JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, March 2011, Vol. 56, No. 2
doi: 10.1111/j.1556-4029.2010.01634.x
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

Elizabeth J. Garvin,[1] *B.S. and Robert D. Koons,*[1,**] *Ph.D.*

Evaluation of Match Criteria Used for the Comparison of Refractive Index of Glass Fragments*,†,‡,§

ABSTRACT: For comparative glass examinations, the refractive indices (RIs) of recovered glass fragments are often compared to a test interval defined by measurements from a broken glass object. RI measurements from five modern float glasses were used via resampling to assess the frequencies of false exclusion errors for eight test criteria as functions of the number of measurements. The test criteria were based on ranges, fixed intervals, and multiples of standard deviations of the known source measurements. The observed error rates for the eight tests studied are between zero and *c.* 35%, depending upon the match criteria, the number of measurements, and the RI distribution for a glass source. The results of this study can be used to predict the false exclusion rate for a test criterion under a given set of conditions or to select test criteria that result in a desired error rate for these typical sheet glasses.

KEYWORDS: forensic science, criminalistics, trace evidence, glass, refractive index, statistical methodology

Comparison of measured refractive indices (RIs) of recovered glass fragments with representative samples of a broken glass object associated with a crime is a well-established forensic trace evidence examination. Most examiners use one of two approaches, either a two-stage approach separating the comparison from a subsequent evaluation of the significance of any matching results or a single-stage probabilistic approach that compares the relative likelihoods of the evidence under two competing hypotheses. The majority of forensic glass examiners in the U.S. currently use some form of a two-stage approach, sometimes referred to as the frequentist approach. The probabilistic approach exemplified by a Bayesian analysis or calculation of a likelihood ratio has been the subject of much recent discussion and is currently used in a number of forensic laboratories primarily in Europe and New Zealand. Curran et al. (1) provide descriptions, historical backgrounds, and their opinions concerning the relative merits of the two approaches as they apply to glass examinations.

This paper reports the results of studies designed to estimate the false exclusion rates (Type 1 errors) obtained when using several of the match criteria currently used in the comparison stage of the two-stage approach. The basis of this approach is testing of the hypothesis of equality of the measurement data from the recovered glass fragments and broken object populations. Essentially, nearly every method used for this approach to glass comparison begins with defining a test interval encompassing measurements from the broken glass object (typically referred to as the known or control sample). We will use the term test interval to characterize the distribution of measurements in the known glass source rather than other terms that are commonly used, such as a confidence interval, because we do not necessarily know the probability associated with intervals determined by some or all of the tests in this study. Measurements from recovered fragments (also referred to as samples of questioned source) are then compared to this test interval to determine whether they "could have come from" or "are consistent with" this known source. Throughout this paper, we will refer to glass fragments as coming from either a known source (K) or a questioned source (Q).

The number and size of fragments collected from the broken glass object are typically much greater than the recovered fragments. As a result, hypothesis tests are often modified from traditional statistical tests, such as the *t*-test. The criterion used for determining whether a match exists when a small number of Q measurements are compared to many K measurements is typically derived either through tradition or from practical considerations. As a result, many specific methods have been advocated for defining the test interval for the RI of the K glass. In a recent informal survey, the Glass Subgroup of the Scientific Working Group for Materials found that a wide variety of match criterion are used by glass examiners for comparison of RI. Differences among these match criteria include the number of K fragments analyzed, the number of measurements made from each K or Q fragment, the method of determining the test interval for these multiple measurements, and the method of combining results when several Q fragments are considered. This paper reports the results of a study comparing the

[1]Visiting Scientist and Research Chemist, Counterterrorism and Forensic Science Research Unit, FBI Laboratory, FBI Academy, Quantico, VA 22135.

*Presented at the 61st Annual Meeting of the American Academy of Forensic Sciences, February 20, 2009, in Denver, CO.

†The Visiting Scientist Program is an educational opportunity funded by the FBI Laboratory and administered by the Oak Ridge Institute for Science and Education.

‡Publication No. 09-07 of the Laboratory Division of the Federal Bureau of Investigation (FBI). Mention of trade names is for information purposes only and does not imply endorsement by the FBI or the federal government.

**Retired. Present address: 2070 Farragut Drive, Stafford, VA 22554.

§The publication does represent work done by the FBI.

Received 27 Mar. 2009; and in revised form 2 Oct. 2009; accepted 10 Oct. 2009.

Type 1 error rates for eight commonly used RI match criteria for various numbers of K and Q measurements using typical sources of float glass.

When comparing glass fragments, the Q fragments may be treated individually or grouped together into sets of like glass. Many examiners, particularly those in the U.S., treat the Q fragments individually, because there is no *a priori* assumption that multiple fragments recovered from a single item of evidence originated from the same source. On the other hand, much effort has been spent on defining the best approach and developing algorithms for grouping Q fragment data in the belief that grouping of Q fragments improves the statistical reliability of comparison results (1–3). In this study, we only considered using individual data points for Q fragments for comparison with an interval of K measurements under two conditions. In the first case, we used only one RI measurement for the Q, as might be observed when the recovered fragment is quite small. In the second case, we considered the mean of three measured RI values for the Q. In casework, the three values could result from multiple measurements on a single fragment or measurements from different fragments that were grouped together. In both cases, we used only a single mean value for comparison to be consistent with the idea of the one Q versus many K match criteria that were tested in this study.

Materials and Methods

Glass Samples

Five sheets typical of modern float glass products were used for this study. These sheets included:

- No. 1 and no. 2: two sheets from a double-paned architectural window, 51 × 71 cm in dimensions, 0.226 cm thick, colorless.
- No. 3: a tempered automobile side window, irregularly shaped, *c.* 46 × 84 cm, light green in color.
- No. 4 and no. 5: two sheets from a laminated windshield from a Toyota Prius, each sheet between 0.210 and 0.214 cm thick, light green in color (replacement windshield).

RI Measurements

The method of temperature variation using a phase contrast microscope and hot stage controlled by a GRIM-3 automated glass RI measurement instrument (Foster and Freeman, Evesham, U.K.) was used to determine RI at a wavelength of 589 nm. Each of the two GRIM-3 instruments used in this study was calibrated for the Locke B oil using a minimum of five measurements of each of seven Locke Series B glass standards: B2, B3, B4, B6, B7, B8, and B9. A calibration response was generated by a linear least squares fit to the plot of the RI at 20°C for each glass standard against the measured match temperatures.

Each glass sheet was sectioned into four quadrants, and five samples were selected from each quadrant. The samples were cleaned by soaking in 30% nitric acid for 60 min, washed with deionized water, dried, and then crushed into smaller fragments for determination of RI. One or more fragments from each of the 20 samples from each sheet were further crushed as needed and mounted on a cleaned microscope slide in Locke B silicone oil (Locke Scientific Ltd., Basingstoke, U.K.) for measurement of RI.

For each sample slide, at least 10 RI determinations were made, for a minimum of 50 measurements per quadrant and 200 measurements per sheet. To avoid RI variations at the production surfaces of float glass, only RIs from the bulk glass were determined by making all measurements using freshly broken edges of fragments, which are easily recognized in the microscope image. Only one measurement per fragment edge was taken; that is, at least 10 fragment edges were measured per slide. In previous studies, we have noted that analysts sometimes have a natural tendency to bias their measurement results by selection of edges having low relief when viewed at temperatures close to the match temperature. To avoid this, the first 10 suitable edges that were found upon scanning over the slide were used for the measurements. Only one edge was measured on each instrument scan to avoid any possibility that precision could be adversely affected by differing focal depths of particles when the simultaneous multiple edge measurement function of the instrument is used. The order with which slides were selected for measurement was rotated among the four quadrants to eliminate any quadrant-to-quadrant biases that could result from instrument drift. Instrument calibration checks were performed five times daily using a glass reference material (designated as NBS 9012, previously provided by NIST, Gaithersburg, MD). Over an *c.* 1-month period, the means and standard deviations of the daily calibration test results were 1.51722 ± 0.00001 and 1.51721 ± 0.00002 for the two GRIM-3 instruments. Because the daily mean results agree well with the accepted value of 1.51722, the observed precision is within the instrumental precision and the largest daily mean deviation was 0.00002; normalizing corrections were deemed unnecessary. Although two GRIM-3 instruments were used in this study, all of the measurements for a given glass source were determined on the same instrument to avoid potential interinstrument biases. The RI measurement procedure used in this study follows the guidelines of ASTM Standard Test Method E1967-98 (4) with the exception that the calibration curve and resulting RI measurements are made at 20°C, instead of at the match temperature. This deviation from the standard procedure has no effect on any of the error rates measured in this study.

Statistical Tests

A total of 1023 measurements were made on glass fragments from the five sheets. All measurements were included in the data sets used to test match criteria. The data for each sheet were kept separate from those of the other sheets during the statistical evaluations. Tests for normality were conducted using Paleontological Statistics (PAST) (University of Oslo, Oslo, Norway), and other data analyses were performed using Microscoft Excel (Redmond, WA). A bootstrapping methodology in which the data points were randomly resampled was used to select subsample sets to test the Type 1 error rates of eight comparison criteria. The specific match criterion tests that were evaluated in this study, listed in Table 1,

TABLE 1—*Match criteria tested in the study.*

Test No.	Description	Test Criterion
1	1 × SD	$\bar{K} - 1 \times s_K \le \bar{Q} \le \bar{K} + 1 \times s_K$
2	2 × SD	$\bar{K} - 2 \times s_K \le \bar{Q} \le \bar{K} + 2 \times s_K$
3	$t_{0.05}$ × SD	$\bar{K} - t_{0.05}^{n_K-1} \times s_K \le \bar{Q} \le \bar{K} + t_{0.05}^{n_K-1} \times s_K$
4	$t_{0.01}$ × SD	$\bar{K} - t_{0.01}^{n_K-1} \times s_K \le \bar{Q} \le \bar{K} + t_{0.01}^{n_K-1} \times s_K$
5	Fixed 0.0001	$\bar{K} - 0.0001 \le \bar{Q} \le \bar{K} + 0.0001$
6	Fixed 0.0002	$\bar{K} - 0.0002 \le \bar{Q} \le \bar{K} + 0.0002$
7	Range	$K_{min} \le \bar{Q} \le K_{max}$
8	Range +0.00005	$K_{min} - 0.00005 \le \bar{Q} \le K_{max} + 0.00005$

Where, $\bar{K}$, mean of K measurements; $\bar{Q}$, either single Q measurement or mean of three Q measurements; s_K, standard deviation of K measurements; n_K, number of K measurements (5, 6, 7…20, 25, 30, or 40 measurements); *t,* value of t at n_K-1 degrees of freedom and stated two-tailed confidence level; K_{min}, minimum of K measurements; K_{max}, maximum of K measurements.

represent variants of the match criteria currently used by many of the glass examiners who use a test interval approach. In each test, if the mean Q value did not fall within the upper and lower limits based on the K measurements as shown in Table 1, a false exclusion was declared. The error rates for each test were measured using combinations of the number of K measurements equal to 5, 6, 7…20, 25, 30, 40 and Q measurements of one and three readings. For each sheet, the appropriate number of K and Q measurements were obtained by random selection with replacement from the 200+ measurements using the Random Selection Tool in Excel. For each test, the RI measurements for the Q samples were averaged and the mean point was compared to the test interval determined using the K measurements. Each subset of K and Q data points was used to test all eight of the match criteria. This process of data resampling and testing was repeated 1000 times for each sheet of glass. These simulations were designed to estimate the frequencies of false exclusions for each of the eight tests for a single questioned RI measurement value and for the average of three questioned measurement values. No cross-source comparisons were made in this study.

All analytical measurements were recorded to the nearest 0.00001. For the match tests, all calculated test interval limits and the mean values for triplicate Q measurements were also rounded to the nearest 0.00001. As indicated by the formulae in Table 1, when Q test points were equal to either limit of a given test interval, the Q was considered to be within the test acceptance range. Any result for which the test formula was not upheld represents a Type 1, or false exclusion, error, because for every comparison in this study, the K and Q measurements were known to have been taken from the same source sheet.

Results and Discussion

Distribution of RI Measurements Within a Sheet

To assess the degree of variability in each sheet, the distributions of measured RI values within each quadrant were compared. For each of the four sheets that were not tempered, the mean value of RI for each quadrant differed from the mean value for the entire sheet by no more than 0.00001. For the tempered sheet (no. 3), the differences between each quadrant mean and the sheet mean were 0.00001, 0.00001, 0.00002, and 0.00004, but the variation measures were greater than for the other sheets. Pairwise comparison of quadrants by analysis of variance (ANOVA) indicated that no significant differences exist between quadrants within a sheet. This result and the fact that the quadrant standard deviations are similar to those of the entire sheet strongly indicate that there are no significant differences between the mean of measurements from any of the quadrants and that of the entire population of measurements for that sheet. Therefore, all data for each sheet were pooled for further statistical analysis.

Histograms of the distributions of the RI measurements for each sheet are shown in Fig. 1, and descriptive statistics are listed in Table 2. Normal distributions based on the means, standard deviations, and numbers of measurements listed in Table 2 are superimposed on the raw data histograms in Fig. 1. Visual comparison of the raw data distributions with the normal distributions and the use of q-q plots (not shown) indicate that there exist some departures from normality for the measured RI for each glass source. Each of the five distributions was tested for normality using the Shapiro–Wilks, chi-squared, and Jarque–Bera tests. The calculated *p*-values for these tests are shown in the lower portion of Table 2. Lower p-values indicate departure from normal distributions. The Shapiro–Wilks, chi-squared, and Jarque–Bera test results all indicate that sheet no. 1, one of the architectural glass sheets, is the only one whose RI data are consistently assessed as being normally distributed at the 95% confidence level.

Several reasons for the deviations from normality are indicated by the descriptive statistics shown in Table 2. Skewness describes the asymmetry of the distribution by indicating the degree to which the data are shifted into one of the tails of the distribution relative to a normal distribution. The sign of the skewness indicates whether the data are shifted into the low tail (−) or into the high tail (+). Relatively low skewness values for sheet nos. 1, 2, and 3 indicate a small amount of data shifting into one of the tails. The greater negative values of skewness in the data for sheet nos. 4 and 5 indicated the presence of the more widely dispersed results in the tail on the low side of both distributions, a result that is supported by visual observation of the histograms shown in Fig. 1. Kurtosis is a measure of the peakedness of the distribution compared to a normal distribution. The greater the positive value of kurtosis, the more the data deviate above a normal distribution at its center. All of the distributions exhibit positive kurtosis, that is, the presence of points at too high frequencies near the center of the distribution. The results of the Jarque–Bera test that are based on skewness and kurtosis are in qualitative agreement with those of the Shapiro–Wilks and chi-squared tests that are based on overall deviations of the raw data from a normal distribution. Close observation of Fig. 1 reveals that most of the measurement histograms have more values in the center and on at least one of the wings and fewer measurements in the intermediate portions of the wings than would be predicted for normal distributions in agreement with the results of the tests for normality and the descriptive statistics shown in Table 2. The same conclusions concerning normality of the data distributions are reached by visual observation of q-q plots of the raw data, rather than the histograms shown in Fig. 1.

The major contributors to the lack of normality for sheet nos. 4 and 5 are the presence of several divergent measurements on the low side of each distribution. In particular, one measurement for sheet no. 4 at an RI of 1.51982 is so low that it is not shown in Fig. 1 to keep the horizontal scale similar to those for the other glasses. A Q test for outliers (5) indicates that this point could be excluded as an outlier value. If this data point were to be removed from the set for sheet no. 4, the standard deviation would decrease from 0.00005 to 0.00004, the range from 0.00064 to 0.00029, the skewness from −3.17 to +0.28, and the kurtosis from 29.20 to 1.68 making the distribution much closer to normal. However, we chose to retain this point for several reasons. First, sheet no. 5, the other pane of the double pane window containing sheet no. 4, has a nearly identical RI distribution and also has several points lying below the majority of the distribution. Second, in casework, the number of RI measurements on a given source will typically be a relatively small number, making it unlikely that outliers of this magnitude would be excluded by a statistical test. Finally, we have no reason to believe that any of the points lying on the low side of the distributions for sheet nos. 4 or 5 are a result of analytical error rather than representing either true variability within the glass or precision of the measurement method. As a result, all measured points were included in the data sets used for testing of match criteria.

Deviations from normality of measured RI values have been previously reported in other studies where measurements comparable to those in this study were made (6,7). Despite these minor deviations from normality, statistical tests that assume normality are used by some forensic laboratories for the interpretation of glass RI data

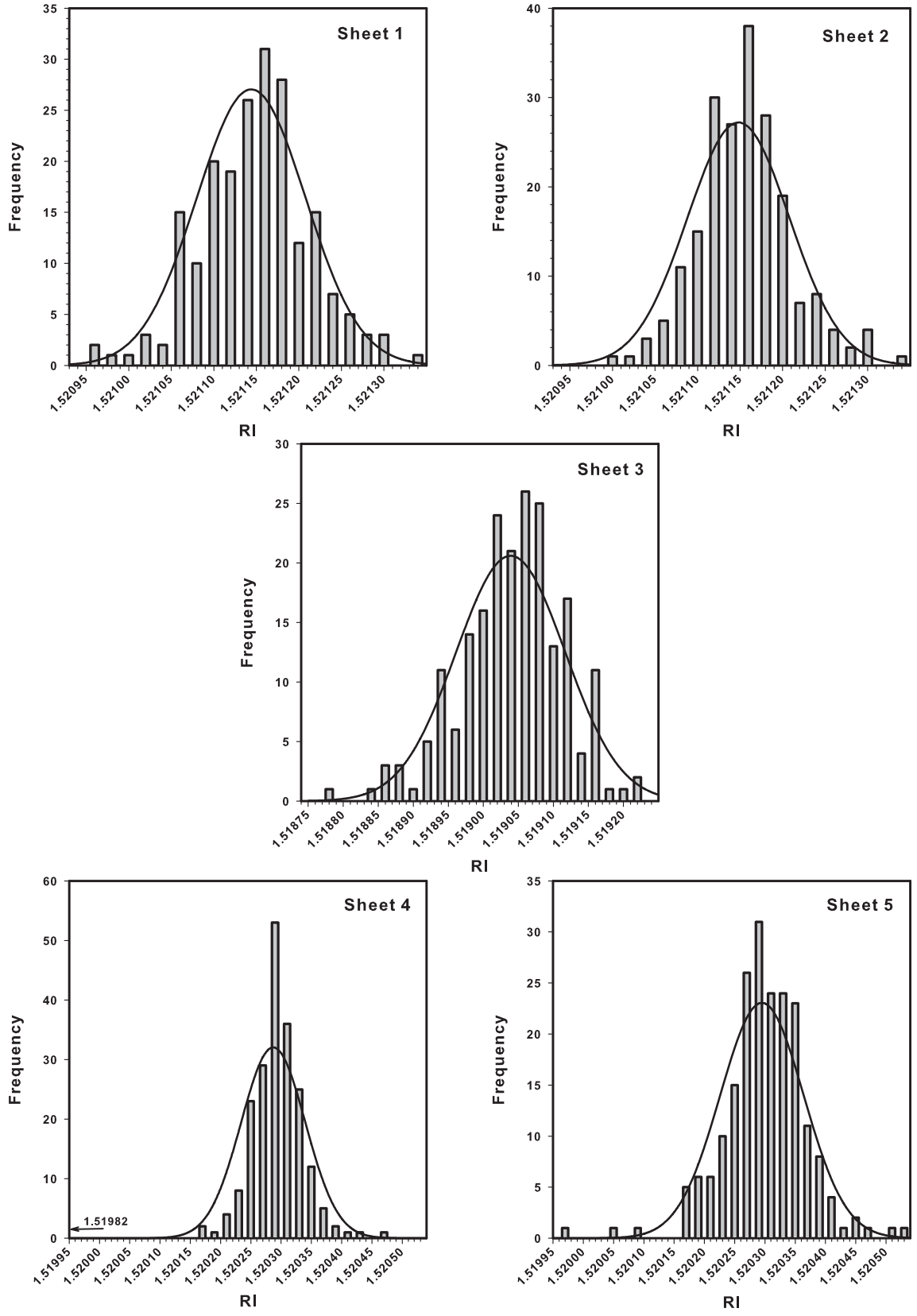

FIG. 1—*Histograms of RI measurement distributions with normal distributions overlaid for five glass sheets.*

in casework. Such match tests can still be used when underlying distributions are not normally distributed. However, deviations from normality may result in actual error rates that are either higher or lower than the theoretical rate. The purpose of this study is not to assess whether the distribution of data collected in casework is normally distributed, but rather to make empirical measurements of

TABLE 2—*Descriptive statistics of RI distributions in five sheets of glass. Normality test results are expressed as* p*-values. Normal distributions at significance above 5% are shown in bold type.*

	Sheet no. 1 Architectural Window	Sheet no. 2 Architectural Window	Sheet no. 3 Automobile Side Window	Sheet no. 4 Automobile Windshield	Sheet no. 5 Automobile Windshield
Number of measurements	204	206	207	204	202
Mean RI	1.52114	1.52115	1.51904	1.52029	1.52030
Standard deviation of RI	0.00006	0.00006	0.00008	0.00005	0.00007
Median RI	1.52115	1.52115	1.51904	1.52029	1.52029
Maximum RI	1.52133	1.52133	1.51941	1.52046	1.52053
Minimum RI	1.52095	1.52090	1.51877	1.51982	1.51996
Range of RI	0.00038	0.00043	0.00064	0.00064	0.00057
Skewness	−0.12146	−0.30918	0.09745	−3.17176	−0.55477
Kurtosis	0.46657	2.39719	2.47228	29.20763	3.57903
Normality test results					
Shapiro–Wilks	**0.204**	5×10^{-5}	0.000233	5×10^{-16}	3×10^{-6}
Jarque–Bera	**0.316**	2×10^{-21}	2×10^{-11}	$<10^{-25}$	8×10^{-25}
Chi-squared	**0.105**	0.00266	7×10^{-5}	5×10^{-5}	0.000916

error rates for several common match tests using real data distributions. The significant fact is that data sets even when collected similarly may or may not be normally distributed. In evaluation of evidence in casework, when the number of measurements is small, it is generally not possible to assess reliably the normality of a data distribution.

The standard deviations for each sheet shown in Table 2 are greater than the precision of the GRIM-3 instrument of between 0.00002 and 0.00003, indicating that at least a portion of the observed variation of RI is a result of heterogeneity within a sheet. Standard deviation is one of the two critical factors controlling discrimination capability of a measured parameter (the other being cross-source differences). As a result, there have been several reports in the forensic literature of standard deviation measurements taken across or throughout a sheet of glass. It is difficult to compare results between studies, because of differences in analytical methods and sampling protocols, particularly the separation of data from surface and bulk fragments. In one study, Locke and Hayes reported several values of standard deviations across a "nontoughened" float glass windscreen but considered a typical value to be 0.00009 compared with 0.00018 for toughened float glass sheets (8). In other studies, Locke and co-workers reported values of 0.00004 and 0.00005 for standard deviations of measurements of typical modern, presumably float, window glass (9,10). In more recent studies, typical reported standard deviations of RI values over sheets of various sizes are 0.00004 for nontempered float glass (7,11), 0.00007 for tempered float glass (6,11), and 0.00006 for nonfloat glass (11). The measured standard deviations shown in Table 2 are generally in agreement with those previous studies in which analytical precisions were similar to those of the method used here. As expected, of the glass sheets in our study, the tempered glass (sheet no. 3) has the greatest variability. The measured standard deviations are relatively insensitive to the presence of a few results lying on the wings of the distributions. A more sensitive measure of the spread of the data is the range of the results. Sheet nos. 3 and 4 have the widest ranges of RIs, both at 0.00064. This result was expected for sheet no. 3 because it is tempered. If the single measurement at 1.51982 was removed from the data set for sheet no. 4, the range would be reduced to 0.00029 as stated previously.

For the two instances where two sheets came from the same window (nos. 1 and 2 and nos. 4 and 5), the *t*-test for equality of the means and ANOVA both indicate no significant differences between the means. This most likely means that the two panes of each pair were cut from a single larger sheet of glass. We could combine the data into *c.* 400 points for each of these sources. However, for this study, we have chosen not to combine the data, but rather to treat each member of the pair separately, giving us five sets of data for testing match criteria.

We note that in casework, the distributions of RI measurements of known glass samples could well be different from those in this study for several reasons. The number of measurements made from broken glass evidence will almost always be <200, the nominal number of measurements in this study. Thus, the data will be collected over a shorter time period, possibly reducing its spread. In our case, it took up to 3 days to make the measurements for each sheet of glass. Although all precautions were taken to obtain consistent results, it is possible that instrument drift over a shorter time period could be less than that observed here. Another difference in casework is that the number and size of fragments available for measurement may be limited, even for the control glass. Although we detected no spatial variations among the samples in this study, this may not always be the case. Along these lines, Curran et al. (1) discuss the common observation in casework that the distribution of results for Q fragments is often wider than for K fragments, an observation that they attribute to unrepresentative proportioning of surface and bulk fragments upon breaking of a window. We prefer to avoid this data spreading problem by comparison of surfaces with surfaces and bulk with bulk. In this study, we excluded surface edges from the measurements, but recognize that this may not always be possible with small recovered fragments. Finally, glass objects other than the float glass sheets used in this study will be encountered in casework and they may have different RI distributions than those noted here. As always, for optimal evaluation of evidence, it is imperative that appropriate sampling of both K and Q fragments is made.

Error Rates for Match Tests

As stated previously, RI values obtained by random selection with replacement of the measurements made from a given sheet were used to test the false exclusion rate for the eight match criteria. The results for each test considered in this study are illustrated for the five glass sources in Figs 2–5. In the figures, each point represents the frequency of Type 1 errors (false exclusions) in 1000 tests displayed against the number of K measurements (n_K). For a given number of K and Q measurements, the same sets of data were used for each of the eight tests to remove any possibility that

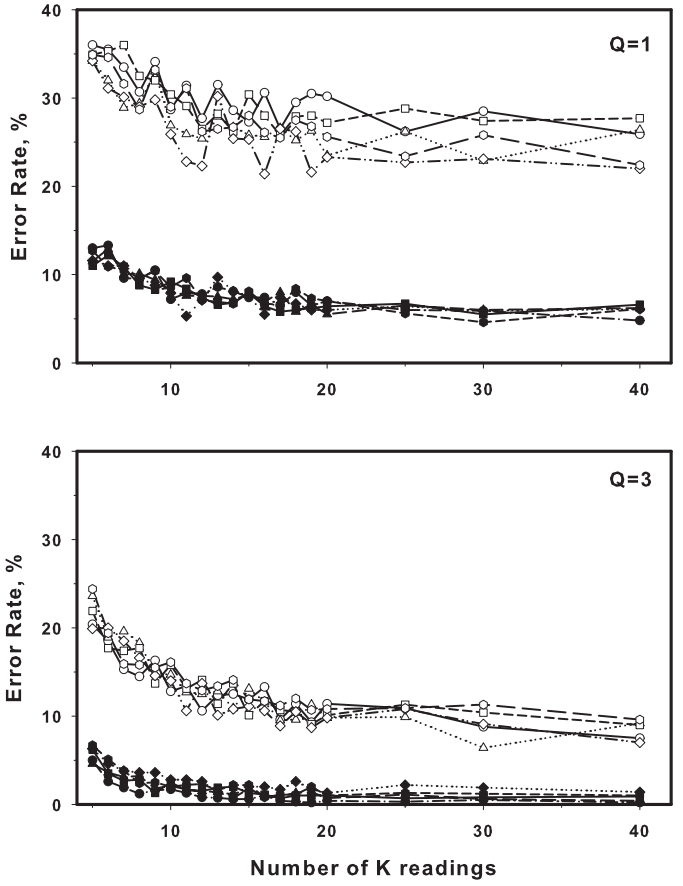

FIG. 2—*Results for Test 1 and Test 2: Fixed Standard Deviation Tests. Open symbols represent results for Test 1, and closed symbols represent results for Test 2 for the five sheets. Symbol legend: sheet no. 1 -circle-, sheet no. 2 -triangle-, sheet no. 3 -square-, sheet no. 4 -diamond-, sheet no. 5 -octagon-. Upper plot shows results for n_Q = 1, and lower plot shows results for n_Q = 3.*

cross-test bias could be introduced by the data selection procedure. Each figure displays the results for two tests for each of the five sheets of glass for the range of K values tested at both $n_Q = 1$ (upper) and $n_Q = 3$ (lower) conditions. Again, we note that any derived numbers, such as mean and standard deviation used in the various tests, were calculated based on the K and Q data points selected for that trial rather than from the 200 values for that sheet. Several notable trends shown in Figs 2–5 are discussed. Note that there is a statistical uncertainty associated with each point in these figures because of the resampling error of the bootstrap. This error was estimated for each measured error rate using a binomial distribution of 1000 trials. This bootstrap uncertainty ranges from *c.* 0.5% at an error rate of 1% to 3% at an error rate of 38%. In all instances where differences are noted, the differences are significantly greater than the bootstrapping error.

Tests 1 and 2: Tests Based on a Fixed Multiple of Standard Deviation—Match criteria based on a fixed number of standard deviations about the mean RI value of the known glass have been used for comparisons by glass examiners for many years. Typical values used for test intervals are 1σ, 2σ, or 3σ about the mean of the K measurements (1). The thinking behind these tests is that for a normal distribution, individual RI measurements from a Q fragment will fall within 1σ, 2σ, and 3σ of the mean value of the correct source sheet *c.* 68%, 95%, and 99% of the time, respectively. This test makes no correction for the number of K or Q measurements other than the fact that as n_K is increased, the estimate of the standard deviation becomes more accurate. We tested intervals about the sample mean of one times the sample standard deviation (Test 1) and two times the sample standard deviation (Test 2).

Results for Tests 1 and 2 are shown in Fig. 2. For each test and value of n_Q, there is a slight trend of decreasing false exclusion rates with increasing values of n_K, but there are no significant differences between the results for the five sheets of glass. The downward trend is more pronounced for low values of n_K and nearly levels off for n_K greater than *c.* 10. For Test 1, at $n_Q = 1$, the average percentage of false exclusions for the five glass sheets decreases from 35% for $n_K = 5$ measurements to 25% for $n_K = 40$ measurements. The numbers of false exclusions for Test 2 are consistently one-third to one-fourth of those for Test 1, reflecting the doubling of the match limits from one to two standard deviations. The average error rates for Test 2 at $n_Q = 1$ are 12% for $n_K = 5$ measurements decreasing to 6% for $n_K = 40$ measurements. The observed decrease in error rates with increasing n_K reflects the lack of correction for the number of measurements in these tests. Tests based on a fixed number of standard deviations nominally have an associated significance level, that is, the probability of a Type 1 error. However, this significance level changes with respect to sample size, a fact reflected in the downward trends shown in Fig. 2. For $n_Q = 1$, the error rates, particularly for small values of n_K, are greater than the assumed 32% and 5% that are often associated with 1σ and 2σ tests of populations. In all instances, the decrease in error rates with increasing values of n_K levels off at *c.* $n_K = 10$ to 15.

The error rates for Tests 1 and 2 when averages of three Q values are used are shown in the lower portion of Fig. 2. The error rates for each test for the $n_Q = 3$ results are approximately one-half to one-sixth those for the $n_Q = 1$ results, because the probability of getting an outlier value with a single measurement is greater than the probability of getting an outlier for the mean of three data points randomly selected from the full data set. For Test 1, at $n_Q = 3$, the mean false exclusion rates for the five glass sheets decrease from 22% for $n_K = 5$ measurements to 9% for $n_K = 40$ measurements. For Test 2 at $n_Q = 3$, the mean false exclusion rates decrease from 6% for $n_K = 5$ measurements to 1% for $n_K = 40$ measurements. The downward trends of error rates with increasing n_K up to *c.* $n_K = 10$ to 15 are also observed in the $n_Q = 3$ results.

Tests 3 and 4: Tests Based on t Times the Standard Deviation—One method of correcting for the effects of n_K on error rates seen when using fixed multiples of standard deviation is to use a test interval based on a t value as the multiplier of the standard deviation. Use of t as a multiplier adjusts the significance level to a more constant value independent of n_K. Note that this is not a true *t*-test, but rather is a test of one Q versus many K measurements. A *t*-test of sample means requires that both samples be normally distributed with approximately equal standard deviations, whereas the tests studied here merely use a t value as a multiplier to define the width of the test interval about the mean K value. The results for tests using t multipliers at significance levels of 0.05 (Test 3) and 0.01 (Test 4) are shown in Fig. 3. As indicated, both Tests 3 and 4 have the advantage of producing error rates that are relatively independent of both the width of the distribution of measured RI values for the source glass and the number of K measurements. This is of practical importance for casework where the examiner may have no prior knowledge about the source glass and sample size may be limited. It might seem somewhat surprising that the error rates are independent of n_K for these tests as the values of t change with

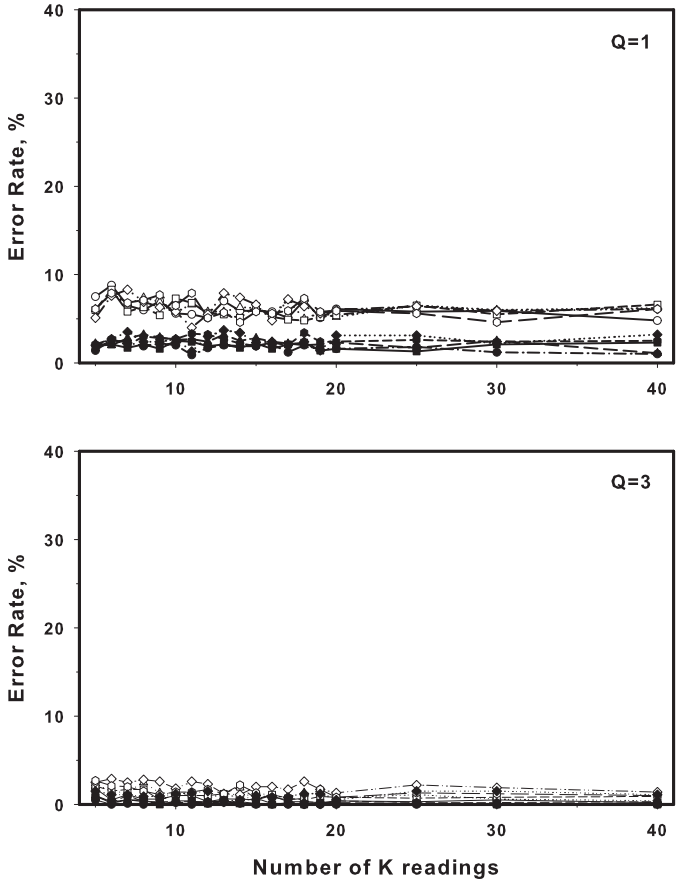

FIG. 3—*Results for Test 3 and Test 4: t Times Standard Deviation Tests. Open symbols represent results for Test 3, and closed symbols represent results for Test 4 for the five sheets. Symbol legend: sheet no. 1 -circle-, sheet no. 2 -triangle-, sheet no. 3 -square-, sheet no. 4 -diamond-, sheet no. 5 -octagon-. Upper plot shows results for n_Q = 1, and lower plot shows results for n_Q = 3.*

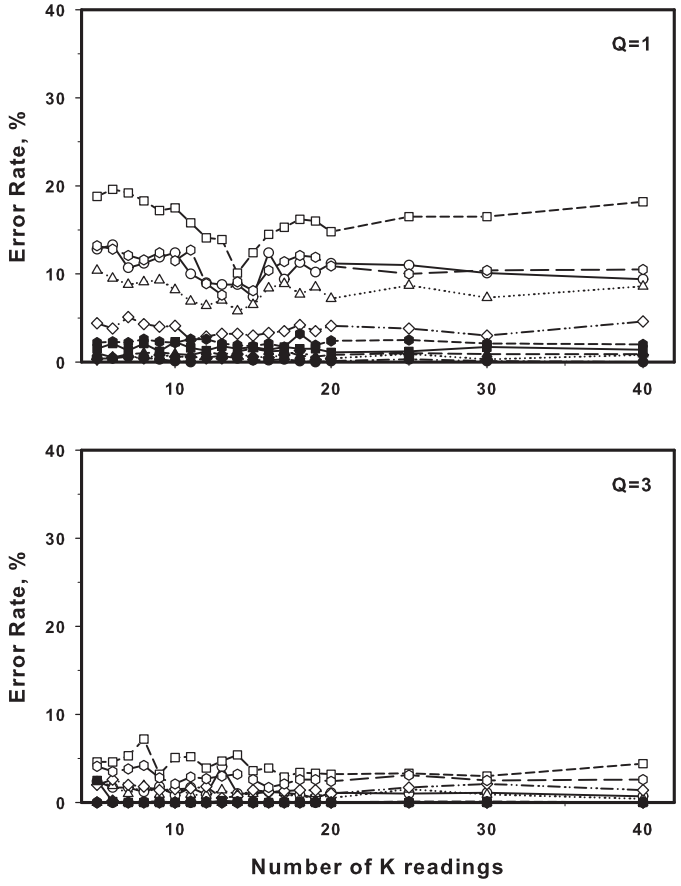

FIG. 4—*Results for Test 5 and Test 6: Fixed Interval Tests. Open symbols represent results for Test 5, and closed symbols represent results for Test 6 for the five sheets. Symbol legend: sheet no. 1 -circle-, sheet no. 2 -triangle-, sheet no. 3 -square-, sheet no. 4 -diamond-, sheet no. 5 -octagon-. Upper plot shows results for n_Q = 1, and lower plot shows results for n_Q = 3.*

values of n_K. The critical values of t used as multipliers in the test intervals decrease from 2.78 to 2.02 at a significance level of 0.05 and from 4.60 to 2.71 at a significance level of 0.01 as the number of degrees of freedom increases from 4 to 39. The narrowing of test intervals should have the effect of increasing the error rates with increasing levels of n_K. Apparently, this increase offsets the decreases shown for Tests 1 and 2 resulting in the nearly constant results seen for Tests 3 and 4.

For Test 3, at n_Q = 1, the error rates average *c.* 6%. For Test 4, at n_Q = 1, the error rates average *c.* 2%. The small differences between measured error rates and the significance levels of the t values for each test may in part be a result of the deviations from normality in the RI distributions.

Results for the n_Q = 3 studies display much scatter as a result of their low values and the resulting small vertical scale in the lower portion of Fig. 4. The error rates at n_Q = 3 are less than 3% for Test 3 and <2% for Test 4. At these low error rates where the bootstrapping error is a significant fraction of the total uncertainty and there is a high degree of scatter among results for different values of n_K, it is difficult to recognize any trends in the data that might be present. However, there is some indication that sample no. 4, the glass with the one extreme RI value, consistently yields slightly higher error rates than the other glass sources.

Tests 5 and 6: Tests Based on a Fixed Interval—Ever since Miller (12) suggested that "a positive opinion of nonidentity" should be made if flat glass specimens fail to match within the limit of ±0.0002, this value has been known as the "Miller criterion" for comparison of RI and has been utilized by many glass examiners as a match criterion for sheet glasses. The value of 0.0002 was derived from an overall average of the combination of the measurement precision and the variation within flat glass sheets, both of float and of nonfloat origin. Miller (12) did not intend for this match limit to be applied blindly or used in all situations, but rather it was given as a general guideline. Some examiners have decreased the allowed deviation from the mean to ±0.0001 to account for perceived improvements in analytical precision and float glass production quality or to limit the number of false inclusion (Type 2) errors. Fixed intervals about the measured mean of the known glass do not consider the number of measurements made nor do they consider any variation among measurements of the glass at hand in a given case.

The results for Tests 5 and 6 are shown in Fig. 4. Of the tests considered in this study, the two fixed interval match criteria result in the greatest differences in error rates among the five glass sheets. For Test 5, at n_Q = 1, the error rates for the tempered glass (sheet no. 3) are *c.* 5% greater and for one of the windshield sheets (no. 4) are *c.* 5% lower than for the other three sources across all levels of n_K. The error rates range from *c.* 5% to 20% for Test 5. The ordering of error rates for the five source glasses follows the rank order of the spread in RI measurements shown in Fig. 1 and Table 2. The larger spread in RIs for tempered glass results in

more errors of exclusion than for nontempered glass when a fixed interval match criterion is used because the match criterion does not consider the distribution of RI values within each glass source. It is interesting to note that the sample with the lowest error rate (sheet no. 4) is the one with the one low RI reading. The error rates are low for this sample because the other data points are more tightly clustered about the mean than are the measurements for any other sample. The error rates for Test 6 are below 5% for all of the glass sources and at all levels of n_K, and they do not display the dependence on glass source seen with Test 5. This probably is a result of the error rates for Test 6 being too low to display any observable trends. For both tests, there is no observable dependence of measured error rates on values of n_K.

For Test 5, the error rate is improved three- to fourfold when increasing n_Q from 1 to 3. Most of the measured error rates are <5% with the exception of the sheet no. 3 results at low values of n_K. Although the differences between samples are not significant when the bootstrapping error is taken into account, there is some indication that even at $n_Q = 3$, the error rates for the tempered glass are higher than those for the other sources. For Test 6, nearly all error rates are zero; only two test conditions at $n_K = 5$ for sheet no. 3 and $n_K = 6$ for sheet no. 4 produced any errors. We note here that 0.0002 is the widest test interval of any of those used in this study. It is included because of its long history, although it is not currently used much, if at all, by glass examiners. The near zero Type 1 errors for a fixed ±0.0002 test interval and $n_Q = 3$ could result in a relatively high number of Type 2 errors, so Test 6 is generally not recommended.

Tests 7 and 8: Tests Based on Range—In range tests, match decisions are based on observing whether the average Q value lies within the range defined by the maximum and minimum values observed in the K measurements. Range tests have appeal because they are nonparametric, meaning that they require no assumptions about the structure of the underlying population distributions. We also suspect that they have acquired some popularity because their lack of mathematical calculations makes them easier to explain to lay jurors and courtroom personnel. Disadvantages to using the range of data are that the range grows larger as more data are collected and the range does not reveal anything about the distribution of data between the two extreme values. The range or extended range tests are currently used by a number of glass examiners in the United States.

Of the tests studied, the range test results shown in Fig. 5 display the strongest dependence of the error rate on the number of K measurements. As n_K increases, the frequency of false exclusions decreases quite rapidly at first and then more gradually at higher values of n_K. This behavior is a direct and predictable result of the fact that the range increases as n_K increases. If all of the K and Q measurement results are placed in rank order, then the Q will be excluded only if it is ranked either first or last. The effect of the magnitude of n_K is readily illustrated, for example, by taking the case of 1 Q and 5 K random measurements from a single glass source. In this instance, the RI of the Q can fall in any one of six positions in the rank order. It will be ranked as either the highest or lowest measurement 2 of 6 times minus an allowance for the frequency at which the value of Q is equal to that of a K at either extreme of the sequence (note that all RI data are rounded to the nearest 0.00001 prior to comparison and that ties go to inclusion or no error). In general, the Type 1 error rate for a simple range test with 1 Q measurement will be slightly $<2/(n_K + 1)$. The values shown in Fig. 5 for Test 7 are consistent with this expectation.

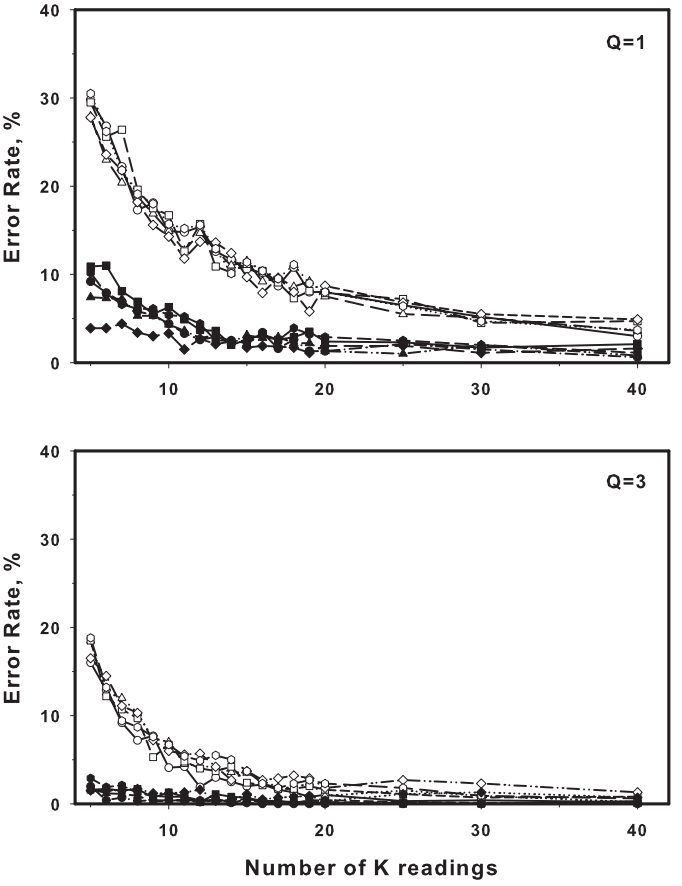

FIG. 5—*Results for Test 7 and Test 8: Range Tests. Open symbols represent results for Test 7, and closed symbols represent results for Test 8 for the five sheets. Symbol legend: sheet no. 1 -circle-, sheet no. 2 -triangle-, sheet no. 3 -square-, sheet no. 4 -diamond-, sheet no. 5 -octagon-. Upper plot shows results for $n_Q = 1$, and lower plot shows results for $n_Q = 3$.*

The trends in error rates for the range tests shown in Fig. 5 are the same for the five source glasses used in this study. This result reflects the fact that range tests are nonparametric and are based on the rank order of the measurements. The ordering of values within a subset of K and Q measurements is not related to the spread of the RI distribution except for the incidence of identical values at the ends of the ranked values. For Test 7, at $n_Q = 1$, the average percentage of false exclusions for the five glass sources decreases from 29% for $n_K = 5$ measurements to 4% for $n_K = 40$ measurements. The error rates are lower for Test 8, the extended range test with its wider test intervals; at $n_Q = 1$, the average percentage of false exclusions for the five glass sheets decreases from 8% for $n_K = 5$ measurements to 1.2% for $n_K = 40$ measurements.

The results for the $n_Q = 3$ case for Tests 7 and 8, shown in the lower portion of Fig. 5, resemble those for the $n_Q = 1$ case except that the error rates are lower. For Test 7, the false exclusion rate decreases from an average of 18% at $n_K = 5$ to 0.6% at $n_K = 40$. For Test 8, the error rates are below 5% for all values of n_K, averaging 2% for $n_K = 5$ and 0.2% for $n_K = 40$.

Conclusions and Significance

In this study, the Type 1 error rates for RI comparison using eight match criteria were empirically determined for five typical sheets of float glass. The trends in measured error rates as functions of the number of measurements and the distribution of RI values in

the glass sources reflect changes in the width of the comparison intervals. As expected, in all cases, larger test intervals and a greater number of Q measurements result in fewer false exclusion errors. In addition to this obvious qualitative conclusion, this study provides quantitative measures of the Type 1 error rates for several comparison criteria that are used or may be considered by forensic laboratories. The significance of the measured error rate values is that, although they compare well with theoretical, statistically based error rates, there are some minor differences for some of the tests. This occurs because either the test does not follow a true statistical test, because an assumption of the test is violated, or the number of measurements is insufficient to give meaningful theoretical error rates.

The relative error rates for the eight tests under selected conditions of n_K and n_Q can be made by side-by-side comparisons of Figs 2–5. For example, at $n_K = 5$ and $n_Q = 1$, the error rates for the tests range from *c.* 2% for Tests 4 and 6 to 35% for Test 1. For each pair of similar tests (1–8), the second test with its wider test interval results in a significantly lower error rate. In general, the error rates for the eight tests increase in the order 6<4<<3,8,2,5<7<<1. The results for Test 5 are strongly dependent on the differences among RI distributions for the various glass sources (resulting from differences in tempering), so the order of Test 5 with respect to Tests 2, 3, and 8 changes with glass type. The observed error rates change upon increasing the number of measurements for some of the tests. Those whose error rates change do so at different slopes across the plots, making some of the trend lines cross. As a result, the ordering of the tests changes at different levels of n_K. Finally, for every test and every level of n_K, the Type 1 error rates calculated using a single questioned measurement are greater than those calculated using the average of three questioned measurements.

Recently, a subcommittee of the National Research Council of the National Academy of Sciences released a report (13) containing a number of recommendations for improving forensic sciences in the U.S. Among their recommendations were the development and establishment of quantifiable measures of reliability of analyses and uncertainty in conclusions. This RI study was completed before the release of the NRC report. Nevertheless, we believe that the results obtained in this study address some of the questions concerning developing quantifiable measures of uncertainty in glass RI comparisons, at least in regard to Type 1 errors, and that this approach may also be applied to other forms of trace or transfer evidence. A discussion of the results of this study and their significance in light of the stated NRC recommendations follows.

We have made no recommendation concerning what is an appropriate Type 1 error rate. It should not be inferred from this paper that a lower Type 1 error rate is somehow "better" than a higher one. In fact, depending upon the circumstances, just the opposite may be true. It is well established that there is an inverse relationship between Type 1 and Type 2 errors. Methods that decrease the incidence of Type 1 errors by widening the test criterion will always increase the incidence of Type 2 errors to some extent. In contrast, methods, such as increasing the number of measurements of the Q sample to decrease the Type 1 error rate, are also likely to decrease the incidence of Type 2 errors. Type 2 errors are generally considered more insidious than Type 1 errors because a false association may lead to incrimination of an innocent subject. Selection of appropriate error rates for a given test is an ethical and administrative decision that has to date been up to each individual forensic laboratory or examiner. This study provides analytical data that can be used to select appropriate tests and evaluate Type 1 error rates for comparison of float glass samples. This study does not provide information concerning the incidence of Type 2 errors whose rates depend not only on the K and Q glass RI distributions and the test used but also on the distribution of RI values among the appropriate populations of glass sources consistent with potential alternative hypotheses.

One encouraging possibility that can be seen from the results of this study is that it is possible under certain conditions to decrease the number of Type 1 errors by a greater amount than the corresponding increase in Type 2 errors. To illustrate, suppose we have a typical population frequency distribution that displays a relatively constant frequency over the short interval corresponding to our test criteria. In this case, the Type 2 error rate is directly proportional to the width of the test criterion. As shown in Fig. 2, at $n_K = 20$ for example, if one were to double the test region from 1σ to 2σ (i.e., from Test 1 to Test 2), the Type 2 error rate would increase twofold, while the Type 1 error rate would decrease fourfold from *c.* 28% to 7%. Of course, a preferable approach when using match criteria similar to those in this study is to increase the number of Q measurements and thereby decrease the incidence of Type 1 errors without increasing Type 2 errors.

Another consideration in the selection of an appropriate test criterion is whether additional tests are being conducted on the evidence. For example, if a second more discriminating test, such as quantitative elemental analysis is being conducted, then the RI comparison may be considered more of a screening test where a greater number of Type 2 errors would be acceptable. When used in this manner, it is more advantageous to use a relatively wide match criterion for RI so that Type 1 errors would be eliminated. Most false associations by RI comparison would then be corrected by subsequent highly discriminating follow-on tests.

The results of this study may be utilized by glass examiners in two ways. First, the Type 1 error rate for a specific comparison criterion being used or under consideration may be determined from the results presented in Figs 2–5. Second, if it is desired to attain a particular Type 1 error rate, such as 5%, a procedure can be selected or modified to obtain conditions under which the desired error rate is obtained.

Regardless of whether one uses a two-stage approach or Bayesian approach, ultimately an estimate of Type 2 error rates is needed to evaluate the significance of the results. We recommend that the best way to utilize the approach used in this study is to first select a protocol based on the Type 1 error rates in the manner of this study. Once a protocol is selected, the appropriate number of RI measurements can then be obtained from many samples to make probability density functions corresponding to glass populations representing various defense hypotheses. These distributions can then be used to calculate Type 2 error rates for the selected test criterion without the need to acquire the 200 RI measurements on each sample in the data base.

As a final comment, we note the need to extend this study to other match criteria and methods of comparison. The eight comparison criteria selected for this study were chosen because these tests are or have been used by practicing glass examiners. As we first reported on the results of this study, many glass examiners have suggested other test criteria that should be evaluated using this RI data. Among these additions to our study are increasing the number of Q measurements, grouping sets of Q measurements, and evaluating tests (such as traditional *t*-tests) that require similar numbers of K and Q measurements. While we have not performed these tests across the ranges of test variables, we do note a couple of preliminary results. When n_Q is increased, the incidence of Type 1 errors decreases rapidly, as was indicated by the results for $n_Q = 1$ and $n_Q = 3$ in this study. When $n_Q > 5$, the error rates for all of the tests except Test 1 become so low or zero in some instances,

thereby requiring a larger number of resamplings to measure them accurately. We also performed a *t*-test with equal variance using $n_K = 10$ and $n_Q = 10$ at two levels of significance. The results of this quick study are that the average measured error rates for the five sheets are identical to the significance level of the test, i.e., Type 1 error rate for a *t*-test at $\alpha = 0.01$ was 0.01 and at $\alpha = 0.05$ was 0.05. The measured error rates differed slightly among the five glass sources, but, in general, the previously discussed deviations from normality in the RI distributions do not cause significant differences between observed and theoretical error rates for the *t*-test. Extension of these preliminary studies to additional test criteria is needed. The data from this study could also be used in a Bayesian approach for the calculation of uncertainties in the probabilities that are included within the numerators of likelihood ratios. We are eager to make the RI data from this study available to anyone who would like to use it for these or other studies. A copy of the analytical data for the five glass sheets may be obtained upon request from the corresponding author.

Acknowledgments

We gratefully acknowledge the advice offered during the conduct of this project and comments on the manuscript provided by Jodi Webb, Maureen Bottrell, David Korejwo, Cary Oien, and JoAnn Buscaglia. We are grateful to three anonymous reviewers who provided a number of comments and suggestions that improved this paper.

References

1. Curran JM, Hicks TN, Buckleton JS. Forensic interpretation of glass evidence. Boca Raton, FL: CRC Press LLC, 2000.
2. Evett IW, Lambert JA. The interpretation of refractive index measurements III. Forensic Sci Int 1982;20:237–45.
3. Curran JM, Triggs CM, Buckleton J, Coulson S. Combining a continuous Bayesian approach with grouping information. Forensic Sci Int 1998;91:181–96.
4. ASTM International. ASTM E1967-98. Standard test method for the automated determination of refractive index of glass samples using the oil immersion method and a phase contrast microscope. West Conshohocken, PA: ASTM International, 2003.
5. Dean RB, Dixon WJ. Simplified statistics for small numbers of observations. Anal Chem 1951;23:636–8.
6. Pawluk-Kǒlc M, Zięba-Palus J, Parczewski A. The effect of re-annealing on the distribution of refractive index in a windscreen and windowpane. Forensic Sci Int 2008;174:222–8.
7. Bennett RL, Kim ND, Curran JM, Coulson SA, Newton AW. Spatial variation of refractive index in a pane of float glass. Sci Justice 2003; 43:71–6.
8. Locke J, Hayes CA. Refractive index variation across glass objects and the influence of annealing. Forensic Sci Int 1984;26:147–57.
9. Locke J, Underhill M, Russell P, Cox P, Perryman AC. The evidential value of dispersion in the examination of glass. Forensic Sci Int 1986; 32:219–27.
10. Locke J. GRIM—a semi-automatic device for measuring the refractive index of glass particles. Microscope 1985;33:169–78.
11. Sandercock PML. Sample size considerations for control glass in casework. Canadian Soc Forensic Sci J 2000;33:173–85.
12. Miller ET. Forensic glass comparisons. In: Saferstein R, editor. Forensic science handbook. Englewood Cliffs, NJ: Prentice Hall, 1982;171.
13. Committee on Identifying the Needs of the Forensic Science Community, National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: National Academies Press, 2009.

Additional information—reprints not available from author:
Robert D. Koons, Ph.D.
2070 Farragut Drive
Stafford, VA 22554
E-mail: rdkoons@verizon.net