Clinical Grading and the Effects of Scaling

Ian L. Bailey,* Mark A. Bullimore,* Thomas W. Raasch,† and Hugh R. Taylor†

In clinical practice, there has been a need to grade the magnitude or the severity of the functions and qualities that are assessed in the examination. It is popular to use a four-step grading scale to categorize the severity of clinical findings. The authors discuss clinical grading scales and their influence on the clinician’s ability to detect change. These principles have been applied to grades or measures derived from either objective measuring instruments, subjective tests, or techniques in which the clinician makes subjective judgments. A hypothetical data set was used to show the problems associated with using grading scales that are too coarse. The authors presented a mathematical model that helps to estimate the benefits of using use of a finer scale. Data were presented from two separate studies, one on visual acuity measurement and the other on grading nuclear opacity, to show the advantages of using finer scales to enhance the sensitivity of clinical measurement. High levels of concordance between independent observations indicated that the grading scale was too coarse and that these scales needlessly reduced the clinician’s ability to detect change in the parameter being assessed. For moderate sensitivity, the size of the scale increments should not exceed one standard deviation of the discrepancy so that the concordance of paired comparisons would not exceed 37%. For fine clinical sensitivity, the size of the scale increments should not exceed one third of the standard deviation of the discrepancy, in which case the concordance of paired comparisons would not exceed 13%. The theory and evidence presented here could prompt re-evaluations of common methods of clinical grading. Invest Ophthalmol Vis Sci 32:422-432, 1991

The use of numeric scales to grade the severity or advancement of clinical signs is becoming widespread. Rather than applying descriptive or qualitative terms, ie, inqipient, slight, moderate, mild, severe, pronounced, or mature, to indicate a stage of development of conditions such as cataract or retinopathy, numeric scaling or grading systems are often used. The clinician makes an observation and assigns a numeric grade. This grade serves as a standard by which any future change may be judged. The same grade might also be used in determining deviations from normality or as a factor contributing to treatment decisions. Clinical grading systems have become important elements in many longitudinal and cross-sectional research studies.

The grading of cataracts is a good example. There have been moves to establish standardized scales to grade selected features of cataracts.1-3 For the grading of nuclear opacity, four standard photographs are commonly used to provide four reference levels of severity. The clinician observes the patient’s crystalline lens with a standardized arrangement of the slit-lamp biomicroscope and, by referring to the standard photographs, assigns a grade for any observed nuclear opacity. A grade of 2 would be assigned to an example of nuclear opacity that appears to be equal to or greater than that shown in the Grade 2 photograph but less than that shown in the Grade 3 photograph. The numeric grading system and the use of standard reference photographs have undoubtedly enhanced the reproducibility of clinical estimates, but there are associated problems that have important clinical implications.

There are difficulties involved in dividing a continuous scale into categories that are determined by subjective impressions or by measurements that are not perfectly consistent. Are the grading increments (or categories) too broad or too coarse? Do they represent approximately equal steps across the range being considered? Do they categorize the parameter being evaluated in a manner that provides information that is useful in the clinical decision processes? For the clinician, the immediate decision usually concerns whether the latest observation grade or measure is different from what was expected. The expected value may be a grade or score recorded at an earlier time or
it may be an established normal value. For all systems used to grade continuous variables, there is a statistical probability that a second observation will result in a grade which is different from the first when in fact, there has been no real difference in the conditions being compared. Similarly, there will be times when the assigned grades remain the same, even though there has been a change in the condition being evaluated.

The clinician has to develop confidence limits to be applied when determining whether there is a difference between the observed and the expected findings. Confidence limits are placed about the expected value and, should the observed value fall outside the limits, the clinician may say that there is a significant difference between the two values. These limits are often set at the 95% confidence level; when there is no real difference between the observed and expected values, there is only a 5% (or 1 in 20) chance that the recorded grades will indicate an apparent difference that is equal to or exceeds these limits.

These confidence limits must depend on the clinician's ability to be consistent in the assignment of grades. In describing consistency of grading, reference is often made to the concordance, or frequency of perfect agreement, between the grades assigned by pairs of independent observations. Most commonly, the pairs being compared are either test and retest scores by the one observer or the scores from two separate observers. If both observers are well trained, there should be similar levels of concordance for the within-observer and between-observer comparisons.

Another method of evaluating agreement between independent measures uses the kappa (κ) statistic, which expresses the agreement between two measures that cannot be accounted for by chance. The weighted κ statistic considers discordant values and applies weights according to the magnitude of the discrepancies. However, we will show that good agreement between independent measures should not be taken as a primary indicator of the value of a clinical grading system. The better the observers' ability to be consistent in their assignment of grades, the better their sensitivity to detect change. Grading systems that yield high levels of concordance between paired independent observations do not necessarily provide narrow 95% confidence limits, which are clearly desirable for the detection of clinical change. The level of concordance represents the interaction between the variance in the observers' assignment of grades and the size of the scale increment. High degrees of concordance may seem desirable because they indicate good consistency in grading, but they also indicate that the confidence limits could be significantly narrower if finer scale increments were adopted. Indeed, to be sensitive in the detection of change, the clinical measures or grades should not show high degrees of concordance.

We discuss clinical grading scales and present a method for estimating the benefit of adopting finer scales. The principles that we discuss can be applied to grades or measures that come from objective measuring instruments (eg, tonometers, keratometers, autorefractors, pupillometers), from tests in which the patient gives subjective responses (eg, visual acuity, contrast sensitivity, subjective refraction) or from techniques in which the clinician makes subjective judgments (eg assessment of lens opacities, anterior chamber depths, cup-disc ratios, grading retinopathies). First, we use a hypothetical data set to show the problems associated with grading scales that are too coarse. High levels of concordance indicate that the grading scale is too coarse and that this coarseness needlessly reduces the clinician's sensitivity to detect change. Second, we present a mathematical model that enables the benefits of adopting a finer scale to be estimated. We then present data from two separate studies, one on visual acuity measurement and the other on grading nuclear opacity, to illustrate how finer scales can substantially enhance the sensitivity of clinical measurement by narrowing the 95% confidence limits for change.

Materials and Methods

Comparing Fine and Coarse Scaling: A Hypothetical Data Set

For this discussion, we generated a hypothetical data set. Imagine that a certain parameter was measured on a group of patients and it was found that 100 patients had scores, in whole numbers ranging from 10–90 points on a 100-point measuring scale. Then, a second independent measure was made on this group of patients, whose condition was unchanged. The second measure of the same parameter might have been a subsequent measurement by the same observer or a separate measurement by a second observer or a measurement using a different instrument. The data set was generated by Monte Carlo methods, and the variances for the two measurements were equal throughout the range. The range of values for the second measurement should be expected to exceed the 10–90 point range of the first measurement.

Results

Figure 1A is a scatter plot in which each point represents the two measurements from one patient. The solid line at 45° is the locus for perfect agreement between the paired observations A and B. To the right
of Figure 1A, a histogram shows the distribution of the discrepancies (Discrepancy = Observation B – Observation A). This distribution has a mean of 0.2 points, which is close to zero, indicating there is virtually no bias between the two observations. The standard deviation of the discrepancy distribution is calculated to be 8.7 points, and the 95% confidence limits may be determined simply by multiplying this...
value by 1.96 (1.96 x 8.7 = 17.1). Given that we are using an integer scale, in order to fall just outside these limits, there must be at least an 18-point difference between the two results being compared. There is a less than 5% chance that any patient who has not undergone any real change will show a change of 18 points or more. In Fig. 1A, the 95% confidence limits are drawn as dotted lines parallel to the unit ratio line (this assumes that the variance of discrepancies is uniform across the scale). Only 4 of the 100 points lie exactly on the locus of perfect concordance, ie, the concordance is 4%.

Now consider what would have happened had the observer been required to assess the same parameter by assigning one of three grades (grade I, II, or III). If the observer(s) were to be perfectly consistent is assigning grades by this coarser system, then all measures between 1 and 33 would have been assigned a grade of I, those 34 to 66 would have been assigned a grade of II, and those 67 to 100 would have been assigned a grade of III. Figure 1B superimposes this coarse grading system on the original data set. Points that lie within the shaded area would be assigned the same grade by both observation A and observation B, ie, all points in the shaded area give perfect concordance when this coarser scale is used. Here, 75% of the data points are concordant and 25% discordant. To apply 95% confidence limits when the measurements are made using this coarser scale, one scale increment (33 points) is not sufficient because this only represents the 75% confidence limits. Two scale increments (66 points) represent the 100% confidence limits, but due to the coarse nature of the scale, this must also represent the 95% confidence limits. This means that only a change from I to III or from III to I should be considered to be significant. Compared with the finer scale, the coarse three-step scale gives much better concordance (75% compared with 4%), but it is substantially less sensitive to change since the width of the confidence limits is 66 points compared with 18 points.

Closer inspection of the individual points on Figure 1B shows some of the basic problems with coarser scaling. Some of the points that lie in the unshaded (discordant) areas are close to the locus of perfect concordance (ie, the 45° line). With the coarse grading scale, a discrepancy of one large grading unit would occur if the observer was inclined to score 32 points but, being obliged to use the coarser scale, assigned a score of grade I. At the second measurement, the observer's inclination been to assign a score of 34 points, a score of grade II would be assigned on the coarser scale, so a change from grade I to grade II would be recorded. At the other extreme, the same change from grade I to grade II would be recorded when the apparent change was from 1 point to 65 points. On the other hand, it is possible for large and easily discerned change to occur but, due to the coarseness of the scale, there might not be any change in grade. A change from 34 to 66 points would not register any change by the coarser grading system. For the data from Figure 1, the most extremely discordant value (x, y coordinates = 59, 34) would be assigned grade II by both observations A and B, despite there being a 25-point difference between the more detailed observations.

Figure 1C shows the results of adopting a less coarse scale that has nine increments across the range. Here, the concordance has dropped to 40% as shown by 40 of the 100 points falling into the nine shaded cells that represent perfect concordance between observations A and B when using this scale. From the distribution of discrepancies as shown on the right, it can be seen that, in order to fall outside the 95% confidence limits, the discrepancies between grades must be at least three increments. This represents 33 points (3 x 11). This intermediate nine-step scale allows greater sensitivity than the coarse three-step scale since the confidence limits are substantially narrower (33 compared with 67 points), but sensitivity would be further enhanced by adopting the finest (100-step) of the three scales for which the confidence limits are 18 points.

Modeling Relationships Between Scale Coarseness, Concordance, and Confidence Limits for Clinical Change

This hypothetical example shows that 95% confidence limits for change are strongly influenced by the coarseness of the measuring scale. If there are no limits to the fineness of the measuring scale, the 95% confidence limits for change become solely dependent on the true standard deviation of the discrepancy distribution. The true standard deviation of the discrepancy distribution reflects the variance in the measurement when all influences of scaling have been eliminated. Adopting a coarser scale causes an artificial increase in the standard deviation of the discrepancy distribution, and this widens the 95% confidence limits. Furthermore, the confidence limits must be expressed as an integral number of increments on the grading scale; this is a second factor that extends the range of 95% confidence limits, often by substantial amounts. With coarse scale increments, situations may arise in which the 95% confidence limits, the 99% confidence limits, and the 99.9% confidence limits will be identical.

The frequency at which perfect concordance occurs in paired comparisons (eg, test–retest or between observers) provides an index of the relation-
ship between the coarseness of the grading scale and the true standard deviation of the discrepancy distribution. From measures of the concordance, one can estimate the ratio of the size of the scale increments to the true standard deviation of the discrepancy. Conversely, the expected concordance can be predicted from knowing the ratio of the increment size to the true standard deviation of the discrepancy.

To develop a simple model, four assumptions have been made. First, it was assumed that the distribution of discrepancies is normal. This may not be true of all grading scales but the clinical examples presented later in this report show discrepancy distributions that are approximately normal as do results from many other clinical studies.5,6 Second, it was assumed that there is no systematic bias, ie, the mean discrepancy is zero. For objective clinical measurements made by trained observers, this is probably a valid assumption, but it should be recognized that there can easily be biases between observers or between instruments and also that for test–retest comparisons, there may be learning or order effects that cause systematic bias.7,8 Third, it was assumed that variance is uniform across the range of the scale. For many clinical tests, there could be larger variances shown in measurements made on patients affected by disease.6,9 The increments on commonly accepted clinical scales have usually been chosen by an informal and evolutionary process which reflects the judgments of experienced and respected clinicians. Although it might not have been a conscious consideration, the chosen scale increments may often produce relatively consistent variance across the range of the scale.10 Fourth, it was assumed that there are no truncation effects caused by end restrictions of the scale. Scales that have intrinsic upper or lower limits usually cause reduced variances when values are near the limits of the range, eg, Snellen visual acuity charts.

The expected concordance C, ie, the frequency (\(v_{d=0}\)) at which there is no discrepancy between paired comparisons, can be predicted from the ratios of the scale increment size to the size of the true standard deviation of the discrepancy distribution. The expected concordance can be calculated from the formula:

\[
v_{d=0} = C = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{0} \left(1 - \frac{z}{R}\right) \cdot e^{-\left(\frac{z^2}{2}\right)} \cdot dz
\]

where C is the frequency of perfect concordance and R is the ratio of scale increment size to the true standard deviation of the discrepancy distribution.

Fig. 2A shows the relationship between expected concordance C and the ratio R. It can be seen that when the scale increment is equal to one standard deviation of the discrepancy distribution (ie, R = 1), then C = 0.37 (ie, concordance equals 37%). Similarly, when the scale increment size is equal to four standard deviations of the discrepancy (R = 4), then C = 0.84 (ie, concordance equals 84%).

A family of equations similar in form to the above equation can be derived to calculate the expected distribution of the discrepancy values according to the size of the scale increment as expressed by the value of R. A table giving the frequency distribution of discrepancy values for a range of values of R is shown in the Appendix. From such tables, the 95% confidence limits may be determined for a range of values of R (Fig. 2B). These confidence limits are expressed as
multiples of the true standard deviation of the discrepancy. Finer scaling causes the 95% confidence limits to trend toward 1.96 true standard deviations. Conversely, with an increase in the ratio of scale increment size to the true standard deviation, there is dramatic broadening of the 95% confidence limits for change. Table 1 shows the expected concordance and the expected 95% confidence limits for a number of selected values of $R$. It shows that coarsening the scale leads to an increased concordance but a broadening of confidence limits. It can be seen that the curve (Fig. 2B) is not smooth, but is jagged, which is a product of using incremental scales (Appendix).

Some clinical studies have reported high levels of concordance between independent observations (eg, 90–99% for grading of cataract, 48–64% for estimating cup-disc ratios, 40–70% for grading of proliferative vitreoretinopathy). These levels of concordance suggest that the scale increments are substantially larger than the true standard deviation of discrepancy and that observers are capable of making much finer judgments than the scale allows. Scales that yield concordance values of more than 80% are likely to have 95% confidence limits that are so broad that they extend to eight or more true standard deviations (Table 1).

### Clinical Examples

#### Measurements of visual acuity

The grading of visual acuity provides a good example of some of the consequences of grading with a coarse scale. Authoritative bodies recommend or permit visual acuity scores to be assigned according to the smallest row in which a specified proportion of the letters (or other optotypes) are read correctly. Then, the only available grades of acuity correspond to the letter sizes available on the chart and there is no interpolation between these sizes. Most serious clinical research studies that involve the measurement of visual acuity use charts that follow the design principles of Bailey and Lovie. These charts standardize the task at each size level so that from one row to the next, the only significant variable is size. In the common format, there are five letters on each row and a logarithmic size progression in which the multiplication ratio 1.2589 (0.1 log units) is used. Within rows, the between-letter spacing is equal to the letter width, between rows the spacing is equal to the height of the letters in the smaller row, and the letter difficulty is standardized at each size level.

A grading system that was introduced with these chart design principles assigns a visual acuity score according to the number of letters read correctly. The visual acuity is expressed as log MAR (the logarithmic value of the minimum angle of resolution). On this scale, a visual acuity of 20/20 corresponds to log MAR = 0.00, 20/40 corresponds to log MAR = 0.30, and 20/200 corresponds to log MAR = 1.00. The letter sizes on the chart progress in 0.10-log unit steps. Since there are five letters on each row, each letter is assigned a value of 0.02 log units. A patient who reads all letters on the 20/50 row (log MAR = 0.40) and then reads two more letters on the 20/40 row (log MAR = 0.30) would be assigned a score of 0.36 log MAR units. The same score of 0.36 would be given if one letter was missed on the 20/50 row and three additional letters were read on the next (20/40) row. The advantage of assignment of credit for each additional letter read has not been widely recognized.

#### Materials and methods

An experiment was conducted in which 21 normally sighted subjects read two versions of each of three different charts (British Letters, Sloan Letters, and Landolt Rings) at three different distances. This allowed 189 test–retest comparisons by pooling across test conditions. Record was made of every letter read correctly, and visual acuity scores were assigned by two different methods. For the first method, credit was given for each letter read. For the other, a row was deemed to be read if at least three of five letters were read correctly.

#### Results

Figure 3 shows the discrepancies between the test and retest scores assigned by the two systems. In Figure 3A, it can be seen that assigning equal credit for every letter leads to perfect concordance (ie, zero discrepancy) between test and retest in 20% of cases. Note that in 97% of the comparisons, the test–retest discrepancy did not exceed four letters (0.08 log MAR units). When the same results were scored on the coarser row-by-row scale (Fig. 3B), 61% of the
test-retest comparisons gave perfect concordance, 38% showed a discrepancy of one full row (0.10 log units), and 1% showed a discrepancy of two full rows (0.20 log units). For the letter-by-letter scoring system, the clinician would assign 95% confidence limits of five letters or 0.10 log units. That is, when there has been no real change, there is less than a 5% chance that the visual acuity score at a second measurement will be different by 0.10 log MAR units or more. For the letter-by-letter grading of visual acuity, the scale increments are effectively five times finer, so the R value is expected to be 0.40. If R = 0.40, it is predicted that there should be 16% concordance and that the 95% confidence limits should be five scale increments which, in this case, is five letters. These predictions are in close agreement with the observed findings (C = 20%, confidence limits = 5 letters).

Grading of Nuclear Opacity

A four-step grading scale is commonly used to grade the severity of nuclear opacity of the crystalline lens from observations made with the slit-lamp biomicroscope. In many research projects, a set of four reference photographs are provided as the benchmark examples of the appearance of grade I, II, III, and IV nuclear opacity. To assign a grade of II, the observer should judge the nuclear opacity to be at least as pronounced as that illustrated in the grade II reference photograph, but less than that shown in the grade III photograph. For most research projects, the only permissible grades are the integer values, and no interpolation is allowed.

Materials and Methods: Slit-lamp photographs of 87 crystalline lenses were shown to two experienced observers who graded the severity of nuclear opacity. The observers were instructed to use decimalization, dividing each integer scale unit into 10 equal parts, to interpolate between the four benchmark grades shown in their reference photographs. Lenses with nuclear opacity that was more pronounced than grade II but less pronounced than grade III could be assigned grades of 2.1, 2.2, 2.3... or 2.9. Any score from 2.0 to 2.9 inclusive on the decimalized scale should be assigned a grade of II when the coarse integer scale is used.

Results

The evaluations of the two observers were compared for each of the 87 lenses. The data were analyzed first using their scores in decimalized form, and second using scores that would have been assigned on
an integer scale. Figure 4 shows the distribution of discrepancies between the 87 paired observations. With the coarser integer scale, the concordance was 74%, and for the remaining 26% of cases, the discrepancy was equal to ±1 integer grade unit. To fall outside the 95% confidence limits, there must be a difference of 2 integer grading units between the paired observations. By the finer decimalized scale, there was only 9% concordance, but the 95% confidence limits would be set at 0.8 grading units. The decimalized scale thus gives confidence limits that are 2.5 times finer than those for the integer scale.

We can examine how the data agree with predictions made by the model presented earlier. The 74% concordance obtained when using the integer grading scale suggests that the increment size is 3.15 times the true standard deviation of the discrepancy distribution, i.e., $R = 3.15$. If this is correct, then for the decimalized scale, which is being 10 times finer, $R = 0.315$. The model predicts that if $R = 0.315$, concordance equals 13% and 95% confidence limits equals seven scale increments, which is 0.7 grading units. Both predictions agree well with the observed findings ($C = 9\%$, confidence limits = 0.8 units).

(a) GRADED BY DECIMAL

(b) GRADED BY INTEGER

Fig. 4. The distribution of the discrepancies between nuclear sclerosis grades as assigned by two observers. (a) decimalized grading system. (b) integer grading system. The 95% confidence limits are shown for comparison.

Discussion

These two clinical examples show that widely used and authoritatively advocated grading systems can be made substantially more sensitive by the adoption of finer scaling. This can substantially improve the clinician's ability to recognize changes in the functions being assessed. This has important ramifications for both the monitoring of change in individual patients and the study of changes within groups of subjects. Making the scale finer will always improve the sensitivity by narrowing the 95% confidence limits for change. Once the size of the scale increment exceeds one true standard deviation of the discrepancy ($R = 1$), there is a sharp broadening of the 95% confidence limits (Fig. 2B and Table 1). However, when the scale increment is less than one third of the true standard deviation of discrepancy ($R = 0.33$), a finer scale achieves only modest improvements in sensitivity.

To set confidence limits for change of a particular parameter, the clinician can make independent measurements of that parameter in subjects for whom it is believed there has been no change. The frequency with which there is perfect agreement between the first and second measure is the concordance. When the concordance is known, the ratio of increment size to the true standard deviation of the discrepancy can be estimated (Table 1 or Fig. 2B). From this information, the size of the true standard deviation of the discrepancy can be estimated. For example, if test-retest comparisons of results from a particular test yield 60% concordance, it is estimated that the ratio of the size of the scale increment to the true standard deviation of the discrepancy ($R$) is 2. If $R = 2$, the 95% confidence limits would be equivalent to four standard deviations of the true discrepancy, twice that which would be obtained by using an unrestricted (very fine) scale. Substantial benefit may be gained by simply making the scale finer by a factor of four. The ratio of the size of the scale increment to the true standard deviation of the discrepancy ($R$) would thus be reduced to 0.5 ($= 2/4$), giving a predicted concordance of 19%. Table 1 predicts that, for the finer scale ($R = 0.5$), the 95% confidence limits would be substantially narrowed; they would be equivalent to 2.5 standard deviations of the true discrepancy.

When using such a procedure to set confidence limits, it is important that the test–retest results represent the most clinically appropriate comparisons. If the common clinical practice was to monitor a particular measurement on a monthly basis, then confidence limits for change should be based on test–retest comparisons that are made 1 month apart for pa-
tients whose condition is believed to be stable. The distribution of discrepancies between such paired comparisons would not only reflect the variance of the clinician’s ability to make consistent judgments and the consistency of the measuring instrument but would also capture variance due to month-to-month fluctuations of that parameter within a population of individuals considered to be normal.

Confidence limits for change may be set on a one-tailed or two-tailed basis. In this study, we have referred to two-tailed confidence limits which are appropriate to use when there is concern about the patient’s score changing in either the upward or downward direction. It is more common, however, for clinicians to be concerned about change occurring in one direction, in which case it becomes appropriate to apply one-tailed confidence limits. Table 2 (see Appendix for full description) may be used to estimate either one-tailed or two-tailed confidence limits.

When clinical measurements are made for the purpose of detecting difference from an expected value, it is desirable to have narrow confidence limits. These may be markedly affected by the size of the increment on the measuring scale. In our view, if it is desirable to have moderate sensitivity in a clinical test, the size of the scale increments should not exceed one standard deviation of the discrepancy. Thus, for any measure for which moderate sensitivity is desirable, the concordance of paired comparisons should not exceed 37%. For fine clinical sensitivity, the size of the scale increments should not exceed one third of the standard deviation of the discrepancy, in which case the concordance of paired comparisons should not exceed 13%. If commonly used grading scales are found to be excessively coarse, it is often easy to interpolate between scale divisions. In general, we would recommend decimalization, dividing the increment into 10 equal parts to interpolate between the integer units. An alternative is to use qualified integer systems (eg, grade 2−, 3+, 1+++). This practice usually represents an improvement of the integer scale, but firm rules need to be established about the number of scale increments per integer value. This allowable number may be either three (eg, 2, 2+, 3−) or five (eg, 2, 2+, 2++, 3−, 3−). There is less risk of ambiguity or inconsistency if decimalization is used rather than a qualified integer scale. For the experiment with the grading of nuclear opacity, the graders reported some initial reluctance to use the decimal grading system because they were rarely confident that they had assigned the most appropriate decimal score, whereas with the integer scoring system they felt highly confident of their integer score most of the time. A little experience was required before the graders became comfortable decimalizing their gradings, and they soon found it easier to assign decimal values than integer values.

The principles discussed here relating to scaling of clinical measurements can be applied to a broad range of clinical measurements. In qualitative grading systems, it is likely that the reference benchmarks of the scale are not evenly spaced. Furthermore, for both qualitative and quantitative grading systems, the true standard deviation of the discrepancy is likely to show some variation across the range of the scale. Although these factors contribute uncertainty to the statistical considerations, the methods described here will still provide a reasonably good basis for predicting relationships between the size of the scale increments and the confidence limits for change. Here, we have dealt with two examples, visual acuity measurements and the grading of nuclear opacity. However, equivalent examination of scaling could be applied to many other clinically relevant parameters such as cup-disc ratios, the severity of diabetic retinopathy, anterior chamber depths, the measurement of refractive error, curvature increments for fitting contact lenses, or the proportion of pupillary area occupied by posterior subcapsular cataract. Finer scaling can substantially enhance the ability of the clinician or the clinical researcher to detect smaller degrees of change.

Fig. 5. Graphic representation of probabilities integrated when determining the frequency with which the discrepancy between Observation B and Observation A is equal to d. The two terms in the general equation are represented by the checked and cross-hatched areas, respectively.
Table 2. Frequency distributions of discrepancies between paired observations

<table>
<thead>
<tr>
<th>Increment size (R)</th>
<th>Discrepancy—in number (d) of scale increments</th>
<th>Confidence limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>01234567891011121314151617181920</td>
<td>Increment size (R) 95% CL's d S.D.s</td>
</tr>
<tr>
<td>.100</td>
<td>.040 .039 .038 .037 .035 .033 .031 .029 .027 .024 .022 .019 .017 .015 .013 .011 .009 .008 .007 .005</td>
<td>.100 21 2.10</td>
</tr>
<tr>
<td>.125</td>
<td>.050 .049 .048 .046 .044 .041 .038 .034 .030 .026 .023 .019 .016 .013 .011 .009 .007 .005 .004 .003 .002</td>
<td>.125 17 2.12</td>
</tr>
<tr>
<td>.160</td>
<td>.064 .063 .061 .057 .052 .046 .040 .034 .028 .023 .018 .014 .010 .007 .005 .004 .002 .002 .001 .001</td>
<td>.160 13 2.08</td>
</tr>
<tr>
<td>.200</td>
<td>.080 .078 .073 .067 .058 .048 .039 .030 .022 .016 .011 .007 .005 .004 .002 .001</td>
<td>.200 11 2.20</td>
</tr>
<tr>
<td>.250</td>
<td>.099 .096 .087 .075 .060 .045 .033 .022 .014 .008 .005 .002 .001</td>
<td>.250 9 2.25</td>
</tr>
<tr>
<td>.315</td>
<td>.127 .120 .103 .080 .057 .036 .021 .011 .005 .002 .001</td>
<td>.315 7 2.21</td>
</tr>
<tr>
<td>.400</td>
<td>.157 .146 .115 .078 .045 .022 .010 .003 .001</td>
<td>.400 6 2.40</td>
</tr>
<tr>
<td>.500</td>
<td>.195 .173 .121 .066 .029 .010 .003</td>
<td>.500 5 2.50</td>
</tr>
<tr>
<td>.800</td>
<td>.303 .227 .096 .022 .003</td>
<td>.800 4 3.20</td>
</tr>
<tr>
<td>1.00</td>
<td>.369 .241 .067 .008</td>
<td>1.00 3 3.00</td>
</tr>
<tr>
<td>1.25</td>
<td>.443 .240 .037 .002</td>
<td>1.25 3 3.75</td>
</tr>
<tr>
<td>1.60</td>
<td>.530 .220 .014</td>
<td>1.60 2 3.20</td>
</tr>
<tr>
<td>2.00</td>
<td>.610 .191 .004</td>
<td>2.00 2 4.00</td>
</tr>
<tr>
<td>2.50</td>
<td>.682 .158 .001</td>
<td>2.50 2 5.00</td>
</tr>
<tr>
<td>3.15</td>
<td>.748 .126</td>
<td>3.15 2 6.30</td>
</tr>
<tr>
<td>4.00</td>
<td>.801 .100</td>
<td>4.00 2 8.00</td>
</tr>
<tr>
<td>5.00</td>
<td>.840 .080</td>
<td>5.00 2 10.0</td>
</tr>
<tr>
<td>6.30</td>
<td>.873 .063</td>
<td>6.30 2 12.5</td>
</tr>
<tr>
<td>8.00</td>
<td>.900 .050</td>
<td>8.00 2 16.0</td>
</tr>
<tr>
<td>10.0</td>
<td>.920 .040</td>
<td>10.0 2 20.0</td>
</tr>
<tr>
<td>12.5</td>
<td>.936 .032</td>
<td>12.5 2 25.0</td>
</tr>
<tr>
<td>16.0</td>
<td>.950 .025</td>
<td>16.0 1 16.0</td>
</tr>
<tr>
<td>20.0</td>
<td>.960 .020</td>
<td>20.0 1 20.0</td>
</tr>
</tbody>
</table>

The predicted concordance (C) for a range of increment sizes (R); multiples of the true standard deviation of the discrepancy. The predicted frequency distribution of discrepancies for each value of R is also tabulated. The table shows only positive discrepancies and it includes only frequency values that are 0.001 or greater and discrepancies up to 20 scale increments. Also shown are the width of the 95% confidence limits in both multiples of the true standard deviation of the discrepancy (SDs) and in terms of the number of scale increments (d). Hence, number of SDs = R \times d.
Key words: clinical grading, sensitivity to change, confidence limits, visual acuity, cataract grading, clinical trials

Appendix

This appendix provides a more detailed description of the mathematical bases of the statistical model used in this article. The expected concordance $C$, or frequency with which comparisons between paired observations show perfect agreement ($d = 0$), is given by the equation:

$$u_{d=0} = C = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{0} \left(1 - \frac{z}{R}ight) e^{-\frac{z^2}{2}} dz$$

where $R$ is the ratio of the scale increment size to the size of the true standard deviation of the discrepancy distribution.

The frequency ($u_d$) with which the value of Observation $A$ exceeds Observation $B$ by $d$ scale increments is given by the general equation

$$u_d = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d} e^{-\frac{z^2}{2}} dz + \frac{1}{\sqrt{2\pi}} \int_{d}^{\infty} e^{-\frac{z^2}{2}} dz$$

In this equation, the two terms represent integrated probabilities that correspond to the regions shown in Figure 5 by the checked and the cross-hatched areas, respectively.

The general equation has been used to generate Table 2, which shows the expected concordance and distribution of discrepancy values for a range of values of $R$ between 0.1 and 20. From such tables the 95% confidence limits may be determined for a range of values of $R$ (Fig. 2B). These confidence limits are expressed as a function of the true standard deviation of the discrepancy. The 95% confidence limits are calculated by determining the discrepancy value that is immediately beyond that which includes the 95th percentile. For example, for a value of $R = 1.60$, there would be a discrepancy of ±1 scale increment for the 54th to the 97th percentiles (from table: 0.530 + [2 X 0.220] = 0.97). Therefore, the 95% confidence limits would be set at two scale increments ($d = 2$). This represents 3.2 true standard deviations. The dotted line in Table 2 shows the locations of the 95% confidence limits for change.

The graph of 95% confidence limits as a function of $R$ is jagged (Fig. 2B). This is a product of using incremental scales and will lead to instances in which slightly coarser scale increments will lead to narrower 95% confidence limits. This can be most easily understood by considering that if the scale increment is 16 times the true standard deviation ($R = 16$), perfect concordance will be expected for 95% of paired comparisons. Thus, for this and all larger scale increments, a discrepancy of ±1 scale unit will be taken as exceeding the 95% confidence limit. For smaller scale increments, the confidence limits will be set at 2 or more scale increments. For example, a scale increment equal to 12.5 standard deviations that Table 2 shows to produce 93.6% perfect concordance. The 95th percentile will show a discrepancy of ±1 scale increment. Here, the 95% confidence limits would be set at a discrepancies of ±2 scale increments, representing 25 standard deviations, substantially greater than the confidence limits (16 standard deviations) when the scale increment was 16 standard deviations.

Similarly, when the confidence limits for $R = 1.25$ and $R = 1.60$ are compared, Table 2 shows that for a value of $R = 1.25$, 44.3% of paired comparisons should be concordant (ie, $d = 0$), 48.0% of comparisons would show a discrepancy of ±1 increment (.240 X 2), and 7.4% would show a discrepancy of ±2 increments (.037 X 2). Hence, 99.7% of comparisons would show discrepancies of ±2 increments or less. The confidence limits would be set at ±3 increments, equivalent to 3.75 true standard deviations of the discrepancy ($3 \times 1.25$). For a value of $R$ of 1.60, it was shown above that the confidence limits should be set at 3.2 standard deviations. The effect of the jaggedness of the function shown in Figure 2B becomes relatively negligible when $R$ is small (<0.3). In practice, the true standard deviation cannot be determined with absolute precision, so for a given scale increment, there will always be some uncertainty of the value of $R$ that it represents.

References


