HETA 90-0340-2659
A TRIAL OF ADDITIONAL COMPOSITE STANDARD RADIOGRAPHS FOR USE WITH THE ILO INTERNATIONAL CLASSIFICATION OF RADIOGRAPHS OF PNEUMOCONIOSES

National Institute for Occupational Safety and Health
Division of Respiratory Disease Studies
Morgantown, West Virginia, USA

Michael Jacobsen
William E. Miller
John E. Parker

PREFACE

The Hazard Evaluations and Technical Assistance Branch of NIOSH conducts field investigations of possible health hazards in the workplace. These investigations are conducted under the authority of Section 20(a)(6) of the Occupational Safety and Health Act of 1970, 29 U.S.C. 669(a)(6) which authorizes the Secretary of Health and Human Services, following a written request from any employer or authorized representative of employees, to determine whether any substance normally found in the place of employment has potentially toxic effects in such concentrations as used or found.

The Hazard Evaluations and Technical Assistance Branch also provides, upon request, technical and consultative assistance to Federal, State, and local agencies; labor; industry; and other groups or individuals to control occupational health hazards and to prevent related trauma and disease. Mention of company names or products does not constitute endorsement by the National Institute for Occupational Safety and Health.

ACKNOWLEDGMENTS AND AVAILABILITY OF REPORT

The following physicians generously donated their time and expertise to provide more than 14 000 classifications of radiographs that are described in this report:

Jacques Ameille
Raymond Begin
Ladislav Benický
Patrick Brochard
Keizo Chiyotani
Dominique Choudat
Alois David
Marc Desmeules
Kurt G. Hering
Marja-Liisa Kokko
Ossi Korhola
Aleksandra Kujawska
Yukinori Kusaka
N. LeRoy Lapp
Marc Letourneux
Li Guowei
Liu Shunging
Liu Yulin
Kazimierz Marek
W. Keith C. Morgan
Hiroshi Morikubo
David Muir
Peter Rathjen
Douglas Scarisbrick
Anthony Seaton
Hisao Shida
Steven Short
Klaus Siegmund
Jiří Slepička
Colin A. Soutar
Aleksander Stachura
František Staník
Andrzej Stasiow
Kristina M. Virkola
Mei-Lin Wang
Susan Weber
Volkmar Wiebe
Paul Willdig
Zhang Cuijuan
Anders J. Zitting

This report was prepared by Michael Jacobsen, William E. Miller and John E. Parker, of the Division of Respiratory Disease Studies (DRDS). Dr. Jacobsen is currently with the Institute of Occupational and Social Medicine at the University of Cologne, Germany.

This report is not copyrighted and may be freely reproduced. Single copies of this report will be available for a period of three years from the date of this report. To expedite your request, include a self-addressed mailing label along with your written request to:

NIOSH Publications Office
4676 Columbia Parkway
Cincinnati, Ohio 45226
800-356-4674

After this time, copies may be purchased from the National Technical Information Service (NTIS) at 5825 Port Royal Road, Springfield, Virginia 22161. Information regarding the NTIS stock number may be obtained from the NIOSH Publications Office at the Cincinnati address.

Health Hazard Evaluation Report 90-0340-2659

A TRIAL OF ADDITIONAL COMPOSITE STANDARD RADIOGRAPHS FOR USE WITH THE ILO INTERNATIONAL CLASSIFICATION OF RADIOGRAPHS OF PNEUMOCONIOSES

Morgantown, West Virginia, USA

Michael Jacobsen
William E. Miller
John E. Parker

SUMMARY

Background

The International Labour Office (ILO) convened a meeting in November 1989 to consider a possible revision of the ILO's International Classification of Radiographs of Pneumoconioses (ILO, 1980). Eleven invited participants, from seven countries, attended the meeting. Among the various suggestions considered was a proposal that sections of some of the standard radiographs that accompany the Classification should be reproduced as quadrants of new standard films. Those new 'QUAD' standards could then replace the full-sized standards involved, thus reducing the total number of full-size (35 x 35 cm) films in the set. This, it was thought, would increase the utility of the standard films and would thus encourage their more frequent use in epidemiological studies. It was agreed that the feasibility and possible effects of the proposed change should be studied by means of an appropriately designed film-reading trial. This report records results from that trial.

A draft protocol for the trial was prepared subsequently, was finalized after correspondence between the experts who had attended the meeting, and was agreed by 12 centers of excellence on pneumoconiosis, in 12 countries, who were invited by the ILO to participate in the trial. The work was administered jointly by the ILO's Occupational Safety and Health Branch in Geneva and the Division of Respiratory Disease Studies of the US National Institute for Occupational Safety and Health (NIOSH) at its Morgantown, West Virginia

laboratory. The Task Force on Pneumoconiosis of the American College of Radiology (ACR) collaborated with NIOSH scientists in the selection and reproduction of existing and modified standard chest radiographs, and of other radiographs needed for the trial. The particular sections of existing full-sized standards that might be suitable for reproduction as quadrant standards were agreed at the 1989 meeting. The spatial arrangement of those sections on the experimental QUAD standards was effected by NIOSH and ACR staff in the light of correspondence between the experts who had attended the meeting in 1989.

The aim of the trial was to determine whether, and to what extent, the proposed modification to the set of standard radiographs would affect film readers' classifications of chest radiographs of persons with histories of occupational dust exposure.

Methods

Thirty-nine physicians from 10 countries repeatedly classified 120 chest radiographs using two different sets of standard films. One of the sets (the "ILO" standards) was that associated with the current (ILO, 1980) edition of the Classification. The other set (the "QUAD" standards) included nine of the films in the existing (ILO) set plus five new (experimental) standard films that reproduce sections from the 13 ILO standards that were not included as full-sized images.

Nineteen of the readers used the ILO standards on the first occasion that they saw the 120 films and the QUAD standards on the second occasion ("Reading Sequence IQ"). The other 20 readers used the standards in the reverse order ("Sequence QI"). All IQ readers, and 18 of the QI readers reviewed the 120 films a third time; on this occasion they classified 60 of the 120 films using the ILO standards and the other 60 using the QUAD standards.

Main findings

Variability between readers in their assessments of small opacity profusion was generally similar when using the two sets of standards.

Variability in readers' repeated small opacity profusion classifications of the same films when using the same set of standards ("within-reader variability") was also similar overall for the two sets of standards, but the size of discrepancies between repeated classifications of the same films was less pronounced when using the QUADs.

Three quarters of 12,929 valid film classifications, from all three viewings of the films, indicated the presence of some small opacities (category 0/1 or higher). There was little difference in this respect between use of ILO and QUAD standards (74.4% and 74.7%, respectively, with category 0/1 or higher; 61.9% and 63.6% with category 1/0 or higher.)

Forty-four percent of 4591 pairs of classifications from the first two viewings of the 120 films showed identical small opacity profusion sub-categories when using the two sets of standards; 29% indicated higher levels of profusion when using the QUADs, and 27% had higher profusion scores when using the ILO standards. The net effect, averaged over all readers and all three viewings amounted to 3 higher sub-categories on the 12-point profusion scale per 100 films classified when using the QUAD standards. This corresponded to about 2% higher overall prevalence of small opacities on the 4-point scale (i.e., categories 1, 2 and 3 combined) when using the QUAD standards with this particular set of films.

The aggregate findings summarized above are the result of partly conflicting tendencies in sub-sets of the data defined by the trial-design-determined sequence of use of the standards, and by the shapes of the small opacities that were recognized.

> IQ readers recorded slightly higher profusion scores, on average, when using the ILO standards (2 sub-categories per 100 films classified). QI readers, however, averaged eight higher profusion sub-categories per 100 films classified when using the QUAD standards.

> Small opacities that were recognized as predominantly irregular in shape tended to be classified into higher profusion categories when assessed with the help of the QUAD standards. This phenomenon was evident irrespective of the sequence in which the standards were used and occurred primarily when viewing films with low small opacity profusion levels (categories 0/1 through 1/2). It explains QI readers' higher average profusion scores when using the QUAD standards because the QI readers identified a relatively high proportion (60%) of the small opacities that they saw as irregular in shape, as compared with 49% thus characterized by IQ readers.

There is some evidence that when IQ readers reviewed the films for the third time, their use of the QUAD standards was associated with higher profusion classifications than when they used the QUADs earlier, that is, at their second viewings of the films.

The distributions of readers' judgements about the predominant shapes of the small opacities that they recognized varied substantially, depending, in part, on the standards used and on the sequence of their use. The ratio of predominant shapes (irregular:rounded) when using the ILO standards was (46:54), on average, among IQ readers (i.e., those who used the ILO standards at the first viewings) and (59:41) for the QI readers. The corresponding ratios associated with use of the QUAD standards were (52:48) for IQ readers and (62:38) for QI readers. The (weighted) average ratios (irregular:rounded) for predominant shapes were (52:48) for the ILO standards and (57:43) for the QUAD standards.

Use of the QUAD standards was associated with a statistically significant lower frequency of classifications referring to the presence of large shadows: 4.5% of the QUAD-guided classifications as compared with 5.1% when using the ILO standards. A large part of this average difference is attributable to classifications from two readers who had used the ILO system relatively infrequently during the 12 months preceding the trial.

Conclusions

The proposed modification to the ILO standard radiographs, involving reproductions of sections from 15 of the ILO standards on five new "quadrant" films, would not increase variability between readers, might improve reproducibility of small opacity profusion classifications in some respects, but could also reduce slightly the frequency with which some readers identify large opacities. Use of the modified set of standards is likely to increase the frequency with which some readers describe the shapes of the small opacities that they see as predominantly irregular, rather than rounded. Classifications of small opacity profusion for films identified as predominantly irregular are likely to be higher using the modified standards when compared to the same classifications using the current ILO (1980) standards.

These results were detectable in the context of a controlled trial setting, involving a contrived high proportion of films classifiable as showing small opacities (about 60% with category 1 or more). The effects described are unlikely to be distinguishable from *inter*- and *intra*-reader variability in most real-life occupational health survey situations, where the prevalence of pneumoconiosis is usually less severe.

Keywords: Chest Radiograph, Crossover Trial, International Labor Office (ILO), Kappa, Observer Variation, Pneumoconiosis, Radiographic Assessment, Standard Chest Radiographs, X-ray Categories.

TABLE OF CONTENTS

INTRODUCTION

The International Labour Office's (ILO) International Classification of Radiographs of Pneumoconioses is the most widely used radiological tool in health surveillance of dust exposed workers and in research on the effects of inhaled dust on the lung. The current edition of the Classification (ILO, 1980)[1] is accompanied by 22 standard radiographs. They include 18 which illustrate various shapes and sizes of small opacities. Collectively, the appearances of these 18 films define the levels of profusion of small opacities that are classifiable into the four middle sub-categories (0/0, 1/1, 2/2, 3/3) of the "short" (4-category) scheme. Those four categories are the basis for the full, 12-category, profusion scale. It follows, therefore, that conscientious adherence to the ILO (1980) recommended procedures for classifying small opacity profusion may require comparisons of any one film under examination with up to 18 different standard radiographs.

The need to refer to so many standard films can be very inconvenient in practice, particularly in studies involving classifications of large numbers of films. It has been suggested, moreover, that the complexity of the procedure effectively acts as a deterrent to regular use of the standard films. This, in turn, may contribute towards the persistence of the well recognized problem of excessive variability in repeated classifications of the same films - both by any one reader ("within-reader variability") and by different readers ("between-reader variability"). One way of tackling this difficulty might be to reduce the total number of standard films. That might be accomplished by reproducing appropriate <u>sections</u> from several of the existing standards onto a fewer number of full size films. This report describes results from a trial that was designed to test the idea.

Twenty of the current set of standards are full sized (35cm x 35cm) *postero-anterior* images of the chest. The other two are composite reproductions of sections of full size films. (One of these defines three different levels of profusion for irregularly-shaped small shadows of size "u"; the other illustrates pleural abnormalities.) At a meeting convened by the ILO in 1989 a group of experts suggested that the total number of standards might be reduced from 22 to 14, while still retaining all the essential illustrations incorporated in the scheme. The

[1] ILO (1980). Guidelines for the use of ILO International Classification of Radiographs of Pneumoconioses. Occupational Safety and Health Series, No. 22 (Rev.) International Labour Office, Geneva.

number of composite, sectional films would be increased to six. It was agreed, however, that an attempt should first be made to determine whether, and in what way, such a change might affect readers' film classifications. The meeting therefore recommended to the ILO that an appropriate trial protocol be prepared and that the ILO invite some 10 to 12 centers of excellence on pneumoconiosis, in different countries, to participate in the film reading. Suitable sections from some of the existing (ILO, 1980) standard films were selected by the experts attending the meeting, for reproduction as new quadrant standards. The work was to be administered jointly by the ILO's Occupational Safety and Health Branch in Geneva and by the Division of Respiratory Disease Studies of the US National Institute for Occupational Safety and Health (NIOSH) at its Morgantown, West Virginia laboratory. The Task Force on Pneumoconiosis of the American College of Radiology (ACR) agreed to collaborate with NIOSH scientists in the selection and reproduction of radiographs needed for the trial.

The aim of the study was to establish whether, and to what extent, use of more composite standard radiographs would affect readers' classifications of chest films from persons who have been exposed occupationally to dust.

METHODS

Design

The design for the film reading trial was detailed in the Protocol which is reproduced in Appendix I. The essential features are as follows. Each participating center was asked to select four film readers familiar with the ILO classification. Two of them were to have classified at least 1000 films during the preceding 12 months; the other two were to have classified fewer than 500 films during the same period. Each reader was asked to classify a set of 120 selected PA chest films on at least two separate occasions: once with the aid of newly produced copies of the existing (ILO, 1980) set of standards (22 films), and on the other occasion using the experimental, modified sets consisting of just 14 films. A "cross-over" arrangement was planned for the duplicate readings, to control for possible temporal trends in reading habits. That is to say, readers at six of the centers were provided with the normal (ILO, 1980 or "ILO") set of standards on the first occasion (Round 1) that they saw the films. Readers at the other centers were sent the experimental ("QUAD") set of standards first. Readers were then sent the other set of standards for their second viewings (Round 2).

Subsequently, all readers were invited to classify the same set of 120 films once more. On this (third) occasion they were asked to classify half of them (that is, 60 films) with the help of the ILO standards, and the other 60 using the QUAD standards. Participation in the third readings was voluntary. These readings were intended to provide separate and independent estimates of within-reader variabilty for both sets of standards.

Film reading method

The procedures that were to be used for the film readings are also described in Appendix I. Films were to be classified in pre-determined (random) sequences and according to rules which specifically disallowed blanks on the reading record-sheet. The appropriate set of standards was to be used throughout. Clerical assistants were instructed to prompt readers if they attempted to classify films without comparing them with one or more films in the set of standards being used.

Films used in the trial

The ACR produced new copies of the existing (22) ILO standard films and the experimental (QUAD) standards. The latter 14 films included five new composite radiographs. Four of them display, in each case, appropriate quadrant sections from the existing (1980) standards, illustrating category 0/0 and the three main categories of small opacity profusion. The types (shape and size) of small opacities illustrated on these four new films are those described as s, u, p, and r in the ILO (1980) booklet. The new composite "u/u" radiograph re-arranges the three images on the ILO (1980) "u/u" standard and also includes a section from one of the ILO (1980) category 0/0 standards in the upper left quadrant. The fifth new composite radiograph displays appropriate sections of the existing standards illustrating large opacities (categories A, B and C). One of the other (nine) films in the modified set is the existing composite film illustrating pleural abnormalities. The other eight are full-size radiographs identical to those currently defining category 0/0 (two films), and categories 1/1 q/q, 1/1 t/t, 2/2 q/q, 2/2 t/t, 3/3 q/q, and 3/3 t/t. Appendix II provides more detailed notes on the appearances of the five experimental composite standards.

NIOSH staff at Morgantown, in consultation with the ACR Task Force, selected the 120 radiographs that were to be classified in the trial from radiographs that had been submitted by the participating centers for this purpose. These films were from persons known to have been exposed to dust, and they illustrated a variety of dust-associated appearances. Fifteen copies of each of these films were reproduced using a computer-based digitization process.

Trial arrangements

The ILO invited participation from 12 centers, in 12 different countries. All 12 accepted. Sets of standard radiographs (ILO or QUAD) as required by the protocol, copies of the 120 study films arranged in predetermined randomized sequences, and appropriately numbered data recording forms, were dispatched to all 12 participating centers during January 1992. Three different randomized sequences of the 120 trial films, designated A, B and C, were used. Any one trial center was sent sets (A, B and C), (B, C, and A), or (C, A and B) for the first, second and third viewings, respectively. All four readers at any one trial center read the films in the same random order (e.g., using randomization B) in Round 1 and, also, in Round 2 (e.g., using randomization C). Delivery of material for the first readings to three of the centers was arranged through the ILO (Geneva) office. All other dispatches of films to the centers were direct mailings from Morgantown, by express delivery. Data entry onto computer files, and preliminary tabulation of the information received, proceded throughout the trial, as results were returned.

In the light of progress made up to July 1992, and following discussion between the principal parties, the minimum time interval between centers' return of first readings and dispatch to them of material for the second readings was reduced, from the 12 weeks specified originally, to 6 weeks.

A letter inviting continuing participation also for the third reading of the test films was sent to each center when the second readings were returned. A short questionnaire accompanying those letters solicited some information regarding readers' adherence to the trial protocol. The questionnaire (Appendix III) also established more precisely the intensity of readers' usage of the ILO scheme during the 12 months preceding the trial. This was recorded as less than 500 films (low usage level), between 500 and 1000 films (medium usage level), or more than 1000 films (high usage level).

For the third readings, the 120 films were split randomly into two sets of 60. One set, designated X, was to be read by each reader using the ILO standards; the other set (Y) with the QUAD standards. Each center was provided with both sets, X and Y, and with four separate lists of randomized X- and Y- serial numbers defining the order in which the films were to be read by each reader. Two readers at each center were asked to read set X first (using the ILO standards); the other two readers were asked to read set Y first (with the QUADs). Readers were asked to allow at least 12 hours between completion of one set of readings before starting work on the other. The Instructions to center co-ordinators and readers for Round 3 are reproduced in Appendix IV.

RESULTS

Available data

When data entry closed, in May 1994, one of the 12 centers had returned results only from the first viewings of the films. Thus no comparisons between the two sets of standards were possible for the four readers concerned. Another center had completed both first and second readings ("Round 1" and "Round 2"), but not the third ("Round 3"). In principle, therefore, this material can be used to study contrasts based on readings during Rounds 1 and 2. However, preliminary inspection of the results showed that the four readers concerned had made their classifications essentially in terms of the ILO 4-point scale of small opacity profusion, rather than the 12-point scale as intended. It would be misleading to include those data with the other material available for analyses. They are therefore not considered in this report, but will be described separately.

The results reported now are thus derived from 10 of the 12 participating centers. They provide 39 sets of independent readings from Rounds 1 and 2. (It was established, by correspondence, that two readers at one center had worked together during film classifications. Only one of those correlated sets of results have been used in the analyses reported here.) At another center, only two of the four readers were able to take part in the third readings of the films. Therefore, a total of 37 independent sets of readings are available from Round 3.

The questionnaires that were returned by co-ordinating centers after the completion of Round 2 established that eight of the 19 readers who used the ILO standards in Round 1 had classified fewer than 500 films in the 12 months preceding the trial. Two had classified between 500 and 1000 films in the same period, and the other nine had classified more than 1000 films. The corresponding distribution for the 20 readers who used the QUAD standards in Round 1 was seven, five and eight.

Data editing

A preliminary examination of returns from the first and second readings revealed many blanks on the readings sheets. Most of these appeared to be due to inconsistencies between specifications for form completion in the trial protocol (Appendix I) and the format of the reading sheet that was actually used (Appendix V). The latter form, which was used in error, specifically solicits blank spaces when a reader considers a film "completely negative" or when a reader believes that there are not "any parenchymal abnormalities consistent with pneumoconiosis."

Rigorous adherence to the trial protocol would exclude from the analysis all records on forms that include blanks. The preliminary inspection of the data indicated that between a quarter and one third of all potentially available pairs of records would thus have been lost. Two edit procedures were therefore devised, and both were applied, separately, to the raw data. The first procedure disallowed all records which included blanks for profusion of small opacities, as intended originally. The second edit procedure coded blank records for small opacity profusion into a new "category naught", but only if those blanks were internally consistent with other information recorded on the data sheet concerned[2]. Existing classifications into categories 0/- and 0/0 were also pooled into the new "category naught", thus generating an 11-point scale of small opacity profusion ("naught", 0/1, 1/0, etc.). All results below refer to data emerging from the second edit which yielded something between 97 and 99% of the raw data, depending on the sub-set of results under consideration. Statistical analyses were focussed on comparisons of individual readers' repeated classifications of the same films. Thus, some further, but otherwise valid, individual records were lost when no corresponding valid record was available with which it could be paired. Supplementary analyses, not described here, have verified that analyses based on the stricter edit, which discards approximately 25% of the raw data, generates broad patterns of results that are consistent with those found using the more relaxed edit.

Between-reader variability

The 39 readers who completed both Rounds 1 and 2 each generated up to 120 valid small opacity profusion classifications on the modified 11-point scale, with each set of standards. Table 1 records the corresponding mean profusion scores, scoring 1 to 11 for categories "Naught" to 3/+, respectively. Overall, readers' mean scores ranged from 2.4 to 5.0 when using the ILO standards, and from 2.7 to 4.7 when using the QUAD standards. Figures 1a and b show the

[2] For instance, blanks in sections 2Bc of the form were classified in the second edit as "category naught" even if question 1C, i.e., "Is film completely negative?" was left blank, provided that "No" was entered in answer to 2A and no large opacities were recorded in section 2C. However, blanks in sections 2Bc, 1C and 2A were treated as missing values in analyses of small opacity profusion. Analogous edit rules were applied to determine the validity of records of the presence or absence of large opacities. Flow-charts describing these edit algorithms in detail are available on request from the authors of this report.

joint distributions of those mean results, separately for readers who used the ILO standards in Round 1 and the QUAD standards in Round 2 (IQ readers) and for those who used the standards in the reverse sequence (the QI readers). Variability between IQ readers was slightly less severe when they used the QUAD standards (in Round 2), as compared to the ILO standards (in Round 1). QI readers (Figure 1b) showed little difference between the standards in this respect.

A more refined analysis, not described here, was also performed using agreement statistics (i.e., kappa statistics). This analysis detected no difference between the standards with respect to between-reader variability.

<u>Within-reader variability</u>

The reproducibility of readers' repeated classifications of the same films when using the same standards can be described only for the 37 readers who completed all three rounds of the trial. Sixty of the films seen by any one such reader during Round 3 were classified previously by that reader (in Round 1 or 2) using one of the two sets of standards. The remainder of the films seen by that reader in Round 3 were classified previously using the other set of standards.

Appendix VI (Table AVI.1) records the percentages of individual readers' classifications that were identical when they used the same set of standards twice. For the 19 IQ readers, these figures ranged from 22.0% to 63.3%, averaging 41.6% and 45.4% for the ILO and QUAD standards, respectively. The standard error of the mean of the 19 differences (% agreed with ILO standards - % agreed with QUADs = -3.8%) was 1.92, indicating that the slightly better reproducibility when using the QUADs could be due to chance ($P \approx 0.065$). QI readers, whose identical classifications ranged from 25.4% to 69.5%, showed a reverse trend, with 48.3% and 43.4% mean percentage identical classifications when using the ILO and QUAD standards, respectively (SE of mean difference = 2.17, $P< 0.04$). Thus, the difference in reproducibility when using the two sets of standards, averaged over all 37 readers, was trivial (0.4%).

Kappa coefficients (κ), reflecting the degree of agreement in each reader's classifications of identical films with the same standards, were calculated separately for those films that were classified twice when using the ILO standards, and for those that were classified twice with the QUADs. The higher (more positive) the numerical value of κ, the lower (less severe) the within-reader variability[3]. The paired values of κ, one pair for each reader, are shown graphically in the scatterplots of Figures 2a and b, separately for IQ and QI readers. If there was no difference between the standards with respect to within-reader variability, we would expect the plotted points to be close to the reference line, with approximately the same number of point above and below the line. Figure 2a, with 7 points close to, but just above, the diagonal 1:1 line, indicates better agreement for those seven readers when using the ILO standards, in Rounds 1 and 3. The 12, more scattered, points below the line in Figure 2a, indicate better reproducibility when using the QUADs, in Rounds 2 and 3. Figure 2b shows a different pattern among the 18 QI readers who completed Round 3: the majority had better reproducibility when using the ILO standards (in Rounds 2 and 3); only four of them show the opposite tendency. The aggregate picture for all 37 readers shows no systematic difference between the standards in these measures of reproducibility, but it does imply that readings involving Rounds 2 and 3 were more similar, in general, than those made in Rounds 1 and 3.

The clustering of symbols used in Figures 2a and b suggest that readers' varying usages of the ILO scheme during the 12 months preceding the trial may have influenced these results. Results from Analyses of Variance on the differences: (κ_{ILO} - κ_{QUAD}) from each reader are summarized in Appendix VI. These show that there were statistically significant differences between the paired values of κ among the 14 readers who had used the ILO scheme relatively infrequently in the 12 months preceding the trial. Eight of them (IQ readers), who used the QUAD standards for the first time in Round 2, and then again in Round 3, generated higher values of κ from films seen on those two occasions, as compared with κ values referring to use of the ILO standards in Rounds 1 and 3 ($P \approx 0.12$). The other six readers who had classified no more than 500 films in the year preceding the trial and who had been allocated to Sequence QI, returned higher values of κ when using the ILO standards in Rounds 2 and 3 than when they used the QUAD standards in Rounds 1 and 3 ($P< 0.002$). These results may be interpreted *either* as conflicting within-reader-variability trends with respect to the two sets of standards, *or* as a consistent tendency for some readers, with relatively

[3] Note that the κ statistic of reproducibility reflects not just the total number of the identical classifications, but also where, on the 11-point ordinal scale, those agreements occur, given the realized distributions of classifications on the scale by the reader concerned.

infrequent usage of the ILO Classification immediately before the trial, to demonstrate better reproducibility for classifications made in Rounds 2 and 3 than for those made in Rounds 1 and 3. (The statistical analysis cannot distinguish between these alternative interpretations.)

Additional analyses (Appendix VI) showed that the *magnitudes* of differences, on the 11-point scale, between repeated classifications of the same films, were significantly less pronounced when using the QUAD standards.

These analyses thus indicate that on those occasions when paired classifications of the same films (and using the same standards) were not identical, there was a tendency to register less marked deviations when using the QUAD standards. However, the data show no consistent trend for reproducibility, as measured by the κ statistics, to worsen or improve when using the QUAD standards.

Profusion of small opacities

The 39 readers who completed Rounds 1 and 2 provided 4591 valid pairs of classifications concerning the presence or absence ("category naught") of small opacities. The percentage distribution on the *4-point* profusion scale when using the ILO standards was 39.6% in category 0, and 37.9%, 17.2% and 5.3% in categories 1, 2 and 3, respectively. The corresponding distribution when the QUAD standards were used were 37.4% into category 0, and 41.0%, 17.2% and 4.4% into categories 1, 2, and 3; i.e., 2.2% higher "prevalence" of category 1+ when using the QUAD standards.

Tables 2a and b elaborate these overall results in terms of the 11-point profusion scale generated by the edit procedure, separately for the 19 readers who used the ILO standards in Round 1 (Sequence IQ, Table 2a), and for the 20 readers who used the QUAD standards in Round 1 (Sequence QI, Table 2b). IQ readers returned almost identical numbers of classifications into category 0 (i.e., "naught" and 0/1 combined) when using the ILO and QUAD: 40.0% and 39.6% for the respective sets of standards. QI readers recorded 39.3% and 35.3% into category 0 respectively, i.e., with 4% higher "prevalence" of category 1+ when using the QUAD standards.

The summary statistics shown immediately below Table 2a record that 44% of all paired classifications by IQ readers were identical on the 11-point scale when using the two sets of standards. The entries above the diagonal (26% of the total) refer to classifications where more profusion was recorded when using the

QUADs. Those below the diagonal (30%) are paired classifications where profusion was higher when using the ILO standards. The *net* effect for IQ readers was therefore to record slightly more profusion on the 11-point scale when using the ILO standards.

In contrast, the corresponding statistics summarising Table 2b show 31.7% of the classifications with higher profusion classifications when using the QUADs and only 23% (below the diagonal) with higher profusion classifications when using the ILO standards. In other words, the patterns of results comparing small opacity profusion classifications when using the two sets of standards differ between the two groups of readers involved. On average, those who were allocated the ILO standards at the first viewings and QUAD standards at the second (Sequence IQ) recorded slightly more profusion on the 11-point scale when using the ILO standards - in Round 1. Those who were allocated the QUAD standards in Round 1 classified more profusion when using the QUADs - also in Round 1. It is not immediately obvious, therefore, whether the results in Tables 2a and b, shown pooled in Table 2c, are to be interpreted as indicating a modest tendency to classify more films into category 0/1 or higher when using the QUAD standards, or whether these results should be interpreted as indicating a tendency to record more profusion on the *first occasion* that the set of 120 test films were seen.

Summary statistics from analogous (11 x 11) Tables that record individual readers' results are shown in Table 3. The percentages of paired classifications where readers repeatedly chose the identical profusion category when using the two sets of standards ranged from 21.7% to 61.9%. Higher profusion classifications when using the QUADs ranged from 10.2% to 61.8% over the 39 readers. Higher profusion categories when using the ILO standards ranged from 6.5% to 67.8%. (The corresponding average statistics, from all 39 readers' pooled results, are shown in Table 2c.)

Inspection of the individual cells in Table 2b shows that most of the classifications by QI readers that indicated more profusion with use of the QUADs, involved films judged to be in the lower part of the profusion scale, up to about the end of category 1. Results from IQ readers are more variable in this respect. Some entries in Table 2a show higher QUAD-associated profusion assessments in the lower part of the scale, but these are matched by others in the same region of the scale that show more profusion when using the ILO standards; and the latter extend also into categories 2 and 3. Figures 3a and b illustrate these different patterns. The distances above and below zero on the vertical axis of these graphs reflect both the numbers and magnitudes (on the 11-point scale) of differences in profusion assessments associated with use of the two sets of standards. Positive values are to be interpreted as more profusion when using the QUADs; negative values as more profusion when using the ILO standards. These *Profusion Difference Indicators* (PDI) are shown in Figures 3a and b in relation to where, approximately, they occurred on the 11-point profusion scale.

Figure 3c is based on the pooled results in Table 2c. It indicates more QUAD-associated profusion up to about category 1/2, and a reverse, but less marked tendency higher on the scale. This is the resultant of the conflicting trends in classifications produced by the two groups of readers. The pooling of those data in Figure 3c obscures the difference between the patterns in Figures 3a and b.

Figure 3d is based on a re-arrangement of the same data, as shown in Table 2d, that is, in relation to whether the classifications were made in Round 1 or Round 2. The graph indicates that most of the higher profusion classifications that occurred in Round 1, as compared with Round 2, were in the lower part of the scale. But this impression, too, is potentially misleading, since Figures 3a and b demonstrate that the overall appearance of Figure 3d is a composite of two distinctly different patterns that depend on which set of standards, ILO or QUAD, was used in Round 1.

All these observations draw attention to an intriguing, but also complicating feature of the results: patterns that emerge by considering the totality of classifications from all readers may hide conflicting tendencies that depend on the sequence in which the standards were used. The following descriptions of the data, and their statistical analyses, are therefore arranged with this potential difficulty in mind. Results from the IQ and QI sequences are considered separately before attempting to generalize from the findings.

Shapes of small opacities

The predominant shapes of the small opacities that were recorded in Rounds 1 and 2 are shown in Table 4, separately for IQ and QI readers. At least three aspects of these results should be noted.

First, IQ readers, in aggregate, recorded predominantly rounded shapes more frequently than predominantly irregular shapes (841 + 900 = 1741 rounded, and 752 + 754 = 1506 irregular; Table 4a). But QI readers (Table 4b) showed the reverse tendency (679 + 702 = 1381 rounded; 1090 + 970 = 2060 irregular).

Second, the number of times that IQ readers recorded predominantly irregular opacities when using the the ILO standards was, in total, almost identical to that found with the QUAD standards (754 and 752, respectively). Rounded shapes, however, were recorded more frequently with the ILO standards (900, as compared with 841 with the QUADs). In contrast, QI readers (Table 4b) again exhibited precisely the contrary pattern: very similar (slightly fewer) notations of rounded shadows when using the QUADs (679, *versus* 702 with the ILO standards) but more records of irregular shadows when using the QUADs (1090, *versus* 970 with the ILO standards). This suggests that the differing results from the two groups of readers, regarding small opacity profusion when using the two sets of standards (Tables 2a and b), may have been due in part to QI readers recording more small irregular opacities when they used the QUADs (in Round 1). The more modest excess of small opacity classifications by IQ readers when using the ILO standards (also in Round 1, Table 2a) refers, presumably, to shadows judged as predominantly rounded, rather than irregular, in shape.

Third, although Table 4a shows that the totals of irregular shape records from IQ readers were almost identical for the two sets of standards (752 and 754), these classifications referred to the same films on only 70% of the occasions when irregular shadows were recorded. The other 30% of irregular shape records in Table 4a refer to classifications of films where (a) predominantly irregular shapes were recorded when using one set of standards but there was no record of any shape when using the other set of standards (93 + 127); and (b) records of predominantly *irregular* shapes of small opacities were entered when using one set of standards and predominantly *rounded* shapes when using the others (142 + 110.)

The latter pair of results indicates that, when IQ readers recorded different predominant shapes for the same films in Rounds 1 and 2, there was a tendency to record irregular shapes more often when using the QUADs. QI readers showed a more marked trend in the same direction (Table 4b.) In this case, therefore, it is reasonable to refer to the pooled results from all 39 readers in Table 4c. This shows that, if attention is restricted to those pairs of film classifications where readers recognized some small opacities in both Rounds 1 and 2, but changed their minds about the shapes of the small opacities that they saw, then use of the QUAD standards resulted in irregular shapes being nominated more frequently (293 times) than use of the ILO standards (217 times; P<0.001).

Table 5 provides more information about the predominant shapes of the small opacities that were recognized by the 37 readers who saw the 120 films a third time, in Round 3. In this case, two sub-sets of results are shown for both groups of readers. In each case, the first sub-set refers to readers' classifications of the 60 films that they viewed with the same standards in Rounds 1 and 3, while the second sub-set refers to films viewed twice with the other standards, in Rounds 2 and 3. (In what follows, the corresponding *film-reading sub-sequences* are referred to as IQI and IQQ for IQ readers, and as QIQ and QII for QI readers.)

Three quarters of the 12,929 classifications included in Tables 5a-d refer to predominant shapes of small opacities. There was little difference in this respect between use of the two sets of standards: 74.4% of viewings with the ILO standards yielding records of shape, and 74.7% of those with the QUADs. (Note, however, that the classifications contributing to the numerators of these percentages do not necessarily refer to the same films.)

Table 5 reinforces the impression gained from Table 4: the distributions of readers' judgements about the predominant shapes of small opacities varied, depending both on which set of standards was used *and* on the sequence of their use. In particular, it is clear that QI readers characterized the predominant shapes of small shadows more frequently as irregular, rather than rounded, irrespective of which standards they were using, and on all three occasions that they saw them. However, a reverse tendency among IQ readers suggested by Table 4, is evident only in Table 5a, for sub-sequence IQI. Table 5b, which describes the IQ results for the sub-set of 60 films that were classified with the QUAD standards in Rounds 2 and 3, shows more frequent recording of irregular rather than rounded shapes.

If all available results from each of the two groups of readers in Table 5 are considered, irrespective of film reading sub-sequence, then it may be seen that there is a general tendency to nominate irregular shapes more frequently when using the QUADs, regardless of which set of standards was used in Round 1. That tendency is more pronounced among IQ readers than among QI readers. This is demonstrated in Table 6, which re-arranges relevant portions of entries in Table 5. QI readers nominated irregular opacities on 59% of the occasions when they saw small shadows while using the ILO standards, and in 62% of their small opacity classifications when using the QUADs. IQ readers recorded 46% as irregular in shape with the ILO standards, and 52% with the QUADs.

These findings prompted further study of results on the profusion of small opacities in relation to the accompanying designations of their predominant shapes.

Profusion of small opacities with different shapes

Appendix VII records more detailed analyses of the small opacity profusion data from Tables 2a and b, with respect to the predominant shapes of the small opacities that were nominated. Table 7 summarizes those analyses. It shows that when readers recorded that small irregular shapes were predominant on a film in both Rounds 1 and 2, they tended to attribute higher profusion levels to them while using the QUAD standards. This is particularly noticeable for QI readers, who recorded higher profusion levels when using the QUADs for 40% of their paired classifications, but only 23% with higher profusion classifications when using the ILO standards (Section 4 in Table 7). The corresponding result for IQ readers is in the same direction but much more muted (33% and 30% above and below the diagonals, respectively, in Section 1).

Of the (517 + 795) = 1312 paired classifications where readers repeatedly classified the shapes of small shadows as predominantly irregular, 906 (69%) also attracted identical classifications regarding their size: 's' was nominated (twice) 536 times, 't' 338 times. and 'u' 32 times. Results from the 's' and 't' sub-sets of classifications are also detailed in Appendix VII. They show that the higher profusion classifications asociated with use of the QUAD standards was evident both when 's'- and when 't'-sized irregulary shaped shadows were recognized twice; and again, this phenomenon was distinctly more marked in the QI than in the IQ readings (Tables AVII.11 to AVII.14).

When IQ readers repeatedly recognized small *rounded* opacities on the films (Section 2 of Table 7) 36% of such viewings resulted in higher profusion classifications using the ILO standards, as compared with 26% when using the QUADs. A similar pattern is evident in IQ readers' classifications of other films, where the primary shape recorded in Round 1 differed from that recorded in Round 2 (Section 3: 27% and 22% below and above the diagonal). The corresponding sub-set of 1042 paired profusion classifications by QI readers (Section 6) shows the opposite trend: more classifications above the diagonal than below.

Table 7 also records the percentage distributions on the 4-point profusion scale that were generated by use of the two sets of standards. Note that IQ readers' more frequent classifications of small rounded opacities with the ILO standards (Section 2 of the Table), resulted in 5% more classifications with the ILO standards into categories 2 and 3 combined; but category 1 attracted 6% more classifications when using the QUADs. Figure AVII.3 in Appendix VII illustrates these different patterns at the two extremities of the radiological scale.

The statistical significance of the various, partly conflicting, trends evident in Tables 2 to 7 was studied by considering differences in profusion scores[4], from individual readers' repeated classifications of the same films with the two sets of standards. Those differences, calculated from data captured in all three rounds, were then grouped according to

(a) the *film-reading sub-sequence* that generated a pair of classifications (IQI and IQQ for IQ readers, QIQ and QII for QI readers) and

(b) the rounds that were involved i.e., (Rounds 1 and 2) or (1 and 3) or (2 and 3).

[4] Again, scoring 1, 2, ...11 for categories "naught", 0/1,...3/+.

The sums of individual readers' score differences, within the four distinct combinations of sub-sequence and pairs of rounds that are relevant to any one reader's comparisons of the standards, were expressed as percentages of the numbers of paired classifications involved. This yielded 4 x 37 = 148 reader-specific *Profusion Difference Scores* (PDS). The PDS were defined, arbitarily, to be positive when the profusion score was higher using the QUAD standards. Negative values of this statistic are therefore to be interpreted as more profusion recorded when using the ILO standards.

Figures 4a and b illustrate both the variability of PDS between readers, and also a relationship between the PDS and readers' judgements about the predominant shapes of small shadows that they identified. The variable S, plotted on the abscissae of Figures 4a and b is

> the number of predominantly irregular small opacity classifications made by a reader when using the QUAD standards, expressed as a fraction of all classifications by that reader involving some small opacities, including those made while using the ILO standards.

The higher the value of S, the greater the tendency for the reader concerned to nominate irregular shapes when using the QUAD standards on films where small opacities were detected. Values of S ranged from 7% to 44%. For QI readers, all but two of the 72 points in Figure 4b show values of S greater than 23%, and the mean value of S for this group was 31%. The mean value of S for IQ readers was 24%. The difference is very unlikely to be due just to chance ($P< 10^{-7}$). Evidently, readers who used the QUAD standards in Round 1 were indeed more likely to categorize the small opacities that they saw as irregular than were the other readers - as was indicated by the results in Table 5.

Figure 4 also indicates the extent of readers' usage of the ILO Classification in the 12 months preceding the trial. No obviously generalisable pattern emerges, but it is noticeable that seven of the eight highest values of PDS in Figure 4b are from readers with relatively low usage.

Table 8 shows the mean values of readers' PDS, within combinations of sub-sequences and rounds involved. The overall mean, 3.1, implies that the average results, from all readings by 37 readers who completed Round 3, amounted to about 3 higher profusion sub-categories per 100 film classifications when using the QUADs. The contrasting mean values for the two groups of readers reflect the pattern apparent from Table 2. They show that, based on results from all three rounds, IQ readers read two more sub-categories per 100 film classifications when using the ILO standards, on average; but QI readers recorded eight more sub-categories per 100 film classifications when they used the QUADs.

Conspicuous in the body of Table 8 is the average of IQ readers' results when they used the QUAD standards in Round 3 and the ILO standards in Round 1. Their profusion assessments with the QUADs, for that particular sub-set of 60 films, averaged about 16 more sub-categories per 100 film classifications than with the ILO standards, in Round 1. This observation is remarkable because, when those readers classified the same 60 films with the QUAD standards in Round 2, they recorded less profusion than they had noted earlier with use of the ILO standards in Round 1 (mean PDS = -9). This suggests that, when IQ readers used the QUADs for the second time in Round 3, they tended to see more small opacities than when they used them earlier in Round 2.[5]

The variability of the PDS, and its dependence on the predominant shapes of the small opacities that were seen (S), was studied in a series of multiple regression analyses. These were pursued separately for the results from Sequences IQ and QI, using a variety of algebraic formulations to reflect the possible importance of various factors that exploratory graphical and tabular analyses suggested might be relevant. The factors considered were:

(a) READER-specific tendencies to classify more or less profusion when using one or other set of standards;

(b) the predominant SHAPEs of the small opacities that readers attributed to small opacities;

(c) the film-reading SUB-SEQUENCES within main Sequences;

(d) the two ROUNDS INVOLVED in any one sub-sequence;

[5] Those two sets of classifications with the QUAD standards are compared in Appendix VII (Table AVII.8 and the associated Symmetry Plot - Figure AVII.8). Further analysis of this particular sub-set of the data indicates that the asymmetry evident in Figure AVII.8 is unlikely to be due simply to chance (P<0.03; see page 3 in Appendix VIII.) There was no analogous tendency to read more profusion in Round 3 than in Round 2 when QI readers used the ILO standards repeatedly (see Tables 8 and AVII.10).

(e) the extent of PRIOR USAGE of the ILO scheme before the trial;

(f) the participating CENTERS with which readers were affiliated; and

(g) the readers' REPRODUCIBILITY (within sub-sequences) of classifications of the same films when using the same standards.

Possible interactions between these factors were also considered, and the effectiveness of different models was assessed by looking for patterns and trends in residual, unexplained variability.

Table 9 summarizes results from analyses that represent the PDS results, within main sequences, as dependent on the SHAPE factor (S) after adjustment for variations attributable to READERS (R), REPRODUCIBILITY (κ), and possible interactions between (R) and (κ). The latter interaction factor reflects differences between readers in any tendency for the PDS to be associated with reproducibility. These analyses explained 63% and 43% of the variability in data from IQ and QI sequences, respectively, and this is more than was explicable by any of the other formulations considered.[6]

Table 9 confirms that the patterns suggested by Figures 4a and b, of associations between the PDS and the SHAPE factor (S), are unlikely to be due simply to chance (P< 0.0001 and P< 0.03 for IQ and QI readers, respectively. The significantly positive regression coefficients are to be interpreted as follows:

> *the higher the proportion of films classified by a reader as showing predominantly irregular (rather than rounded) small opacities when using the QUADs, the more likely that reader was to record higher profusion scores for those films when using the QUAD (rather than the ILO) standards.*

The magnitudes of these effects, as measured by the regression coefficients, did not differ significantly between IQ and QI readers. The (weighted) mean gradient in PDS with respect to S (averaged over both Sequences) amounts to 58 units increase in the

[6] For instance, models providing for possible effects of SHAPE of small opacities, USAGE of the ILO scheme before the trial, CENTERS, and all first and second order interactions between these factors, accounted for 30% of the total variability in the IQ readers' results, but for only 3% in those from QI readers.

PDS for any 10% increase in S. Thus the 7% higher mean level of S generated by QI readers (see Figures 4a and b) is more than sufficient to explain the higher mean level of PDS recorded by the QI readers (Table 8).

Additional analysis (not presented here) confirms that some IQ readers' higher Profusion Difference Scores were associated with better reproducibility of small opacity profusion classifications when using the same standards, but that this tendency was not uniform among the IQ readers. Among QI readers, on the other hand, within-reader reproducibility when using the same set of standards (as reflected in the κ statistics) contributed little to explaining the variability between readers in their differing PDS results.

Table 10 shows the extent to which the statistical model represented by Table 9 explains the variability in average results depicted in Table 8. Variations in IQ readers' consistency when using the same standards, that is, variations in the κ statistics, accounts for most of the difference between the two IQQ mean PDS results in Table 8 [15.9 - (-9.3) = 25.2] The mean residuals concerned, in Table 10, are ±1.9. The corresponding pair of mean residuals from an analysis that did not include the REPRODUCIBILITY factor were considerably higher (±7.0).

In summary, these analyses have shown that the comparisons of small opacity profusion when using the two sets of standards were influenced primarily by

(a) which reader viewed the films;

(b) the sequence in which the sets of standards were used; and

(c) readers' opinions about the predominant shapes of small opacities that they recognized.

In practice, these three factors operated in a way that generated a relatively small overall average result with regard to the profusion of small opacities: about three more profusion sub-categories per 100 films classified when using the QUAD standards. Nevertheless, Table 11 shows that each of four distinct sub-sets of results, corresponding to four different sequences in which the standards were used, resulted in more classifications into categories 1, 2, and 3 combined when using the QUAD standards. In total, the QUAD standards were associated with 4.1% more classifications into category 1, and 2.3% fewer classifications into categories 2 and 3 combined. The net effect was that the prevalence of categories 1, 2 and 3 combined in the particular set of films used for this trial was about 2% higher when using the QUAD standards.

Large opacities

The 39 readers who completed Rounds 1 and 2 provided a total of 4529 pairs of valid records regarding the presence or absence of large opacities (Table 12). In aggregate, large opacities were recorded at 5.1% of viewings with the ILO standards and at 4.5% of those using the QUADs. Individual readers' judgements regarding the presence of large opacities ranged from 2.5% to 12.5% of the pairs of viewings in Rounds 1 and 2. Seventeen of the 39 readers noted large shadows at up to 4% of their viewings; another 17 at between 4 and 6% of their viewings. Two of the other five readers recorded 6.2%. The three highest reader-specific proportions recorded in Rounds 1 and 2 combined were 10.1%, 11.8% and 12.5%.

Table 12 shows that 4270 of the paired readings in Rounds 1 and 2 (94.3%) recorded no large opacities on both occasions. Among the remainder, 29 indicated the presence of large opacities when using the QUAD standards but not when using the ILO standards, and 55 showed the reverse, that is, some large opacities when using the ILO standards but none when using the QUADs. The disproportion (55:29), was apparent in both reading sequences (Table 13), was more marked among IQ readers, and is very unlikely to be due just to chance (P<0.0005). All 84 paired classifications concerned were therefore studied individually with respect to the co-ordinating centers and readers involved. Twenty of these 84 paired classifications were by just two readers, from the same (IQ Sequence) center, and both reported relatively low usage of the ILO scheme (less than 500 films classified) in the 12 months preceding the trial. Eighteen of the latter (20) paired classifications recorded large opacities when using the ILO standards, but not with the QUADs. Of the other 64 pairs of anomalous classifications, 37 nominated large shadows with the help of the ILO standards, and 27 when using the QUAD standards.

Supplementary analyses showed that on 22 of the 55 occasions when large shadows were nominated with the help of the ILO standards but not with the QUADs, the symbol indicating the presence of coalescence of small shadows ("ax") was recorded with the QUAD classifications; cancer ("ca") was noted on five further occasions. Similarly, 13 of the 29 paired classifications in which large opacities were recorded when using the QUADs, but not the ILO standards, included records of "ax" when using the ILO standards.

DISCUSSION

The ILO (1980) Classification of Radiographs of Pneumoconioses is used in many countries to assess the prevalence of radiographic abnormalities in the lungs of workers occupationally exposed to dusts, and to monitor radiographic changes in workers' lungs over time. Any major change in the procedures or conventions associated with the Classification could artefactually alter the patterns emerging from such investigations. The ILO therefore decided that possible effects of a proposed change to the set of standard radiographs that accompanies the Classification should be studied in a suitably controlled trial. The change proposed was to remove 12 of the existing full-sized standard radiographs, and to replace them with four new films which reproduced only selected sections of the standards that had been removed.

It can be argued that no very dramatic effect on levels of small opacity profusion was to be expected as a result of the proposed change, because the six standard films that currently define profusion for the most commonly reported sizes of small rounded and small irregular shadows (q and t) were all retained unaltered in the experimental ("QUAD") set, as were the existing two full-size radiographs that illustrate the range of appearances classifiable as category 0. Two such films, rather than just one, were first introduced with the 1980 revision of the scheme, in an effort to emphasize that classifications into category 0 were not to be restricted to totally unexceptional appearances that show no small shadows whatsover; the ILO (1980) Guidelines to the Classification defines category 0 as applicable to appearances where "small opacities [are] absent or <u>less profuse than the lower limit of category 1</u>" (emphasis added). The experimental set of standard films used in this trial reinforces that point, not only because it retains both full-size existing category 0 films, but also because it reproduces selected sections from both of them onto quadrants of four of the new composite films. But perhaps the latter change could, by itself, alter readers' perceptions of what profusion category should be allocated to a film? If so, then that fact can be established only empirically. The trial was designed to answer questions of this type.

The other change incorporated in the experimental QUAD standards is the compression of three current standard radiographs that illustrate large shadows classifiable as categories A, B and C, onto a single composite film. The latter reproduces only appropriate sections from the existing films, where the large shadows are visible.

This change too, would not, *a priori*, have been expected to result in any change in film reading habits, since the scheme defines large shadows strictly according to their dimensions; the standard radiographs in this case are described in the ILO (1980) Guidelines as examples, not definitions, of large opacities.

Nevertheless, the analyses described above have identified a distinct and statistically significant trend for some readers to classify films more frequently as showing irregularly shaped small shadows when using the QUAD standards. That trend was accompanied by a tendency for those readers to categorize such films as having levels of profusion that are higher than those recorded when using the existing (1980) set of standards.

The tendency to attribute higher levels of profusion with the QUAD standards to films thought to show predominantly irregular opacities was evident both when those films were classified as primarily of size 's' and when they were classified as 't'. Yet the standard films defining the profusion levels for t-sized small shadows were precisely the same in the QUAD set as in the ILO (1980) set of standards. It seems, therefore, that the apparent difference between results when using the two sets of standards is not explicable simply in terms of the appearances of the standard films involved. Other factors, perhaps associated with the circumstances in which the trial was conducted, seem also to have played a role.

Also conspicuous in the data are fairly persistent differences in patterns of results from readers grouped according to which of the two sets of standards they used on the first occasion that they saw the trial films. The distributions of prior usage of the ILO Classification in the 12 months preceding the trial were similar for the "IQ" and the "QI" readers, but it is possible nevertheless that the reading habits in the two groups happened to differ systematically. Alternatively (or additionally) it could be that the sequence in which the two sets of standards were presented to readers was an important contributory factor in determining their classifications. One may speculate about possible reasons for such a sequence effect. Perhaps QI readers, faced in Round 1 for the first time with an unfamiliar set of standards and with a batch of films that they had never seen before, were particularly careful to classify all appearances consistent with the standard films, including some that they might otherwise have disregarded. If so, then it would be understandable that, presented in Round 2 with a familiar set of standards, and 120 films that they had seen a few weeks earlier, and perhaps recalling in broad terms the high proportion of films that were likely to show abnormalities, they would be more relaxed, less anxious, and less inclined to classify doubtful appearances that were recorded assiduously on the first occasion. To a limited extent, results from the IQ readers are consistent with this idea, since they too recorded slightly more profusion in Round 1 (with the ILO standards).

A further fairly clear-cut difference between results generated by the two groups of readers was that the tendency for the IQ-Sequence readers to classify more profusion when using the ILO standards, in Round 1, was expressed primarily as more classifications higher on the profusion scale, in categories 2 and 3, and the shapes of the small opacities involved were predominantly rounded. In contrast, the QI readers' more frequent classifications of small shadows when using the QUAD standards occurred mainly lower on the scale, in categories 0/1 and 1, and they involved opacities that were characterized as predominantly irregular in shape. Even if these different patterns are interpreted as driven primarily by which set of standards were used at the first viewing of the films, the fact remains that IQ readers also identified more irregular opacities in the lower part of the profusion scale when using the QUADs. The consistency of this observation over both groups of readers strongly suggests that it reflects a real effect associated with the QUAD standards.

Whatever the reason for the apparent difference between results from the two groups of readers, the design of the trial enabled the phenomenon to be identified, and allowances were made for it in the analysis of results. The availability of the third set of readings was particularly helpful in this respect because the way that the films were distributed for Round 3 provided (partial) replication of individual readers' classifications of films when using the same set of standards. Analysis of all the available data, including those from Round 3, verified that the main trends suggested by results from the first two rounds on their own were sustained, irrespective of which set of standards was used in Round 1. In particular, it was possible to quantify

- a tendency to characterize small shadows as irregular in shape more frequently when using the QUADs (Table 5),

- an approximately linear relationship between this trend and a tendency to record higher profusion levels when using the QUADs (Figure 4 and Table 9), and

- a higher proportion of films attracting classification into category 1 when using the QUADs (Table 11).

Again, the latter effect was more apparent in results from QI readers than in those returned by IQ readers. On average, however, it amounted to 1.7% more frequent classifications into category 1 or higher while using the QUADs with the particular set of films chosen for this trial. This average result may be expressed as about three more profusion subcategories per 100 film classifications with the QUADs.

All average results should be interpreted with caution, particularly in the context of a trial such as this. In the first place, the artificiality of the trial setting may itself be expected to affect results. The "sequence effects", discussed above, are an example. But in general, the directions and magnitudes of such artifacts are not necessarily predictable or easily quantified.

Second, selection of the 120 films for study was deliberately contrived to include a high proportion (about 60%) with signs likely to be classified as abnormalities. In consequence, the wide range of appearances that are associated with the pneumoconioses was simulated successfully in a relatively small and manageable batch of films. But the very high prevalence of small shadows represented in this contrived film set is atypical of most real situations where the ILO Classification is used for surveillance or research purposes. It can be argued therefore, that the relatively small effects associated with the use of the QUAD standards that were found in this trial are unlikely to be detectable in most real-life survey situations.

Third, the additional 3 sub-categories per 100 film classifications that were associated with use of the QUAD standards were found in the context of a sophisticated, and statistically powerful, trial design. The magnitude of that estimated effect should be considered in relation to the considerable variability between the readers who generated the average result. Reader-specific average profusion scores for the 120 trial films ranged over the equivalent of 2.2 sub-categories per 100 film classifications when using the ILO standards and 1.7 sub-categories per 100 film classifications when using the QUADs. The within-reader reproducibility of classifications on the 11-point scale when using the same standards also varied considerably, ranging from 22% to 70%. And while 21 of the 39 readers read more profusion with the QUADs, the other 18 read more profusion with the existing ILO standards.

Fourth, there is strong evidence that the relatively modest average tendency to classify films with the QUAD standards into higher levels of profusion than with the current set, is specific to films where readers feel that small irregular opacities are predominant. Therefore, in situations where small irregular, rather than rounded, shapes are likely to occur (for instance, in surveys of asbestos workers) the apparent difference between the two sets of standards is likely to be more important than in situations where the majority of radiological abnormalities are likely to be small rounded opacities (for instance, in surveys of coal miners).

Fifth, subsumed in the overall average results, there was a trend among IQ readers to record more profusion of small opacities with the QUAD standards when they used them for the second time, in Round 3. The estimated increase in profusion levels was relatively large, amounting to 4.6% more frequent classifications into catgeory 1 or higher on the second occasion that the IQ readers used the QUADs. If this was due to increased familiarity with the QUAD standards, then it might be interpreted as indicating that the average results from this trial understate the kind of effect that some readers' use of the QUADs is likely to have in the future.

Neither between- nor within-reader variability was affected importantly by which set of standards was used in the trial. But there was an unexpected tendency for readers to omit notations of large opacities when using the QUADs with (the few) films where they did note such shadows when using the ILO standards. Many, but not all, of of these discrepancies were in results from just two readers, both of whom had used the ILO scheme relatively infrequently before the trial. It may be that in some of the cases where readers recorded large opacities when using one set of standards, they described the same shadows as coalescence of small opacities, or perhaps as cancer, when viewing them with the other set. There is some evidence suggesting that this may have occurred, but it cannot explain all the discordant classifications. In any case, there is no obvious reason why records of large opacities should have been made significantly more frequently when using the ILO, as distinct from the QUAD, standards. As noted above, large opacities are defined in terms of their dimensions, rather than according to the similarity of appearances with those on standard radiographs or to their suspected aetiology.

It is a fact, however, that the introductory General Instructions with the ILO (1980) Classification includes an injunction that readers should not classify any appearances which are "definitely not pneumoconiosis". (The booklet does not explain how such a definite diagnostic decision can be made simply from the appearance of a chest radiograph.) We were unable to determine, from the available data, whether readers made such decisions more frequently when viewing films with the QUAD, as opposed to the ILO, standards. This fact strengthens the case that has been argued by some to the effect that all shadows consistent with those appearing on the standard radiographs, or the definitions in terms of dimensions, should be recorded, irrespective of whether the appearances are thought to be due to pneumoconiosis. The General Instructions with the Classification would then have to be amended accordingly, and the already obligatory use of Symbols and Comments would have to be re-emphasized to ensure that, where appropriate, readers' opinions about shadows that they think are not attributable to dust would be on record and available for clinical use, as well as statistical analysis (see paragraph 9 in "Recommendations for Future Research" on page 33 of the ILO (1980) Guidelines booklet).

CONCLUSIONS

It seems to us that the complexity of results from this study vindicates the decision to conduct it, the care that went into its design, and the scale of effort that was required for its completion. The likely effect of replacing the existing 22 standards by the smaller set of 14 films has been quantified.

Of course, that information is specific to the particular set of 14 experimental standards that were used in the trial. There is no reason to suppose that similar results would have been found if alternative sections of the existing standards had been chosen to construct the new quadrant films, or if further full-size standards had been replaced by additional sectional composites. On the contrary, the results have demonstrated that intuitive expectations of null effects may be negated in practice. And if the most frequently used full-sized standards (for types q/q and t/t) had been replaced by quadrants, then it would have been reasonable to expect that at least some readers' classifications would have been affected - but in an unpredictable way. Perhaps the most important outcome from this investigation, therefore, is that it has demonstrated the wisdom of the ILO's insistence that major changes to its Classification of Radiographs of the Pneumoconioses must first be justified by appropriately conducted trials before they are adopted.

It remains then for the ILO to decide whether the evidence presented here does justify introduction of the particular modified standards that were tested in this trial. No doubt advice will be sought from various experts, particularly those who read the films. Our view is that nothing that we have reported here contra-indicates the general principle of introducing more sectional ("quadrant") standards to replace or supplement some of the present full-sized standards. The greater ease with which such films may be manipulated should encourage their more frequent use and may, in turn, help to reduce observer variability in film classifications.

We feel that the relatively small average effects identified in the trial setting are unlikely to be distinguishable, in practice, from the between- and within-reader variability commonly found in epidemiological studies (and demonstrated also in the results reported here). But we note that early signs of effects on the lung that are manifested as small irregular opacities, may be detected more readily with the modified standards than they would be with the existing set. From the point of view of occupational health screening and surveillance, and the associated objective of protecting workers from the effects of exposure to dust, this may be regarded as a slight increase in the sensitivity of the Classification to the early effects of dust exposure, and therefore to be welcomed. This implies, however, that if the modified set is used, then care will have to be taken to distinguish between that increased sensitivity and any real increases in prevalence or incidence of small irregular opacities that may occur in some populations. It would be important, therefore, for the ILO to emphasize to users of the Classification that, when reporting their results, they must always state clearly which set of standards was used.

REFERENCES

Cleveland, W.S. [1994]. The Elements of Graphing Data (Revised Edition), Summit, New Jersey: Hobart Press.

Cohen, J. [1960]. A coefficient of agreement for nominal scales. Educational Psychological Measurement, 20, 37-46.

Conover, W.J. [1980]. Practical Nonparametric Statistics (2nd edition). New York, NY: John Wiley and Sons.

Duke, S.P., and Carpentieri, A.C. [1989]. Calculating the Wilcoxon signed-ranks test using Conover's method. In Proceedings of the Fourteenth Annual SAS Users Group International Conference, 1335-1336.

Guidelines for the Use of ILO International Classification of Radiographs of Pneumoconioses, Revised Edition 1980. Occupational Safety and Health Series, No. 22 (Rev.), International Labor Office, Geneva.

SAS/STAT User’s Guide, Version 6, Fourth Edition 1990, SAS Institute, Inc., Cary, NC.

Table 1. Mean small opacity profusion scores from duplicate classifications of up to 120 films, by each of 39 readers in Rounds 1 and 2, using ILO and QUAD standards

19 "IQ" READERS (ILO standards in Round 1)						20 "QI" READERS (QUAD standards in Round 1)					
		ILO		QUAD				ILO		QUAD	
Center/ reader	U[**]	Mean	n[*]	Mean	n	Center/ reader	U[**]	Mean	n	Mean	n
02 C	M	5.0	120	4.2	119	01 A	H	3.4	119	3.1	117
D	L	4.0	120	4.0	120	B	M	3.9	120	3.6	118
E[***]	H	3.5	117	3.2	120	C	M	2.9	119	3.0	120
		-	-	-	-	D	H	2.6	120	2.7	119
04 A	H	3.3	119	3.5	118	06 A	H	3.9	120	3.9	120
B	H	4.4	120	4.5	120	B	H	3.2	120	3.4	120
C	L	3.7	120	3.4	120	C	L	3.8	120	4.5	120
D	L	3.2	120	3.0	120	D	L	3.8	120	3.8	120
05 A	H	3.3	118	3.8	116	07 A	H	4.1	118	4.2	120
B	H	3.8	120	3.7	116	B	H	4.1	120	4.6	120
C	H	3.2	107	3.9	118	C	M	4.3	120	4.7	118
D	M	3.7	119	3.9	118	D	L	4.3	120	3.8	120
09 A	H	3.6	120	3.3	120	08 A	M	3.5	120	4.1	113
B	H	2.4	119	2.8	120	B	M	3.7	120	3.3	118
C	L	3.8	120	4.2	120	C	L	3.0	120	4.2	110
D	L	5.0	118	3.7	120	D	L	4.0	111	4.1	118
12 A	H	3.0	119	2.7	120	11 A	H	3.3	120	3.6	120
B	L	3.2	120	3.0	120	B	H	2.8	118	3.2	120
C	L	3.3	120	3.7	118	C	L	4.0	120	3.5	118
D	L	4.7	120	4.1	115	D	L	2.8	119	3.3	120

[*] n = number of valid classifications

[**] U = Usage of ILO Classification in 12 months preceding the trial:
L = <500 films; M = betw. 500 and 1000 films; H = >1000 films

[***] Readers A and B at Center 02 worked together during some reading sessions. Only one of those correlated sets of results, selected at random and here designated "E", has been used in the analyses.

Table 2a: ILO vs. QUADRANT Small Opacity Profusion Classifications
Tabulated for Sequence 1 (IQ) Paired Classifications of Rounds 1/2

Frequencies for paired classifications	**QUADRANT**											
	NAU-GHT	**0/1**	**1/0**	**1/1**	**1/2**	**2/1**	**2/2**	**2/3**	**3/2**	**3/3**	**3/+**	**TOTAL**
ILO												
NAUGHT	444	70	31	25	11	2	4	4				591
0/1	100	89	72	29	7	4	2		1			304
1/0	51	44	81	71	16	10	5	2				280
1/1	29	27	58	177	49	22	16	2	1			381
1/2	12	8	10	49	29	32	14	1				155
2/1	2	4	12	44	33	34	18	7				154
2/2	3		5	24	22	29	62	22	7	3		177
2/3				5	2	9	29	13	7	4		69
3/2		1			1	1	9	13	11	7		43
3/3	2			1		1	10	10	13	32	3	72
3/+						1			1	3	6	11
TOTAL	643	243	269	425	170	145	169	74	41	49	9	2237

Diagonal	Above Diagonal	Below Diagonal
43.7 %	26.0 %	30.3 %

Frequencies for single classifications	**Major Categories**			
	0	**1**	**2**	**3**
Standards				
ILO	895	816	400	126
QUADRANT	886	864	388	99

Table 2b: ILO vs. QUADRANT Small Opacity Profusion Classifications
Tabulated for Sequence 2 (QI) Paired Classifications of Rounds 1/2

Frequencies for paired classifications	QUADRANT											
	NAU-GHT	0/1	1/0	1/1	1/2	2/1	2/2	2/3	3/2	3/3	3/+	TOTAL
ILO												
NAUGHT	483	94	62	30	4	3	4					680
0/1	50	71	67	48	6	3						245
1/0	29	49	140	103	33	12	5					371
1/1	17	21	85	141	72	27	23	2				388
1/2	2	6	16	35	58	26	18	3				164
2/1	2	5	13	26	19	37	24	5	1			132
2/2	1	1	10	15	26	25	70	28	7	6		189
2/3			1	1	4	10	16	23	8	4	1	68
3/2							5	14	13	9	1	42
3/3				1	1	1	12	6	11	22	6	60
3/+									1	6	8	15
TOTAL	584	247	394	400	223	144	177	81	41	47	16	2354

Diagonal	Above Diagonal	Below Diagonal
45.3 %	31.7 %	23.0 %

Frequencies for single classifications	Major Categories			
	0	1	2	3
Standards				
ILO	925	923	389	117
QUADRANT	831	1017	402	104

Table 2c: ILO vs. QUADRANT Small Opacity Profusion Classifications
Tabulated for Paired Classifications of Rounds 1/2

Frequencies for paired classifications	**QUADRANT**											
	NAU-GHT	**0/1**	**1/0**	**1/1**	**1/2**	**2/1**	**2/2**	**2/3**	**3/2**	**3/3**	**3/+**	**TOTAL**
ILO												
NAUGHT	927	164	93	55	15	5	8	4				1271
0/1	150	160	139	77	13	7	2		1			549
1/0	80	93	221	174	49	22	10	2				651
1/1	46	48	143	318	121	49	39	4	1			769
1/2	14	14	26	84	87	58	32	4				319
2/1	4	9	25	70	52	71	42	12	1			286
2/2	4	1	15	39	48	54	132	50	14	9		366
2/3			1	6	6	19	45	36	15	8	1	137
3/2		1			1	1	14	27	24	16	1	85
3/3	2			2	1	2	22	16	24	54	9	132
3/+						1			2	9	14	26
TOTAL	1227	490	663	825	393	289	346	155	82	96	25	4591

Diagonal	Above Diagonal	Below Diagonal
44.5 %	28.9 %	26.6 %

Frequencies for single classifications	**Major Categories**			
	0	**1**	**2**	**3**
Standards				
ILO	1820	1739	789	243
QUADRANT	1717	1881	790	203

Table 2d: Round 1 vs. Round 2 Small Opacity Profusion Classifications
Tabulated for Paired Classifications of Rounds 1/2

Frequencies for paired classifications	**Round 2**											
	NAU-GHT	**0/1**	**1/0**	**1/1**	**1/2**	**2/1**	**2/2**	**2/3**	**3/2**	**3/3**	**3/+**	**TOTAL**
Round 1												
NAUGHT	927	120	60	42	13	4	5	4				1175
0/1	194	160	121	50	13	9	3		1			551
1/0	113	111	221	156	32	23	15	3				674
1/1	59	75	161	318	84	48	31	3	1	1		781
1/2	16	14	43	121	87	51	40	5		1		378
2/1	5	7	24	71	59	71	43	17		1		298
2/2	7		10	47	40	53	132	38	12	15		354
2/3				7	5	14	57	36	21	10		150
3/2		1			1	2	16	21	24	18	1	84
3/3	2			1		1	16	14	22	54	9	119
3/+						1		1	2	9	14	27
TOTAL	1323	488	640	813	334	277	358	142	83	109	24	4591

Diagonal	Above Diagonal	Below Diagonal
44.5 %	24.5 %	31.0 %

Frequencies for single classifications	**Major Categories**			
	0	**1**	**2**	**3**
Round				
1	1726	1833	802	230
2	1811	1787	777	216

Table 3. Summary statistics from 39 readers' classifications of 120 radiographs using the ILO and QUAD standards

(a) SEQUENCE 1 READERS

Center/ Reader	PAIRED CLASSIFICATIONS FROM ROUNDS 1 AND 2			
	n[*]	% with identical profusion	% with more profusion when using QUADs (in Round 2)	% with more profusion when using ILO films (in Round 1)
02 C	119	24.4	18.5	57.1
D	120	38.3	30.0	31.7
E[**]	117	57.3	14.5	28.2
04 A	118	58.5	28.8	12.7
B	120	45.0	31.7	23.3
C	120	45.0	21.7	33.3
D	120	51.7	18.3	30.0
05 A	114	53.5	33.3	13.2
B	116	38.8	24.1	37.1
C	107	47.7	45.8	6.5
D	117	34.2	35.0	30.8
09 A	120	55.0	15.0	30.0
B	119	52.9	34.5	12.6
C	120	46.7	37.5	15.8
D	118	22.0	10.2	67.8
12 A	119	50.4	16.8	32.8
B	120	43.3	24.2	32.5
C	118	44.1	30.5	25.4
D	115	21.7	25.2	53.0

[*] n = number of paired classifications

[**] Readers A and B at Center 02 worked together during some reading sessions. Only one of those correlated sets of results, selected at random and here designated "E", has been used in the analyses.

Table 3. Summary statistics from 39 readers' classifications of 120 radiographs using the ILO and QUAD standards

(b) SEQUENCE 2 READERS

Center/ Reader	PAIRED CLASSIFICATIONS FROM ROUNDS 1 AND 2			
	n[*]	% with identical profusion	% with more profusion when using QUADs (in Round 1)	% with more profusion when using ILO films (in Round 2)
01 A	116	55.2	19.0	25.9
B	118	41.5	19.5	39.0
C	119	56.3	22.7	21.0
D	119	60.5	26.1	13.4
06 A	120	45.0	31.7	23.3
B	120	60.8	28.3	10.8
C	120	35.8	52.5	11.7
D	120	40.8	27.5	31.7
07 A	118	42.4	33.0	24.6
B	120	37.5	47.5	15.0
C	118	36.4	44.9	18.6
D	120	44.2	17.5	38.3
08 A	113	45.1	39.8	15.0
B	118	45.8	12.7	41.5
C	110	29.1	61.8	9.1
D	110	37.3	32.7	30.0
11 A	120	36.7	42.5	20.8
B	118	61.9	23.7	14.4
C	118	39.0	18.6	42.4
D	119	52.9	32.8	14.3

[*] n = number of paired classifications

Table 4a: ILO vs. QUADRANT Primary Shape Classifications for Sequence IQ*
For Paired Classifications in Rounds 1 and 2

Frequencies for pairs of classifications	QUADRANT			
	Blank	Rounded	Irregular	Total
ILO				
Blank	438	52	93	583
Rounded	79	679	142	900
Irregular	127	110	517	754
Total	644	841	752	2237

Table 4b: ILO vs. QUADRANT Primary Shape Classifications for Sequence QI*
For Paired Classifications in Rounds 1 and 2

Frequencies for pairs of classifications	QUADRANT			
	Blank	Rounded	Irregular	Total
ILO				
Blank	483	55	144	682
Rounded	34	517	151	702
Irregular	68	107	795	970
Total	585	679	1090	2354

Table 4c: ILO vs. QUADRANTS For Primary Shape Classifications in Rounds 1 and 2

Frequencies for pairs of classifications	QUADRANT			
	Blank	Rounded	Irregular	Total
ILO				
Blank	921	107	237	1265
Rounded	113	1196	293	1602
Irregular	195	217	1312	1724
Total	1229	1520	1842	4591

*Sequence IQ readers are those who classified films using the ILO standards in round 1 and using the experimental QUADRANT standards in round 2; sequence QI readers are those who classified films using the QUADRANT standards in round 1 and using the ILO standards in round 2.

Tables 5a-5d: Round By Primary Shape Tables for the Four Subsequences of Readings*

Table 5a: Subsequence IQI Frequencies with Row Percentages

ROUND	**PRIMARY SHAPE**			
	Blank	Rounded	Irregular	Total
Round 1 (ILO)	269 24%	498 45%	343 31%	1110
Round 2 (QUAD)	318 29%	466 42%	326 29%	1110
Round 3 (ILO)	279 25%	448 40%	382 35%	1109
Total	866	1412	1051	3329

Table 5b: Subsequence IQQ Frequencies with Row Percentages

ROUND	**PRIMARY SHAPE**			
	Blank	Rounded	Irregular	Total
Round 1 (ILO)	311 28%	395 36%	406 36%	1112
Round 2 (QUAD)	325 29%	367 33%	420 38%	1112
Round 3 (QUAD)	292 26%	329 30%	491 44%	1112
Total	928	1091	1317	3336

*Subsequence IQI readings comprise the triad of classifications by sequence 1 readers where the round 3 reading is made using the ILO standards. Subsequence IQQ readings comprise the triad of classifications by the same sequence 1 readers where the round 3 reading is made using the QUADRANT standards. Similarly for sequence 2 readers.

Table 5c: Subsequence QIQ Frequencies with Row Percentages

ROUND	PRIMARY SHAPE			
	Blank	Rounded	Irregular	Total
Round 1 (QUAD)	250 24%	283 27%	522 49%	1055
Round 2 (ILO)	272 26%	304 29%	479 45%	1055
Round 3 (QUAD)	244 23%	321 30%	490 47%	1055
Total	766	908	1491	3165

Table 5d: Subsequence QII Frequencies with Row Percentages

ROUND	PRIMARY SHAPE			
	Blank	Rounded	Irregular	Total
Round 1 (QUAD)	212 20%	330 32%	491 48%	1033
Round 2 (ILO)	268 26%	333 32%	432 42%	1033
Round 3 (ILO)	255 25%	324 31%	454 44%	1033
Total	735	987	1377	3099

Tables 6a and 6b: Standards by Primary Shape Tables For Each Sequence When Primary Shape Is Specified*

Table 6a: Sequence IQ Frequencies with Row Percentages

Standards Used	**PRIMARY SHAPE**		
	Rounded	Irregular	Total
ILO	1341 54.3%	1131 46.7%	2472
QUADRANT	1162 48.4%	1237 51.6%	2399
Total	2503 51.4%	2368 48.6%	4871

Table 6b: Sequence QI Frequencies with Row Percentages

Standards Used	**PRIMARY SHAPE**		
	Rounded	Irregular	Total
ILO	961 41.3%	1365 58.7%	2326
QUADRANT	934 38.3%	1503 61.7%	2437
Total	1895 39.8%	2868 60.2%	4763

*The classifications where primary shape was left blank are excluded from these calculations. Also, these are tabulations of single classifications, not of paired classifications.

Table 7. Summary of paired small opacity profusion classifications from Rounds 1 and 2, and distributions onto the 4-point scale, by Sequence and readers' nominations of predominant shapes of small opacities in both Rounds; data from Appendix VII, Tables AVII.1-6

Sequence	Predominant shape in both Rounds	No. of paired classifications	Percent with profusion higher using -			Standard used	% distribution to categories				Data in Appendix VII, Table:	This Table Section No:
			-QUADs (above diagonal)	-ILO (below diagonal)	- NEITHER (on diagonal)		0	1	2	3		
IQ (ILO in Round 1)	irregular	517	33.3	29.8	36.9	ILO	18.0	55.9	21.1	5.0	1	1
						QUAD	14.9	58.4	22.2	4.4		
	rounded	679	25.9	35.5	38.6	ILO	9.6	43.4	33.3	13.7	3	2
						QUAD	8.2	49.8	31.1	10.9		
	remainder	1041	22.4	27.2	50.4	ILO	70.8	22.3	6.2	0.7	5	3
						QUAD	72.3	21.5	6.0	0.2		
QI (QUAD in Round 1)	irregular	795	40.1	23.1	36.7	ILO	15.3	63.3	17.4	4.0	2	4
						QUAD	10.2	68.0	18.2	3.5		
	rounded	517	30.2	29.8	40.0	ILO	6.4	39.4	37.9	16.2	4	5
						QUAD	4.8	41.0	39.6	14.5		
	remainder	1042	25.9	19.7	54.4	ILO	73.9	20.7	5.3	0.1	6	6
						QUAD	69.6	25.3	5.0	0.12		
Both Sequences; all shapes		4591	28.9	26.6	44.5	ILO	39.6	37.9	17.2	5.3	This report: Table 2c	7
						QUAD	37.4	41.0	17.2	4.4		

Table 8. Mean Profusion Difference Scores (PDS*) per 100 film classifications, for 8 combinations of Film reading Sub-sequences and Rounds Involved

Rounds Involved	Sequence 1 (19 reader)		Sequence 2 (18 readers)		All 4 sub-sequences
	IQI	IQQ	QIQ	QII	
1 & 2	-7.0	-9.3	6.2	11.8	0.2
1 & 3	-	15.9	-	11.2	13.6
2 & 3	-5.8	-	2.6	-	-1.7
All Rounds	-6.4	3.3	4.4	11.5	
	-1.6		7.9		3.1

*Positive values of PDS indicate more profusion when using the QUADRANT standards.

Table 9. Analyses of Covariance of Profusion Difference Scores (PDS) and SHAPE (S).

Source of Variability	Sequence 1 (IQ)					Sequence 2 (QI)				
	DF	SS*	MS	F	P	DF	SS*	MS	F	P
Readers (R)	18	87379	4854	3.73	<.0001	17	56041	3297	2.41	<.006
Reproducibility (κ)	1	1322	1322			1	166	166		
Variations between readers in the effect of reproducibility (R x κ)	18	35721	1985			17	24176	1422		
Shape (S)	1	17094	17094	19.97	<.0001	1	5007	5007	5.25	<.03
Residual	37	31669	856			35	33398	954		
Percent variance accounted for					62.9					43.0
Residual SD					29					31
SHAPE regression coeficient (and standard error)					598.5 (133.9)					513.2 (224.0)
Weighted mean SHAPE coefficient (standard error)					576.1 (115.0)					

* "Sequential (Type I)" Sums of Squares, calculated using SAS® Software. Explanatory variables were entered into the equation in the order indicated.

Table 10. Mean PDS residuals, by Film Reading Sub-sequence and Rounds Involved

ROUNDS INVOLVED	SEQUENCE 1 (19 reader)		SEQUENCE 2 (18 readers)	
	IQI	IQQ	QIQ	QII
1 & 2	-0.12	-1.90	-3.57	-0.30
1 & 3	-	1.90	-	0.30
2 & 3	0.12	-	3.57	-

Table 11a-d: Frequencies by Major Categories for the Four Subsequences*

Table 11a: Subsequence IQI

ROUND	**MAJOR CATEGORIES**			
	0	1	2	3
Round 1 (ILO)	419	389	222	80
Round 2 (QUAD)	411	415	220	64
Round 3 (ILO)	427	374	249	60

Prevalence of Small Opacities on Four-point Scale

ILO	QUADRANT
1374/2220 (61.9%)	699/1110 (63.0%)

Table 11b: Subsequence IQQ

ROUND	**MAJOR CATEGORIES**			
	0	1	2	3
Round 1 (ILO)	473	417	176	46
Round 2 (QUAD)	471	439	167	35
Round 3 (QUAD)	420	452	192	48

Prevalence of Small Opacities on Four-point Scale

ILO	QUADRANT
639/1112 (57.5%)	1333/2224 (59.9%)

*Subsequence IQI readings comprise the triad of classifications by sequence 1 readers where the round 3 reading is made using the ILO standards. Subsequence IQQ readings comprise the triad of classifications by the same sequence 1 readers where the round 3 reading is made using the QUADRANT standards. Similarly for sequence 2 readers.

Table 11c: Subsequence QIQ

ROUND	MAJOR CATEGORIES			
	0	1	2	3
Round 1 (QUAD)	360	475	184	36
Round 2 (ILO)	384	449	182	40
Round 3 (QUAD)	358	478	186	33

Prevalence of Small Opacities on Four-point Scale

ILO	QUADRANT
671/1055 (63.6%)	1392/2110 (66.0%)

Table 11d: Subsequence QII

ROUND	MAJOR CATEGORIES			
	0	1	2	3
Round 1 (QUAD)	336	468	168	61
Round 2 (ILO)	377	414	173	69
Round 3 (ILO)	380	412	173	68

Prevalence of Small Opacities on Four-point Scale

ILO	QUADRANT
1309/2066 (63.4%)	697/1033 (67.5%)

Overall Prevalence of Small Opacities on Four-point Scale
Combining All Four Tables

ILO	QUADRANT
3993/6453 (61.9%)	4121/6477 (63.6%)

Table 12: Tabulation of ILO Vs. QUADRANT Classifications With Respect To Large Opacities For Paired Classifications in Rounds 1 and 2

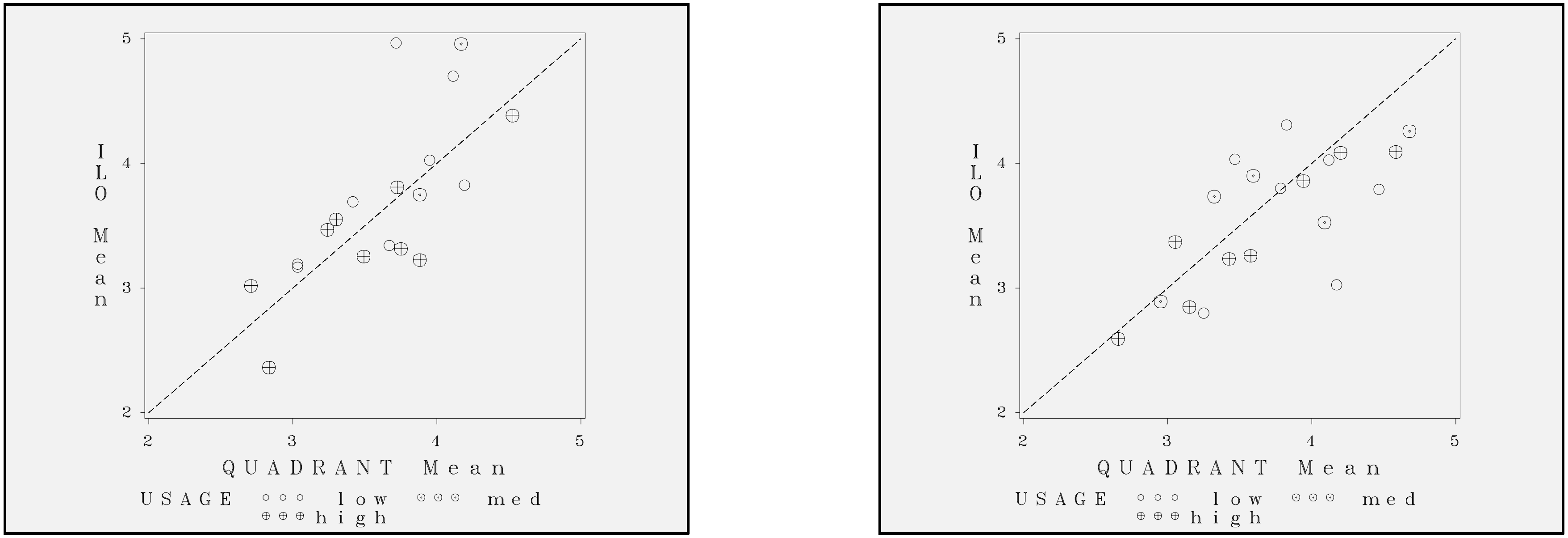
ILO (1980) Standards	**QUADRANT Standards**				
	NO LARGE OPACITIES	A	B	C	TOTAL
NO LARGE OPACITIES	4270	27	2		4299
A	51	39	7		97
B	4	8	35	8	55
C			8	70	78
TOTAL	4325	74	52	78	4529

Percentage On Diagonal	Above Diagonal	Below Diagonal
97.5	1.0	1.6

Table 13. Readers' classifications of large opacities with one set of standards but not the other; Rounds 1 and 2

SEQUENCE	NUMBER OF READERS INVOLVED	LARGE OPACITIES RECORDED WITH -		ALL DISCORDANT CLASSIFICATIONS
		- ILO standards, but not with QUADs	- QUADs, but not with ILO standards	
1 (IQ)	15	31	14	45
2 (QI)	15	24	15	39
ALL READERS		55	29	84

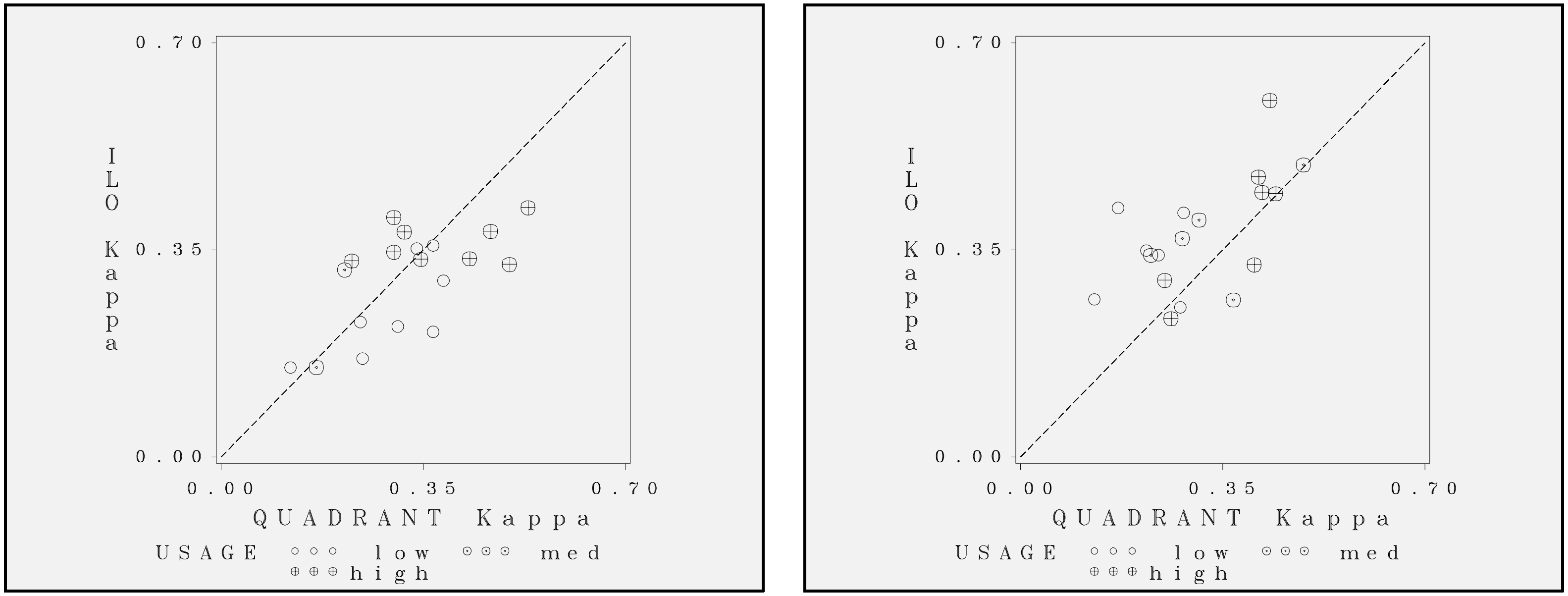
Figures 1(a) and 1(b). Mean Small Opacity Profusion Scores from Paired Classifications of Rounds 1 and 2 for the 19 Sequence IQ and the 20 Sequence QI Readers.



(a) Sequence IQ

(b) Sequence QI

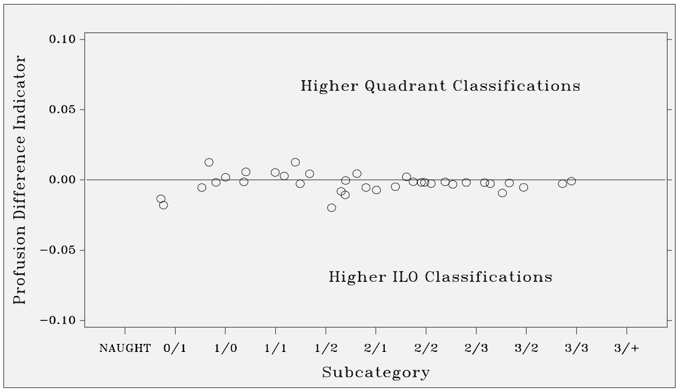
Figures 2(a) and 2(b). The ILO Kappa Vs. the QUADRANT Kappa for the Readers Who Completed Round 3.



(a) Sequence IQ

(b) Sequence QI

Figure 3(a). Symmetry Plot for Table 2(a).



An Explanation of the Symmetry Plot

The symmetry plot is a variant of the Tukey mean-difference plot (see Cleveland 1993, page 130) and is useful for indicating whether a systematic difference in readings is occurring uniformly across the subcategories of the classification scheme. Each Profusion Difference Indicator (PDI) is a weighted difference between the frequencies of two cells which are positioned symmetrically above and below the main diagonal of Table 2 when that table is rotated anti-clockwise through 45°. First, differences are calculated as (numbers vertically above) minus (numbers vertically below) the rotated diagonal. (Zero differences are not used or shown on the graphs.) The weighting factor associated with each PDI is a ratio of the absolute difference (on the 11-point scale) of paired classifications and the total number of paired classifications. The weights are proportional to the distance between the mid-points of the two cells involved. Each PDI is plotted against a (theoretical) point on a linear representation of the 11 categories identified by the rotated diagonal. The point is defined by the intersection of (a) a line joining the mid-points of the two cells and (b) a line drawn through the mid-points of the diagonal cells.
AN EXAMPLE: The second entry in the first row of Table 2a shows 70 paired classifications where category "naught" was recorded using the ILO standards and category 0/1 with the QUADs. The symmetrically positioned cell below the diagonal shows 100 pairs of classifications where category "naught" was recorded using the QUADs and category 0/1 with the ILO standards. The corresponding PDI is calculated as (70 - 100) multiplied by the weight of |1 - 2| / 2237 to give -0.0134. This is the y-value of the first point plotted at the left of the figure, where the x- co-ordinate is shown as a point between categories "naught" and 0/1. In order to avoid overplotting, the plot symbols have been ‘jittered’ (that is, a small random quantity has been added to the x- and y-values of each plotted point).

Figure 3(b). Symmetry Plot for Table 2(b).

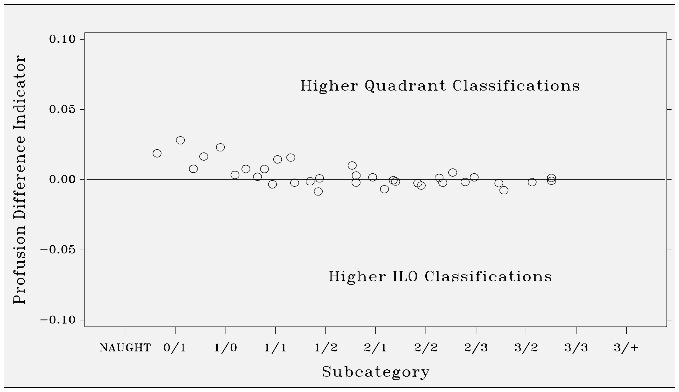


Figure 3(c). Symmetry Plot for Table 2(c).

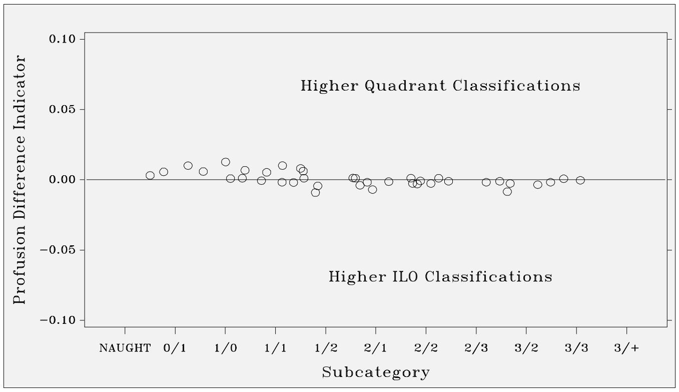
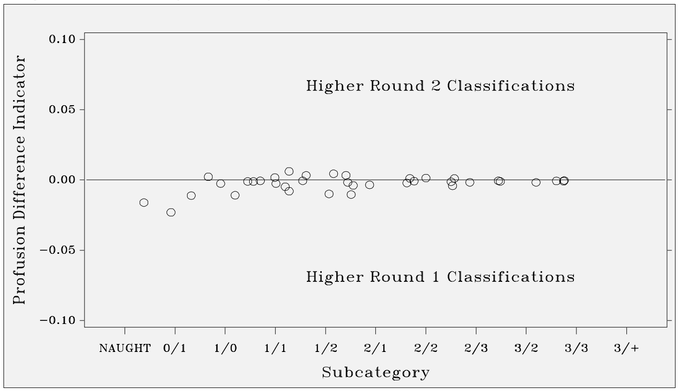
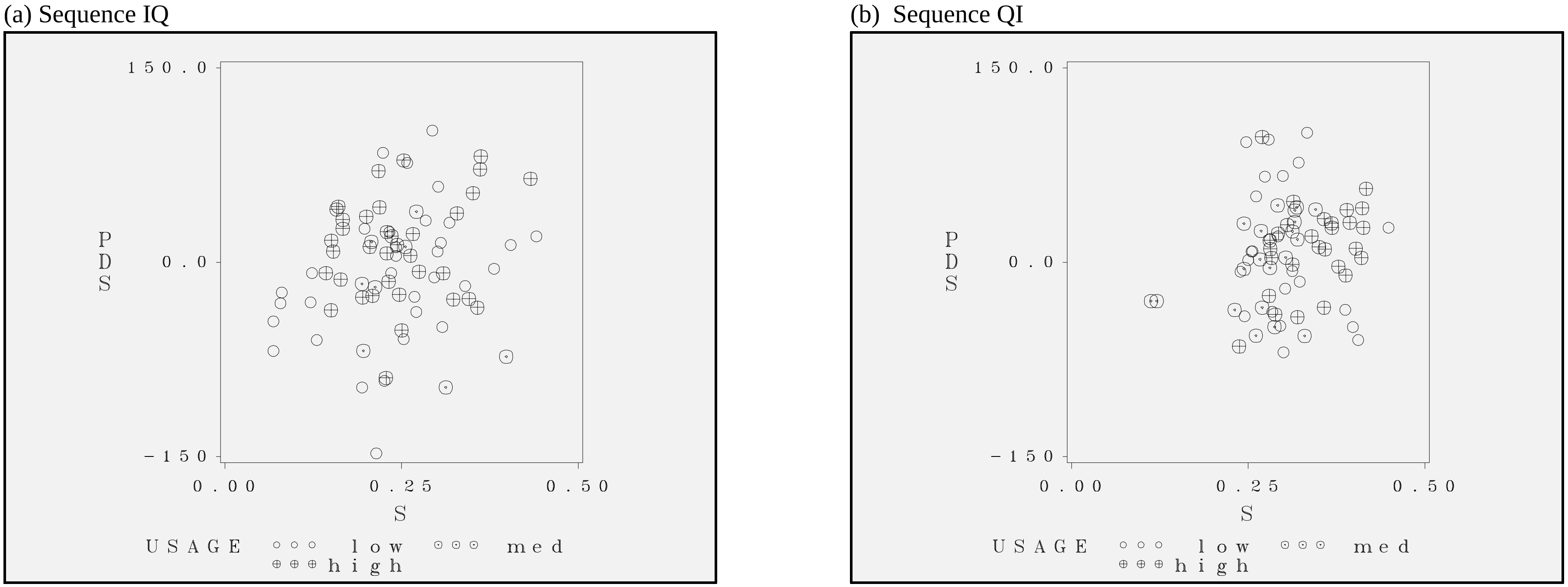


Figure 3(d). Symmetry Plot for Table 2(d).



Figures 4(a) and 4(b). Scatterplots of Profusion Difference Score (PDS)* and Shape Statistic S** for Each Sequence Group.


(a) Sequence IQ

(b) Sequence QI

*Positive values indicate more profusion with the QUADs, negative values more profusion with the ILO standards.

**S = number of predominantly irregular small opacity classifications by a reader when using the QUAD standards expressed as a fraction of all classifications involving some small opacities.

The labels indicate readers' usage of the ILO Classification during the 12 months preceding the trial.

Delivering on the Nation's promise:
Safety and health at work for all people
Through research and prevention