On the Statistical Reliability of Letter-Chart Visual Acuity Measurements

Aries Arditi and Ron Cagenello

Purpose. To determine boundary values of test-retest reliability and sensitivity to acuity change that are unlikely to be exceeded in any clinical situation, for a popular visual acuity chart (Lighthouse/ETDRS) with three scoring methods; to discuss general methodological issues associated with statistical accuracy of optotype chart testing; and to link measures of test reliability to measures of sensitivity to change.

Methods. Five highly practiced subjects were tested using a computer-controlled acuity testing system in a procedure designed to reduce measurement error. Subjects read the computerized chart 156 times, yielding a sample of 78 test-retest comparisons.

Results. Under conditions likely to minimize variability, visual acuity may, with 95% confidence, be ascertained only within ±0.1 log units, using this chart with the recommended letter-by-letter scoring. Detecting a significant change in visual acuity requires about ±0.14 log units for the same degree of confidence.

Conclusions. These measurements may be viewed as approaching the upper limit of reliability of this letter chart. Reliability probably is considerably less in typical usage.

Visual acuity testing with optotype charts and projection systems is the most widely used test of visual function. It is used nearly universally in ophthalmology and optometry for screening, refraction, and monitoring of disease progression; in licensing for vision-intensive tasks such as driving and flying; and in occupational testing. Optotype acuity testing also is used in both basic and clinical vision research, as a way to characterize observers' visual resolving capacities.

Accuracy of visual acuity scores may be less critical for some applications, such as screening for gross deviations from normal acuity, or refraction, where the acuity chart is used more as a stimulus with which to compare potential lens corrections than as a formal assessment. However, accuracy of obtained acuity scores is important in clinical diagnosis, in monitoring change in disease progression or remission of the effectiveness of a particular intervention, and for assessing relative performance under different test conditions. In addition, the use of visual acuity scores in driver's licensing and occupational testing may have a strong impact on an individual's quality of life or vocational options.

Snellen is generally credited with inventing letter-chart acuity testing, and although many variants of his original test exist, most aspects of the test have withstood the test of time, retaining nearly universal clinical and public acceptance. Although we may place a special trust in a test that has been with us for some 130 years, the statistical accuracy of optotype acuity has received only cursory study, appearing as an adjunct in just a few studies. Because recent years have

From The Vision Research Laboratory, The Lighthouse Inc., New York, New York.
Supported in part by National Aeronautics and Space Administration cooperative agreement NASA-AMES NCC2-541.
Proprietary interest category: N.
Submitted for publication: March 10, 1992; accepted July 6, 1992.
Reprint requests: Aries Arditi, Vision Research Laboratory, The Lighthouse Inc., 800 Second Avenue, New York, NY 10017.

Copyright © Association for Research in Vision and Ophthalmology
seen rapid improvement in and specialized variants to the basic Snellen design, it seems especially worthwhile to examine the statistical accuracy of such tests. In particular, we are concerned with their test-retest reliability—that is, the correlation of the test with repeated measurements.

However, there are problems in the design and typical use of these tests that hinder the proper assessment of their test-retest reliability. First, the smallest line available at the recommended test distance generally corresponds to −0.3 log MAR (20/10) acuity. Although adequate for most clinical purposes, this limit can truncate the distribution of obtained visual acuity scores, distorting its true form.

Second, the a priori probability of correctly guessing the identity of a test letter is ambiguously specified, because the subject may select responses from among the 26 letters of the alphabet, even though only a subset (10 for the test examined in this report) are represented in the test. Assumption of an a priori probability of 1/26, even for naive subjects, will ignore the possibility that the subject is cognizant of or learns of the reduction in optotype set size, perhaps even within a single reading of the chart. The assumption that the guessing probability is 1/10 similarly leads to error, because the subject is generally allowed to make responses, in error, of letters that fall outside the optotype set. This indeterminacy of the guessing rate adds statistical noise to the acuity measurement as a whole and will tend to reduce the reliability of the test. More important, it makes within-subject test-retest comparisons awkward to interpret.

Third, conditions generally vary widely from examiner to examiner in test scoring procedure. Most examiners score the test in whole line increments; others adopt the letter-by-letter procedure recommended by Ferris et al3 that has been shown to provide a more accurate index of acuity by Bailey et al.4 Anecdotally, many clinicians believe they can obtain a higher and more accurate acuity score by “pushing” the patient to respond to smaller letter lines than they would otherwise respond to. Yet another variant, preferable in our view, is to require the patient to read the entire chart with letter-by-letter scoring to avoid penalizing those who choose not to guess.5 Such variations in procedure make any test-retest analysis less meaningful, because it is likely to apply only to the specific test conditions adopted in the analysis.

In light of these problems and variations in test conditions, such as chart luminance, test distance, and other sources of statistical error, assessments of test-retest reliability are nearly always open to criticism in that they do not accurately reflect common test usage. Indeed, attempts to typify actual usage and departures from the same may be doomed to failure, given this wide variability.

In the present study, we have taken a different approach. Instead of attempting to evaluate the acuity test under typical conditions, we have assessed what we consider to be the highest test-retest reliability that can be obtained—of a popular chart developed by Ferris et al.5 for the Early Treatment Diabetic Retinopathy Study (ETDRS) and now marketed as The Lighthouse Distance Acuity Chart. Using carefully controlled laboratory conditions that would be expected to minimize variability, and psychophysically sophisticated, practiced subjects, our assessment is intended to estimate what is surely a less variable upper bound to acuity test accuracy.

MATERIALS AND METHODS

Optotypes were Sloan characters displayed on a 15" Mitsubishi (Torrance, CA) color monitor driven by an X-windows (X11) server residing on a Silicon Graphics (Mountain View, CA) IRIS 4D/25 graphic workstation via a StereoGraphics Corporation (San Rafael, CA) stereo display system, with a final pixel resolution of 1280 horizontal X 512 vertical. The stereo display system, which is responsible for the unusually high ratio of horizontal to vertical resolution, allows us to present separate stimuli to each eye and is used in experiments other than those described here. Client software that actually ran the experiments, via a local network, resided on a Sun Microsystems (Mountain View, CA) SparcStation 1.

To reduce the size of the letters at the eye, the display was viewed at a distance of 95 cm through the objective lenses (ie, in reverse) of a pair of Minolta (Ramsey, NJ) 10 × 25 wide-angle pocket binoculars, yielding an optical distance of 950 cm. The luminance of the background of the display was 88.3 cd/m². We measured the transmittance of the binoculars on an optical bench to be 65%. Thus, the luminance of the background through the binoculars was 0.14 log units lower than the level typically recommended for acuity testing.6,9 The binoculars were mounted on a clamp that rested on the table holding a headrest to stabilize the subject’s head. They had adjustable interocular distance, and the image was carefully centered for each eye through its sighting tube at the start of each session.

We intended to mimic as closely as possible the Lighthouse/ETDRS chart, which incorporates ac-
cepted principles established by Bailey and Lovie, and to adhere as closely as possible to accepted standards for visual acuity testing. Letters were displayed in lines of five characters, separated from each other and from the edge of the screen by one character width, as shown in Figure 1.

The Lighthouse/ETDRS chart uses lines of letter combinations determined to be of equal confusability and not containing words or acronyms. With repeated measurement, subjects could, in this case, memorize letter sequences, so in our procedure the computer generated sequences of 5 of the 10 Sloan letters randomly without repeats for each line. (The additional variability resulting from randomly selected letters is discussed at length in the Discussion.) We used an algorithm that ensured letters were displayed with roughly equal frequency in the set of lines that made up a single reading of the chart.

Letters were generated using the font design language METAFONT, which facilitates translation of geometric definitions of type into pixel representation. The advantage of using METAFONT is that a single program can be used to generate optotypes of any size. Our optotypes were intended to vary in size by 0.05 log units between “lines,” as compared with the 0.1 log unit step size used by the charts emulated. Digitization error caused sizes to depart slightly from their intended values, of course. However, for all letters, the size along the (vertical) dimension of lowest resolution differed from its intended value by no more than 3%, and in most cases by no more than 1.5%. The smallest letters were 40 X 20 pixels, and the smallest letters corresponding to the highest obtained visual acuity score subtended 50 X 25 pixels.

Five highly practiced subjects, including the authors, were tested repeatedly over several weeks. The tenets of the Declaration of Helsinki were followed, and informed consent was obtained. Two of the subjects were emmetropic, two were slightly myopic, and one was slightly hyperopic. The myopic subjects wore their corrective lenses. All subjects were aware of the identity and number of characters in the Sloan set (C, D, H, K, N, O, R, S, V, and Z).

Subjects identified letters verbally to the experimenter, who typed responses into the computer. Subjects were allowed to make responses only within the Sloan set and were not allowed to repeat letters in their response for a single line. Tests were scored and data were analyzed by computer.

Acuity initially was scored by the method recommended by Ferris et al, wherein the subject reads the entire “chart” of 15 lines and is given 0.01 log MAR units for each letter read correctly. The log MAR acuity score for a single trial, then, was $0.25 - 0.01T_c$, where $T_c$ is the total number of letters read correctly of the 75 in the test. The 0.25 constant is a bias term that is the acuity score obtained if none of the letters on the chart is correctly identified. Our smallest letters corresponded to −0.5 log MAR (20/6.3 Snellen). Data subsequently were rescored in two additional ways.

Although data were collected from only five practiced observers with normal and supranormal acuities, the range of acuity scores was artificially increased by testing monocularly and binocularly at luminance contrasts ranging from 0.109 to 0.994, where contrast between foreground $L_{fg}$ and background $L_{bg}$ was defined as $L_{bg} - L_{fg}$.

The variability introduced by these manipulations helps prevent the well-known sampling problems associated with restriction of range in correlational analysis and does not artificially increase the kind of variability (measurement error) we are assessing with reliability analysis.

RESULTS

Figure 2A shows scatterplots of acuity scores of even-numbered repeats against odd-numbered repeats of tests within the same test condition (eg, tested eye, contrast), using size steps of 0.05 log units between lines. Each data point thus represents two entire readings of the chart, and the scatterplot therefore represents 156 complete readings of the chart. Test-retest reliability as measured by the correlation between tests was 0.895. Using Fisher’s $z$ transformation, if the population reliability coefficient is equal to our observed value, samples of our sample size (n = 78) should yield correlation coefficients within the range 0.84–0.93 95% of the time. The standard error of measurement ($\sigma_{meas}$), which reflects observed variability inherent in the test, was 0.0324 log MAR and is itself accurate to

**FIGURE 1.** A line of five Sloan characters, similar to that used in the Lighthouse/ETDRS chart and in the experiment.
about ±0.006 log MAR, taking into account the statistical uncertainty of the reliability coefficient itself. To the extent that our data accurately reflect the underlying variance and covariance of test and retest and that the true distribution of acuity scores is normal, we conclude that our test is accurate to within about ±0.06 log MAR 95% of the time.

Because our size steps between lines were one-half those of the Lighthouse/ETDRS chart, we rescored the data using only the eight lines corresponding to 0.1 log MAR steps. Acuity in log MAR units in this case is 0.30–0.02Tc. Data are shown in Figure 2B and are similar to but somewhat more scattered than with the 0.05 log unit size increments. In this case, the σ_meas is 0.044 ± ~0.009, with 95% confidence limits around acuity scores equal to ±0.09, nearly a line of the chart. This is the scoring condition that corresponds most closely to the actual chart and its recommended use, in that the size increment is 0.1 log MAR and the scoring procedure corresponds to that described by Ferris et al., the developers of the chart.

Finally, we rescored the data using the 0.1 log MAR increments, but adopting the line scoring procedure used by many clinicians and other examiners in their use of the chart. Visual acuity was taken to be the smallest line at which at least 3 of 5 letters were correctly identified. The data, shown in Figure 2C, show considerably higher variability, with an σ_meas of
0.067, itself accurate only to about ±0.01 and a 95% confidence interval about the true acuity score of ±0.13 log MAR.

The 95% confidence intervals about our reliability coefficients computed from Fisher's $z'$ transform, about our $\sigma_{\text{meas}}$, and about the underlying true visual acuity score estimated from $\sigma_{\text{meas}}$ assume normality of the underlying distribution. Figure 3 plots quantiles of the visual acuity scores obtained with 0.05 log MAR size increments and letter-by-letter scoring, against quantiles of the normal distribution. The data fall close to a straight line, providing support for this assumption.

In addition, visual inspection of Figure 2A indicates that variance was uniform over the range of acuities tested.

**DISCUSSION**

Previous attempts to assess test-retest reliability of optotype acuity measures have been few and have not used correlation coefficients as an index of reliability. With the suggestion of Bland and Altman\(^{11}\) that correlation coefficients were inappropriate measures of repeatability in clinical measurements, two such studies\(^{4,5}\) chose to use the coefficient of repeatability as a reliability index. This statistic is defined as 1.96 times the standard deviation of the distribution of differences between test and retest. It corresponds to the 95% confidence interval around discrepancies between the tests. It is believed by some investigators\(^{5,11,12}\) that this statistic is superior to the traditional reliability coefficient, because the reliability coefficient is sensitive to the range of values obtained.

We disagree with this rejection of classical psychometric theory for two reasons. First, although it has long been established that reliability coefficients alone are not and should not be compared across different ranges of test values,\(^{13,14}\) it nevertheless is a useful statistic for comparing reliability among tests that do share the same range. This is clearly reflected in the three scoring conditions of our test-retest data, whose reliability coefficients range from 0.68–0.90, whereas the standard deviations of the data range only from 0.10–0.12.

Second, the reliability coefficient is required to compute a statistic, $\sigma_{\text{meas}}$, that does allow useful comparison of test precision across different ranges of scores. The $\sigma_{\text{meas}}$ estimates the standard deviation of scores obtained on an infinite number of administrations of the test.\(^{14}\)

Given the similarity of assumptions that underlie the coefficient of repeatability statistic and the standard correlation model of reliability, we sought to link the two models mathematically. We show in the Appendix that the standard deviation of the distribution of discrepancies between test and retest ($\sigma_{\text{meas}}$) is equal to the correlation model's $\sigma_{\text{meas}}$ times $\sqrt{2}$. Thus, the traditional correlation model and the normal discrepancy distribution models are closely related. Exceeding the 95% discrepancy distribution confidence interval represents a high degree of confidence that a score will depart from that of a previous testing. Exceeding the 95% measurement error distribution confidence interval, on the other hand, represents a high degree of confidence that an obtained score will depart from the true acuity score. The former is most useful in detecting that change has occurred, whereas the latter is most useful in estimating the accuracy of the score.

Bailey et al\(^4\) used the distribution of differences between test and retest as an index of test precision. They developed a model of repeatability that uses the concordance rate, or frequency of observed zero discrepancies between test and retest, as a statistic for estimating the true standard deviation of the underlying distribution of differences. Because any real set of test-retest data contains an entire sample distribution of such differences from which the standard deviation of differences can be estimated more directly and precisely, it is not clear what purpose the concordance
FIGURE 4. Effect of scale coarseness on visual acuity, showing assumed error distributions sampled by observable scale values. Left panels are distributions of measurement error; right panels are distributions of measurement discrepancies. The shaded regions show the 95% confidence region for accuracy (left panels) and for detection of change (right panels).
statistic might serve as an index of repeatability, although Bailey et al found it useful in demonstrating that high rates of exact agreement between test and retest indicate that a test can benefit from the use of a finer scale. Indeed, that the concordance estimator uses such a small part of information available in the sample suggests a problem with its statistical efficiency.

In Bailey et al's study, they also assessed repeatability of optotype acuity, using two versions of each of three different acuity tests with different optotypes (British letters, Sloan letters, and Landolt rings) by examining the distribution of test and retest differences among repeated administrations of the tests. Using estimates of the standard deviation of this distribution based on their reported concordance rates and statistical model, and using the result proven in the Appendix of the present report, we can estimate the standard error of measurement of the underlying error distribution by dividing by √2. The estimated standard error of measurement is 0.028 log MAR for letter-by-letter scoring of the chart with standard 0.1 log MAR size increments, and 0.035 for the line scored chart. These values indicate accuracy of optotype acuity some 60–90% greater than we found. However, they used the concordance rate to estimate the standard deviation of that distribution, rather than estimating it from the sample standard deviation or from √2σ_{mean}. Our concern is that their estimate of discrepancy variability itself was unreliable and limited by the low statistical efficiency of their model.

Another aspect of their methodology can help explain the discrepancy between our results and theirs. Although it is not explicitly stated in their report, they used the same set of two charts for test and retest at three different distances for each optotype. This possibly increased accuracy of the test artificially by introducing memorization effects.†

Although not an explanation for the discrepancy between Bailey et al’s results and ours, another problem with their estimates of test-retest reliability is that they pooled data from three different charts. Aside from the ambiguity of defining just what is the test whose reliability is being evaluated, there are statistical reasons to believe that both acuity and reliability should differ at least between the 4-alternative Landolt test and the 10-(or 26) alternative Sloan and British letter tests they used. More letters should be read correctly on the basis of chance on the 4-alternative test than on the 10/26-alternative test. Thus, without a correction for guessing, acuity scores should be higher. Even with the conservative assumption of 10 alternatives for the Sloan/British chart, the Landolt test should lead to a (0.25/0.1) • 0.02 = 0.05 log MAR better performance on chance factors alone for a chart scored with 0.02 log MAR per correctly identified letter.

Thus, although we have no complete explanation for the discrepancy between our results and those of Bailey et al, low statistical efficiency of their concordance model and the possibility of memorization artifacts could have played significant roles. It should be noted that the main point of the Bailey et al report was to determine how scale coarseness affects detection of change in clinical tests; visual acuity was one of two didactic examples.

Figure 4 helps visualize the effects of scale coarseness within the paradigm of our test reliability analysis. The left column of graphs shows the estimated underlying normal measurement error distributions with standard deviations equal to our estimated σ_{mean} for each of the three scoring conditions we used and sampled by the possible scores that can occur in the test under those scoring conditions. It is easy to see that increasing scale coarseness adds measurement error within this range.

The right column shows sampled normal distributions of discrepancies for the same scoring conditions. The underlying normal distributions here have standard deviations equal to σ_{mean} (which estimates the underlying standard deviation of discrepancies) for each condition.

* Using Monte Carlo methods, we assessed the relative efficiency of three estimators of the true standard deviation of discrepancies: the sample standard deviation of discrepancies (σ_{mean}), the standard deviation of discrepancies computed from the concordance rate using the Bailey et al model, and √2σ_{mean}. Our data were sets of 1000 pairs of samples of known correlation from a normal distribution.15 Relative efficiency of two estimators is the ratio of their variances about the population parameter to be estimated.15 As might be expected, the sample standard deviation of the discrepancy distribution was the most efficient estimator. Although the relative efficiency depended on the correlation coefficient and the scale increment size relative to the true standard deviation of discrepancies (Bailey et al's R), we found the estimator based on concordance to be roughly 500–2000% worse than σ_{mean}, in estimating true underlying standard deviation of discrepancies, for reliability coefficients ranging from 0.75–0.95, and for R = 0.4—the value Bailey et al had empirically determined. √2σ_{mean} was only about 5–20% worse over the same range. For the 0.829 correlation coefficient we observed in the condition that corresponded most closely with Bailey et al (0.1 log MAR increment with letter-by-letter scoring), the concordance estimator was about 1300% less efficient than σ_{mean}. deviation. Ironically, as the scale increments get finer and finer, the concordance estimator becomes less and less efficient. For increments 1/10, the true standard deviation of discrepancies, it was about 2000–8000% worse than σ_{mean}, whereas √2σ_{mean} remained only about 5–20% worse, over the same range of reliability coefficients.

† In our original reading of the report, we had assumed that Bailey et al had allowed pairings of test with each optotype with retest using a different optotype. We believe this is a valid alternative interpretation of their published methodology. Such a procedure would have the benefit of obviating the potential effects of memorization, but would lead to deflated estimates of concordance if there are constant errors among the charts that are not exact integer multiples of the scale increment. However, this apparently is not the methodology they used (T. Raasch, personal communication).
The dark-shaded regions show the 95% confidence region to be exceeded for accuracy of the test in the left panels, and for detection of change in the right panels. Bailey et al. observed that as a practical matter, one must use as confidence limits the acuity scores that fall on an integer number of scale increments and that lie just beyond the shaded region. The coarseness of the scale, of course, adds uncertainty to the location of these “usable” confidence limit values and forces one to adjust the confidence limit to a probability that happens to fall on an increment boundary.

There is a potential pitfall, however, in simply using as a confidence limit the scale increment value that falls just beyond the standard deviation of the discrepancy distribution, when the concern is detecting the significance of a change. To illustrate, consider the middle row of graphs in Figure 4, which show 95% confidence limits for the Lighthouse/ETDRS chart with the recommended scoring. In the right graph, there is a scale increment (±0.12) that falls just beyond the 95% confidence limit of ±0.119. However, because we know that the standard deviation of discrepancies is equal to $\sqrt{\frac{2}{\sigma_{\text{meas}}^2}}$, we also could use ±1.96$\sqrt{\frac{2}{\sigma_{\text{meas}}^2}}$ = ±0.122 as another estimator for constructing our confidence interval. This latter value falls just beyond the scale increment value of 0.12 and suggests that we use the next increment, ±0.14, as our criterion value for the 95% limit. This option, of course, is the more conservative one and ensures that the 95% boundary is exceeded.

Elliott and Sheridan also analyzed discrepancies between test and retest of the Lighthouse/ETDRS chart, obtaining a $\sigma_{\text{meas}}$ of 0.107 for pairs of letter-by-letter-scored visual acuities obtained from 20 subjects with no refractive correction. Transforming to $\sigma_{\text{meas}}$ by dividing by $\sqrt{2}$ gives 0.076. Computing $\sigma_{\text{meas}}$ directly from their published correlation coefficients and standard deviations yields 0.078, which agrees nearly perfectly with the transformed $\sigma_{\text{meas}}$.

It is notable that their $\sigma_{\text{meas}}$ is nearly twice the value that we observed under similar scoring conditions. There are many possible reasons for this. First, their sample size is small. Second, their procedure included many of the features we specifically tried to avoid, such as testing at the recommended 4 m, which may have resulted in some truncation of the distribution of scores; requiring subjects to read only down to the line at which they made no correct responses; and presumably allowing subjects to make letter responses outside the Sloan set, etc. Third, they used naive subjects whose mean age was 64 yr. Finally, much of the variability in their scores is the result of refractive error, whereas in our study most of the variability came from our manipulation of stimulus contrast.

However, it is possible that the much lower statistical reliability they obtained more realistically reflects the accuracy of the test as typically used in the clinic than do our measurements, because like clinicians, they controlled for fewer variables that would add measurement error than we did. Given the wide variation in clinical usage of acuity tests, however, it ultimately may be impossible to characterize the “typical” acuity assessment. In fact, their measurements may be, like ours, far better controlled than most actual clinical measurements and thus fail to reflect clinical realism. Our approach, on the other hand, estimates reliability that is unlikely to be exceeded in any clinical situation. Even with such controls, accuracy of acuity using this test probably is worse than many clinicians and researchers would expect.

Our experiment, of course, departs from typical conditions of acuity testing in several significant ways. Our subjects were a small, select set of highly motivated and practiced observers who were familiar with the set of stimulus letters and other constraints on test responses and who had very high visual acuities. Our testing used formal protocols for scoring that often are passed over for the sake of brevity by vision practitioners. Our test luminance was somewhat lower than typical and slightly below the minimum recommended by commonly accepted standards. However, all of these departures from typical testing conditions, possibly except for lower luminance and our use of lines composed of randomly chosen letters, can be reasonably expected to decrease measurement error and increase test accuracy. The high visual acuities shown by our subjects would argue against our slightly lower luminance having but a negligible effect.

The lines being composed of randomly chosen letters, however, might be considered a significant additional source of variability that could reduce, rather than enhance accuracy of the test. Ferris et al specifically selected lines for the charts that made their overall identifiability highly uniform, having an average probability of correct identification ranging from 79.16–84.4%, based on the data of Sloan et al. Selecting letters randomly produces average probabilities ranging from 74.92% to 89.16%.

Figure 5 is a histogram of average probability of correct identification for all possible combinations of 5 of the 10 Sloan letters based on these same data from the Sloan et al report. We estimated the potential increase in variability introduced by randomly selecting letters as follows. First, we assumed that the variance introduced into the acuity data by variability in which particular line is chosen is equal to the square of the standard deviation of the histogram in Figure 5. This standard deviation is 0.0276 and is in units of probability. This value corresponds to acuity variation of 0.0069 log MAR units for the finest scoring condition we used and 0.0138 log MAR units for the other two scoring conditions. Thus, the variance in acuity ex-
Assuming the variability introduced by line letter composition is independent of other factors that affect our visual acuity scores allows us to subtract this variance from our acuity score variance to obtain the variance expected on the basis of no variation in line letter composition. After subtracting out the increased variance associated with letter composition of the lines in this way, we can estimate corrected standard error of measurement for our three conditions (finest to coarsest scoring) of 0.0323, 0.0436, and 0.0672. These values, which represent the measurement error expected if the variance attributable to variations in line composition were removed entirely, are virtually identical to our original standard error of measurement and suggest that chart designers need not be particularly concerned with differences in overall identifiability of lines that are based on differences in identifiability of the letters in the Sloan set.

There is one additional difference in our reliability assessment from what often is considered ideal in test evaluation. Rather than comparing test and retest scores over a wide range of the test, our measurements are confined to a 0.45 log unit range of acuities. The chart test itself, however, varies over the range of some 1.3 log units. One might argue that our assessment is subject to problems associated with restriction of range and that we might have obtained higher reliability coefficients had we tested acuity over a broader range. On the other hand, the use of the correlation coefficient in reliability (and discrepancy) analysis requires the assumption of uniform variance throughout the range of scores. We believe this assumption probably is unjustified over the full range of visual acuities represented on most visual acuity charts, given the fact that best-corrected acuities corresponding to the upper lines of the chart generally indicate conditions more severe than common refractive error. That is, while uncorrected refractive error may result in very poor visual acuity, the acuity value obtained in this way is used only to guide the practitioner to a better refraction. Poor best-corrected visual acuities, on the other hand, are associated with significant visual impairment, which is widely believed to be associated with increased variability in visual acuity measurements.

The range of our measurements is as large as the range of “normal” best-corrected visual acuities with any reasonable criterion of normal acuity; hence we believe our measurements reflect the best accuracy obtainable with visually normal or supranormal subjects and the tests and procedures described above. By modifying standard procedures, measurement error can be further decreased. For example, we found staircase methods to provide far better reliability in acuity measurements, using the apparatus and optotypes described in the present report, and we use them in experiments where short test duration is not critical. Another way of decreasing measurement error, of course, is to use letter size increments smaller than 0.1 log units, as we have shown in the present study. Unfortunately, charts with smaller increments are not commercially available.

In summary, we believe our measurements may be usefully viewed as approaching the upper limit of reliability of the Lighthouse/ETDRS chart, achievable when the strictest testing conditions are met. Their importance, however, is not necessarily to demonstrate that the reliability of this rapid and conveniently administered test can be improved (which it obviously can), but rather to provide clinicians, researchers, and others who use the test with boundary specifications for its accuracy.

APPENDIX

The following demonstrates that the standard deviation of the distribution of discrepancies between test and retest is equal to the standard error of measurement times \( \sqrt{2} \).

Let \( \sigma^2_t \) and \( \sigma^2_r \) be the variances of the test and retest scores, respectively, and \( \sigma_{\text{meas}} = \sigma \sqrt{1 - r_{tr}} \) be the standard error of measurement, where \( r_{tr} \) is the product-moment correlation coefficient between test and retest. The variance of the discrepancy distribution is

\[
\sigma^2_{t-r} = \sigma^2_t + \sigma^2_r - 2r_{tr}\sigma_t\sigma_r
\]

By assumption, however, \( \sigma^2_t = \sigma^2_r \) so

\[
\sigma^2_{t-r} = 2\sigma^2_t(1 - r_{tr})
\]
Reliability of Visual Acuity Measurements

and

\[ \sigma_{\text{err}} = \sqrt{2} \sigma_1 \sqrt{1 - r_{\text{tr}}} = \sqrt{2} \sigma_{\text{meas}}. \]

Key Words

acuity, letter-chart, optotype, reliability, vision.

Acknowledgments

We thank Kenneth Knoblauch for helpful discussion and comment on this report and for suggesting the Monte Carlo test of statistical efficiency.

References