crescent size becoming up to 3.9 mm larger or up to 1.9 mm smaller as the subject’s fixation moved off axis. These results indicate that it is difficult for raters to judge whether a photograph meets the direction of gaze criterion for interpretability. Furthermore, the results indicate that if a rater chooses to interpret a photograph in which the subject is fixating off-axis, changes in crescent size related to off-axis fixation can result in incorrect scoring (pass or refer) of the photograph.

In addition to including photographs scored as retake in the analysis, the present study differed from most previous studies by examining intra-rater as well as inter-rater reliability for nonexpert raters (see Simons et al. for an intra-rater reliability analysis for
specific diagnoses). Comparison of the $\kappa$-coefficients in Table 3 with those depicted in Figs. 2 and 3 indicate that intra-rater reliability was higher than inter-rater reliability in the present study. Overall, the intra-rater reliability values were in the substantial agreement range for pass and refer scores and in the moderate agreement range for retake scores, whereas inter-rater reliability values were in the moderate agreement range for pass and refer scores and in the slight agreement range for retake scores. This suggests that ambiguities in the interpretation of MTI Photoscreener photographs are less likely to cause a given rater to disagree with himself or herself across time than to cause disagreement among different raters.

In conclusion, the data of the present study suggest that the outcome of vision screening in a population with a high prevalence of astigmatism, conducted using the MTI Photoscreener, is highly dependent on the rater who scores the photographs. Raters with considerable experience tended to agree with one another in the scoring of interpretable photographs (pass or refer) (Table 1), whereas agreement among nonexpert raters was considerably more variable (Table 1 and Figs. 2 and 3). A review of previously published findings is consistent with the current results, in that raters with more experience had higher rates of agreement. However, it is possible that a comparison of inter-rater reliabilities across studies is not appropriate because several factors varied across studies, including the prevalence of amblyogenic factors and the sensitivity and specificity of the individual raters, which can have a large impact on the value of the obtained $\kappa$-coefficient. Therefore, considerable caution must be exercised when generalizing reliability results from specific studies to different populations and scoring situations.

ACKNOWLEDGMENTS

We thank the people of the Tohono O’Odham Nation and the Tohono O’Odham Head Start Program for their help in making this research possible. We would also like to thank Tom Leonard-Martin, Ph.D., M.P.H., Director of the Fundus Photo Reading Center, Vanderbilt University Medical Center, for providing inter-rater reliability data from three expert raters, Erin Siegel for her help with the statistical analyses, and the nonexpert raters who were not also coauthors: P. Broyles, S. M. Delaney, PhD, J. Funk-Weyant, BA, COA, H. Leising Hall, MS, C. Lopez, F. Lopez, RN, and J. T. Schweigerling, PhD. Finally, we would like to thank the anonymous referees for their helpful comments and suggestions. This work was supported by National Eye Institute grant U10-EY11155 to JMM.

These data were presented in part at the annual meeting of the Association for Research in Vision and Ophthalmology, Fort Lauderdale, Florida, May 11, 1999.

Received November 15, 1999; revision received February 8, 2000.

REFERENCES


K. M. Mohan

University of Arizona, Department of Ophthalmology

655 N. Alvernon Way, Suite 108

Tucson, AZ 85711-1824

email: kmohan@u.arizona.edu