Classification of radiological appearance and the derivation of a numerical score

A. J. FOX*
Employment Medical Advisory Service, Department of Employment, Baynards House, London

Fox, A. J. (1975). British Journal of Industrial Medicine, 32, 273-282. Classification of radiological appearance and the derivation of a numerical score. Approaches to the use of the International Labour Office U/C Classification of radiological appearances are considered in the context of multiple readers recording different degrees of abnormality. The Panel on Survey Radiology has been set up to provide a basis for standardization. An example of how the panel may be used in this role is given. The approach will depend on the standards maintained by the panel and on a greater understanding of the appearances associated with differing exposures.

The image obtained when a chest is interposed between a sensitive film and an x-ray source depends among other things on the spectrum and energy of the radiation source, the absorbing, scattering, and secondary radiation produced by integument, skeleton and viscera, as well as effects of posture, quality of film, and development techniques. On inhalation, dust and fumes may be retained and deposited locally in the lungs. Local or diffuse changes may occur subsequently in lung parenchyma or pleura. The extent and nature of response vary between industries with the dose and physico-chemical nature of the dust. The host response will also depend on the sex and age of the subject. Consequently, a straightforward dose-response relationship may not always exist. The primary dust effect may result in secondary distortion of airways and embarrassment of cardiopulmonary haemodynamics which may not be seen radiographically. The visualization of discrete small opacities in the lung fields does not necessarily imply that each represents the image of a particular pathological focus either of dust or of fibrous tissue. Identical small opacities in films obtained from different industries or within the same industry do not imply identical underlying pathological processes. The severity of disease is not necessarily reflected in the profusion of small opacities. Nevertheless, within certain industries, dust measurements in air and in lung (post mortem) have been shown to be related to the radiographic profusion and even the type of small opacities (Rivers et al., 1960).

Aims of classification

A classification of radiographic appearances should record relevant information contained in the radiograph. Conflict often arises in the definition of 'relevant' because the definition must depend on the use that is to be made of the information. Clinicians, with their responsibility to the examinee, aim to extract different information from that sought by the epidemiologist. In parallel with the changing attitude and new advances in many aspects of clinical medicine, radiological decision making is changing. In the past, the clinical diagnosis was primarily concerned with the advanced stages of disease. Currently, the identification and quantification of relatively early changes are of great importance in order that advice concerning continued exposure to potential hazards may be given to patients. Epidemiologists have long been concerned with the quantification of all the changes related to exposure to potential hazards. Emphasis is now placed on the

*New address: Office of Population Censuses and Surveys, Medical Statistics Division, Kingsway, London WC2B 6JP.
detection of the changes occurring before disease is irreversible.

**Background to International Labour Office (ILO) U/C International Classification (1971)**

Davies and Mann (1949) postulated that the number of small opacities seen is the radiological equivalent of the severity of simple pneumoconiosis. Categories of simple pneumoconiosis were introduced corresponding closely to the International Labour Office (1970) categories in which the profusion of small opacities is partitioned into four ranges. Fletcher and Oldham (1949), when studying the problem of consistent radiological diagnosis in coalworkers' pneumoconiosis, suggested that consistency could be improved by introducing accepted standard films with which the films under study might be compared. The committee of experts (International Labour Office, 1958) recommended the introduction of accepted standard films when they reviewed the scheme originally introduced in 1950. It was shortly after this meeting that standard films were introduced as an aid to film readers. Although in the original experiments with standard films it was thought that films depicting the boundaries between categories would be of most benefit, the ILO introduced films which were 'typical' of each category. It was only in the most recent set of standard films that a 'normal' film was introduced.

In 1968, in the light of the experience in the preceding decade, the ILO reviewed the classification and proposed that both a short and an extended classification be introduced (ILO, 1970). The short classification was introduced because it was felt that the detail required for some epidemiological purposes, where, for example, the radiographic appearances were to be related to indices of dust exposure, different aspects of pulmonary function, and other clinical features, was greater than that required by the majority of clinicians.

The initial introduction of a classification was primarily concerned with the radiological description of coalworkers' pneumoconiosis. Modifications to the original scheme have been made on the basis of experience gained because there has been a progressive widening of the types of pneumoconiosis covered, and because more features required detailed description.

In 1964 the International Union Against Cancer (UICC, 1965) recommended that to facilitate studies of asbestos workers 'a scheme based if possible on an extension of the International Labour Office (ILO) 1958 classification be developed'. The resultant classification was published as the UICC/Cincinnati in 1970.

The ILO U/C International Classification 1971 modifies the ILO (1970) scheme making the short classification compatible with the extended classification and incorporating improvements based on experience with the UICC/Cincinnati scheme over the preceding years.

The history of the development of the current classification indicates how it has been extended to incorporate different types of appearance that have in the past been associated with exposure to different dust hazards. Recently the ILO (1970) described the scope of the classification as 'International classification of persistent radiological opacities in the lung field provoked by the inhalation of mineral dusts, including coal and carbon'. The 1971 revision (ILO, 1972) extends its scope to cover 'all types of mineral dusts, including silica, coal, asbestos, and beryllium'. In extending the classification, the description of small opacities has been separated into small rounded and small irregular opacities. In the original classification, rounded opacities were graded whereas irregular opacities were described as increased lung markings (Z) or linear shadows (L). The type of opacity, whether rounded or irregular, is separated into three further subgroups; the locations of opacities are described; and other features such as pleural thickening, pleural calcification, and ill-defined diaphragm are incorporated. Rossiter (1972a) reported a trial to evaluate the repeatability of the UICC/Cincinnati classification. His observations concerning observer variation are similar to those reported by Liddell and May (1966) who were primarily concerned with coalworkers.

**Inter- and intra-observer variation**

Much early work on the use of these classifications was aimed at measuring observer variation and subsequently methods of reading standard films were introduced in order to reduce this variation. Observer variation was primarily separated into observer variability and observer error (Liddell and May, 1966). Observer variability is concerned with the variations in categorization between observers in the readings of individual films. In the present context, this can be separated into inter-observer variation and intra-observer variation. The former is the variation that arises from using different observers, whereas the latter arises using the same observer at different points in time. The initial study of observer variation in the classification of radiographic appearances was that by Fletcher and Oldham (1949). Ashford (1959) and Fay and Ashford (1960) discussed two further models for measuring observer variation. The first model is based on the assumption that a film taken at random from a particular category has a constant probability, dependent on the observer, of being placed in a specific second category. The performance of readers can consequently be monitored by estimating the matrices of probabilities. This model suffers from the disadvantage of applying only to average films and the assumptions...
are such that it manifestly does not provide an entirely realistic representation of the reading process on any particular film. The second model introduced by Fay and Ashford suggested that the disadvantages of the first model might be overcome if each film were considered to represent 'a unique degree of abnormality located on a continuum, or abnormality scale, which covers the whole range of simple pneumoconiosis'. Rivers et al. (1960) also suggested the idea of a continuous scale of abnormality. Fig. 1 illustrates the relationship between the various scales that have been considered. Starting with the ILO categories, the assumption that each of the categories was of uniform width gave rise to the x-scale. Rivers et al. (1960) suggested that if the x-scale were converted to a new y-scale using the transformation

\[ y = 1 + \frac{1}{4} (x^2 - x^2) \]

the distribution of radiographs would be approximately Gaussian (normal). Fay and Ashford (1960) assumed a continuum and a normal distribution of abnormality. As the choice of normal distribution does not affect the results, they defined the boundary between category 0 and category 1 and that between categories 1 and 2 as being at 0.0 and 1.0 respectively and then permitted the cumulative percentage of films in each category to define the boundary between category 2 and category 3 at \( x \). The method was used by Fay and Ashford to measure observer variation, although the validity of its use for this purpose was subsequently questioned by Wise and Oldham (1963).

In 1963 Liddell noted that readers would often describe films as 'high category 2' or 'low category 1'. He set up an experiment to measure whether or not it would be valuable to record these comments in an organized way. The result of the experiment (Liddell, 1963) was the National Coal Board (NCB) elaboration of the ILO classification suggested by Liddell and Lindars (1969). The current NCB elaboration separates the scale into 12 categories: 0/-, 0/0, 0/1, 1/0, 1/1, 2/1, 2/2, 2/3, 3/2, 3/3, 3/4. The elaborated classification provides a more sensitive measure of disease than the original classification. It was consequently incorporated by the ILO into the 1968 and subsequent classifications.

**Use of the classification in different industries**

The early studies involving descriptions of radiographs concentrated primarily on the problem with respect to coalworkers' pneumoconiosis. The radiographs were used to measure the degree of simple pneumoconiosis and to distinguish those subjects who had appearances consistent with progressive massive fibrosis (PMF). Since the degree of simple coalworkers' pneumoconiosis was related to the profusion of small rounded opacities, the studies considering observer variation have concentrated on this aspect of the classification. The classification or one of its derivatives has been used in various industries. Lloyd Davies (1971), Harries et al. (1972), Rossiter et al. (1972), and Lloyd Davies et al. (1973) in particular have studied foundry workers, asbestos workers, and granite workers. The radiographic appearance has not been adequately related to lung pathology or to other measures of disease for these industries. Some studies have been reported relating the profusion of small rounded opacities in coal workers to their pathology at death (Gough, 1947; Rivers et al., 1960; Rossiter, 1972b). In view of the range of radiological features currently recorded, more extensive radiological and pathological studies incorporating measures of progression are required.

Various methods (Rivers et al., 1960; Fay and Ashford, 1960; Wise and Oldham, 1963; Oldham, 1971, Rossiter, 1972b) have been used to summarize the profusion of small opacities in the form of a score on a continuum and the resultant scores have been related to various measures of exposure and effect. Since the classification has been applied to different industries without adequate validation, the significance of these scores has not been fully understood. In particular, little use has been made of the differentiation between small rounded and small irregular opacities. In the current classification the reader is given the option of recording the combined profusion, although standard films with which to compare are not provided.

Throughout the past two decades studies have used more than one reader to classify films in order to minimize the effect of observer variation on the results of the survey. Although this purpose is served, new problems are created when the readers do not agree on the classification of individual films. In early studies using multiple readers, agreed readings were obtained. A move was subsequently made to reading randomized batches of films independently.

Despite warnings by Fay and Ashford (1960) that direct averages of different readers should not be used, Oldham (1971) and Rossiter (1972b) introduced methods which were primarily based on a system of direct averaging. They also incorporated a method of weighting their results according to the degree of confidence in the score an individual
reader gave. The degree of confidence was not based on an assessment made by the reader but on the width of the category in which the film was placed. This approach can introduce apparent anomalies in the derived scores. For example, if few films were placed in category 1/0 as compared with category 1/1, a film read 0/0, 0/0, 1/0 might receive a higher score than one read 0/0, 0/0, 1/1 (as in Lloyd Davies et al., 1973).

**Pottery workers survey**

A survey of approximately 10% of the pottery industry, stratified according to product, was recently undertaken by the Employment Medical Advisory Service (HM Chief Inspector of Factories, 1971). A member of the Stoke-on-Trent Pneumoconiosis Medical Panel, a chest physician, and an Employment Medical Adviser were trained to read radiographs according to the ILO U/C International Classification 1971. The readers were subjected to three day-long training sessions to become familiar with the classification. Radiographs of 5714 subjects, 86.3% of the nominal roll, were read by these three readers (1, 2, and 3) and questionnaires were accurately completed. The radiographs were read independently in a different random order by each reader. As in Lloyd Davies et al. (1973), the maximum profusion of each reading of a film is recorded as the profusion of small opacities. Fig. 2 gives the distribution of films placed in each category by each reader. The distributions for readers 1 and 2 appear to be smooth.

**Fig. 2. Distribution of the maximum profusion for individual readers.**

![Figure 2](http://oem.bmj.com/)

**Fig. 3. An illustration on an arbitrary scale of the relative ranges and locations of the ILO (NCB elaboration) categories as seen by each of the three readers.**

![Figure 3](http://oem.bmj.com/)
curves, whereas reader 3 shows sharp rises for mid-categories. Reader 2 read films further along the continuum than either of the other two. This was partially due to an anomaly in the instructions concerning the use of category 0/1.

In order to derive a score, it will be assumed that the distribution of profusions in this population is continuous and normal. Having made the assumption of normality, the choice of parameters is a matter of convenience. In Fig. 3 the scale derived for each reader is shown transformed by a fixed normal. The borders for the ILO categories on these scales are estimated from the cumulative functions for each reader, using probability graph paper and a straight line drawn across the page. The line must be the same for each reader, otherwise different distributions are being fitted. The line fitted in this instance connects the bottom left corner with the top right corner of the particular print of probability graph paper used (Fig. 4).

Since category 3/4 was not recorded and reader 2 did not record 0/-, the upper boundaries of category 3/3 have not been shown in Fig. 3 and the lower boundary of category 0/0 for reader 2 is also not shown. This representation of readers' distributions on a scale demonstrates the same properties as the histograms shown in Fig. 2. In particular, it is simple to see that, because reader 2 placed the boundary between category 0/0 and category 0/1 to the left of the position used by readers 1 and 3, he saw more films above that border. Similarly, reader 3 demonstrates his reluctance to use outside categories, since these are narrow on his scale.

We had no measure of how these readers compared with other readers and were therefore reluctant either to take a direct average or to ignore some of the readings. The search for a single index is pursued primarily to facilitate the presentation of the results. The problem is probably exaggerated here because of the choice of readers. The viewpoints are those of the clinician, the doctor responsible for compensation, and the epidemiologist. Different criteria are important in each of the fields and each is used to reading different degrees of abnormality.

Panel on Survey Radiology
In 1971 the Department of Employment and the Medical Research Council jointly established a reference panel to which samples of films from different surveys might be submitted and adjustments made for reader differences. The Panel on Survey Radiology (PSR) was selected to encompass a wide range of experience and viewpoints. A list of members is given in Appendix 1. The functions of the panel are:

1. to standardize the classification of films, eliminating the dependence on individual readers;
2. to render the findings in different studies comparable;
3. to ensure the comparability of studies at different points in time;
4. to act as an arbiter on national problems, for example, those of radiological technique and interpretation; and
5. to provide a national view in international discussions.

The panel would not be used as an arbiter on unusual films and would not be used as a reference panel for legal decisions.

After early meetings at which the discussion was mainly concerned with how the panel should operate and initial questions about the effect of film quality and technique, an introduction to the ILO U/C International Classification 1971 was given. Subsequently, the panel read batches of films. The first main reading session was organized in April 1973 at which approximately 300 films from the pottery survey were read by 12 of the 13 readers. The readers did not know the origin of the films, which were supplemented by films of asbestos, granite, and foundry workers, as well as a few films demonstrating the appearance of cryptogenic fibrosing alveolitis.

As above, the cumulative distribution of pottery
FIG. 5. Relative ranges and locations of NCB elaborated ILO categories on an arbitrary scale of abnormality as recorded by members of the panel on survey radiology for a sample of 300 pottery films.

### TABLE 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Reader 1</th>
<th>Reader 2</th>
<th>Reader 3</th>
<th>Reader 4</th>
<th>Reader 5</th>
<th>Reader 6</th>
<th>Reader 7</th>
<th>Reader 8</th>
<th>Reader 9</th>
<th>Reader 10</th>
<th>Reader 11</th>
<th>Reader 12</th>
<th>Overall Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>6-90</td>
</tr>
<tr>
<td>0/1</td>
<td>9-12</td>
<td>9-18</td>
<td>9-23</td>
<td>9-30</td>
<td>9-35</td>
<td>9-40</td>
<td>9-45</td>
<td>9-50</td>
<td>9-55</td>
<td>9-60</td>
<td>9-65</td>
<td>9-70</td>
<td>9-75</td>
</tr>
<tr>
<td>2/1</td>
<td>11-96</td>
<td>12-50</td>
<td>11-102</td>
<td>12-38</td>
<td>12-63</td>
<td>12-88</td>
<td>12-113</td>
<td>12-139</td>
<td>12-166</td>
<td>12-193</td>
<td>12-221</td>
<td>12-250</td>
<td>12-281</td>
</tr>
<tr>
<td>2/2</td>
<td>12-60</td>
<td>13-00</td>
<td>12-127</td>
<td>12-52</td>
<td>12-87</td>
<td>12-116</td>
<td>12-144</td>
<td>12-173</td>
<td>12-201</td>
<td>12-230</td>
<td>12-259</td>
<td>12-289</td>
<td>12-322</td>
</tr>
<tr>
<td>3/3</td>
<td>14-28</td>
<td>13-28</td>
<td>14-32</td>
<td>14-23</td>
<td>14-06</td>
<td>14-03</td>
<td>14-02</td>
<td>14-01</td>
<td>14-00</td>
<td>14-00</td>
<td>14-00</td>
<td>14-00</td>
<td>14-00</td>
</tr>
</tbody>
</table>

Note: The table includes scores for each reader in each category, with the overall score calculated as the average of all scores for that category.
TABLE 2

Scores for Each Pottery Reader Based on the General Scale and the Assumption of a Normal Distribution Over This Scale

<table>
<thead>
<tr>
<th>Category</th>
<th>Reader 1</th>
<th>Reader 2</th>
<th>Reader 3</th>
<th>Overall readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td></td>
<td>0/-</td>
<td>1/6</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td></td>
<td>0/1</td>
<td>9/28</td>
<td>5/62</td>
<td>9/36</td>
</tr>
<tr>
<td>1</td>
<td>1/0</td>
<td>11/00</td>
<td>9/06</td>
<td>10/27</td>
</tr>
<tr>
<td></td>
<td>1/1</td>
<td>11/81</td>
<td>10/16</td>
<td>11/03</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>12/40</td>
<td>11/24</td>
<td>12/17</td>
</tr>
<tr>
<td>2</td>
<td>2/1</td>
<td>13/00</td>
<td>12/07</td>
<td>12/38</td>
</tr>
<tr>
<td></td>
<td>2/2</td>
<td>13/78</td>
<td>12/79</td>
<td>12/92</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>14/41</td>
<td>13/82</td>
<td>14/19</td>
</tr>
<tr>
<td>3</td>
<td>3/2</td>
<td>15/02</td>
<td>14/19</td>
<td>14/36</td>
</tr>
<tr>
<td></td>
<td>3/3</td>
<td>16/00</td>
<td>14/83</td>
<td>14/80</td>
</tr>
<tr>
<td></td>
<td>3/4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

workers' films defined a scale for each member of the panel and the overall distribution ignoring reader differences defined an overall scale for the panel. Fig. 5 shows the scales that were derived and the location of category boundaries for each reader. Scores for films placed in each subcategory by each reader were estimated from the cumulative distributions using the median percentage in that category. Table 1 gives the scores for each reader in the Panel on Survey Radiology and Table 2 gives those for the readers of pottery workers’ films. The final column in Table 2 shows the scores derived from the overall distribution of films read by these readers.

The panel score for a film read by the PSR was estimated from the readings of its members. In Appendix 2 a method is given for adjusting the scales of individual readers of pottery workers' films so that they compare better with the panel score. The approach is to look at all films placed by each of the readers 1, 2, and 3 in each of the ILO categories. The distribution of panel scores for films in each category is then considered as an estimate of the relationship between each reader's interpretation of the category boundaries and the PSR scale. A method of averaging is described and a method of weighting also suggested.

**Comparison of methods**

Consider fitting a normal distribution to all readings and then taking a direct average of the three readings of a particular film weighted according to the inverse of category widths (method 1). If the readers’ distributions are reasonably similar, the scores are based on the overall distribution (last column in Table 2). The results using this approach were compared with those using a different scale for each reader (method 2) and with the estimated panel score. (It should be noted that the comparison was made using equation 10, not equation 11, see Appendix 2). A film read as 0/0, 0/0, and 0/0 by readers 1, 2, and 3 respectively would be scored as \((6.78 + 5.62 + 6.98) \div 3\) by method 2 and as 6.45 by method 1. The 300 films read by the PSR were scored by each method. It should be noted that each film was read by at least nine readers. The correlations between the three methods and the panel scores were 0.87 for method 1, 0.86 for method 2, and 0.90 for the estimated panel scores. The mean errors were \(-0.0623\) for method 1, \(-0.1378\) for method 2, and \(-0.0028\) for the estimated panel scores. The standard deviations of the errors were \(\pm 0.8153\), \(\pm 0.8708\), and \(\pm 0.6076\) respectively.

Since the estimated panel score is based on the same films as those read by the PSR, it is not entirely surprising that the evidence favours this method. A more valuable test would have been to compare the estimated scores with the panel scores on a second series of pottery films. The PSR has as yet not read a second set of pottery films as it is engaged in other trials of its bias and consistency. This paper is intended only to indicate that an improvement in comparability between studies might be obtained using the PSR. Further evidence on the ability of the PSR to maintain its standard and to measure the improvement in comparability is still required.
Discussion and conclusions

The ILO U/C classification has gone a long way towards establishing international standards for reading radiographs. It has, nevertheless, been thought necessary to create a reference panel to supplement the findings of small groups of readers. The panel representing various aspects of radiology should not be subject to the wide individual variations previously observed. It should consequently provide a sounder basis for comparisons between different studies.

This is the first time the PSR has been employed to define a scale. Many of the assumptions inherent in the use of the panel, particularly in terms of its ability to maintain its standard, can be evaluated only by its continued use. An initial attempt to quantify the changes that might be built into a panel such as this is being made by analysing repeat readings on 150 of the above films.

The interest of the statistician is such that he not only wants a scientific method of analysing his data but he also needs to be able to convey the results to his customers, the medical practitioners in this field. The ILO U/C scale is well known and understood, but this is not the case for scales such as those described in this paper. The approaches to radiograph scoring fall into two distinct classes. The first is concerned with the use of a simple classification. The method of obtaining a consensus reading affects the meaning of the simple categories and consequently affects comparisons between studies.

The second approach arises from the desire to measure exactly the changes that are occurring. It is believed that the step-wise model as represented by categories 0, 1, 2, or 3 is perhaps an oversimplification of these changes. The simple classifications are particularly helpful in prevalence studies and in studies of the interrelationships between radiographic changes and other factors. They are of less value in longitudinal studies. The statisticians have consequently sought an approach using the concept of change being a continuous process.

Since the methods described all involve fitting distributions to observations that are discrete, the discrete scale can be regenerated after a consensus reading has been obtained. For the method described here, a film can be placed in the relevant ILO category by referring the score to the scale defined in Fig. 5. This method is preferred to those previously used because the PSR lays the foundations for a more objective average and results from different surveys can more easily be compared.

When readers differ as widely as they did in the study of pottery workers, one might well question whether they know what they are reading. This assumption is fundamental whichever method of analysis is adopted. The solution is found partially in the training they received and in the frequency with which they read. An absolute answer to this question can, however, be obtained only from a post mortem comparison of recorded appearances with lung pathology.

The method of scoring advocated for the analysis of data utilizes an arbitrary scale supposedly measuring abnormality. Reports should be made in terms of ILO categories which form the basis of comparison. Problems exist in interpreting the significance of changes from one point in the scale to a second point but these cannot be solved until these changes have been related to other measures of function/dysfunction and to measures of exposure. This paper describes the development of a more objective scale. In the analysis of survey data it will be necessary to develop a similar scale for each type of appearance, including large opacities, pleural thickening, and the other appearances recorded.

The author wishes to acknowledge his indebtedness to Dr T. A. Lloyd Davies for his advice, and with Dr J. C. Gilson, for setting up the Panel on Survey Radiology to Dr E. Posner, Dr M. K. Coles, and Dr M. Greenberg for patiently reading over 6000 pottery survey radiographs; and to members of the PSR for providing a basis for comparison. Professor J. R. Ashford, Dr P. D. Oldham, and Mr M. Jacobsen, made many helpful comments and criticisms in discussion as well as on early drafts of this paper. Mr P. F. Collier ably assisted in the statistical analyses.

References


APPENDIX 1

PANEL ON SURVEY RADIOLoGY

Dr J. C. Gilson—Director, Pneumoconiosis Unit
(Chairman)
Dr C. S. Darke—Consultant chest physician
Dr J. A. Dick—Medical Officer, National Coal Board
Dr A. C. Douglas—Consultant chest physician
Dr M. Greenberg—Senior Employment Medical Adviser

APPENDIX 2

Consider the score associated with the Panel on Survey Radiology reader i placing a category j as \( s_{p(i)} \) and the corresponding scores for the pottery survey readers as \( s_{s(i)} \). The \( s_{p(i)} \) were estimated from the films read by each member of the Panel on Survey Radiology and \( s_{s(i)} \) were estimated from all pottery films read by the panel of three readers. The true score for a particular film (k) is \( s_{t(k)} \). For the films read by the Panel on Survey Radiology the true score was estimated by

\[
S_{p(k)} = \frac{i=1}{12} (\delta_{ki}) \frac{j=1}{12} (\delta_{kj}) \quad (1)
\]

where

\[
\delta_{ki} = \begin{cases} 0 & \text{if reader } i \text{ did not read film } k \\ 1 & \text{if reader } i \text{ read film } k \end{cases} \quad (2)
\]

and \( s_{p(i)}(k) \) is the score associated with the category in which the \( i \)th reader placed film \( k \).

The distribution of true scores for films read as category j by Pottery Survey reader i was estimated for each j and i from the films read by the Panel on Survey Radiology. The mean true scores \( s_{t(ij)} \) and the standard deviations \( \sigma_{s(ij)} \) for pottery readers are estimated from

\[
S_{t(ij)} = \sum_{k} \delta_{*ki} S_{p(k)}/n_{ij} \quad (3)
\]

and

\[
\sigma_{s(ij)} = \left[ \left( \sum_{k} (\delta_{*ki} S_{p(k)^2} - n_{ij}S_{t(ij)^2}) / (n_{ij} - 1) \right) \right] \quad (4)
\]

where

\[
n_{ij} = \sum_{k} \delta_{*ki} \quad (5)
\]

and

\[
\delta_{*ki} = \begin{cases} 0 & \text{if reader } i \text{ did not place film } k \text{ in sub-category } j \\ 1 & \text{if reader } i \text{ placed film } k \text{ in sub-category } j \end{cases} \quad (6)
\]
Estimates for some categories were not available because no films in the samples read by the Panel on Survey Radiology were placed in these categories by the pottery survey readers. If estimates of the neighbouring categories were available the estimate for that category was

\[ S_{sij} = S_{si(j-1)} + \frac{(s_{sij} - s_{si(j-1)})}{(s_{si(j+1)} - s_{si(j-1)})} (S_{si(j+1)} - S_{si(j-1)}) \]  

(7)

If no estimates were available for the upper boundary the estimate used was

\[ S_{sij} = S \]  

(8)

Where \( S \) is the score allocated on the true scale, that defined by the overall distribution from the Panel on Survey Radiology. Table 3 gives the estimated means and standard deviations associated with each sub-category for each of the three pottery readers. The true score estimated for film \( k \) from each reader \( i \) was taken to be

\[ S_{sij(k)} = S_{sij} \]  

(9)

if film \( k \) was placed in sub-category \( j \). The 'true score' estimated from the panel of pottery readers was the mean of the three readers' estimates,

\[ S_{sk} = \frac{1}{3} \sum_{i=1}^{3} S_{sij(k)} \]  

(10)

So for example a film read by the pottery survey readers as 0/0, 0/1, and 0/0 (readers 1, 2, and 3 respectively) was given a score of

\[ (7.36 + 7.61 + 7.37) / 3 = 7.45 \]  

(from Table 3).

It was noted above that Oldham (1971) introduced a method of weighting results according to the degree of confidence on the reader's score. The method introduced was to weight the average according to the inverse of the category width. We suggest that the category width is not only a reflection of the confidence in the score but primarily of the number of films in the category. The weighting suggested here is the inverse of \( \omega_{sij} \) as it is this that defines the variation associated with a particular score. If the variation is high this means that one can have little confidence in the score. Consequently one might estimate \( S_{sk} \) from the three readings from

\[ S_{sk} = \sum_{i=1}^{3} \frac{S_{sij(k)}}{\omega_{sij}} \]  

(11)

where

\[ \omega_{sij} = 1/\delta_{sij} \]  

(12)

and \( \delta_{sij} \) is given in equation (4) above.