Quality control of Kato slide counts for *Schistosoma mansoni*: a review of 12 years’ experience in Kenya

R.F. Sturrock,¹ J.H. Ouma,² H.C. Kariuki,² F.W. Thiongo,² D.K. Koech,³ & A.E. Butterworth⁴

A total of 19 annual or biannual audits were performed over a 12-year period by an independent microscopist on randomized subsamples of Kato slides examined for Schistosoma mansoni eggs by Kenyan microscopists from the Division of Vector-borne Diseases (DVBD). The recounts were invariably lower than the originals owing to some deterioration of the preparations between counts, but the two were strongly correlated: significant regressions of recounts on counts taking up 80–90% of the observed variance. Observer bias differed significantly between microscopists but remained stable over time, whereas repeatability of recounts on counts dropped slightly in periods of maximum work load but did not vary systematically with time. Approximately 7% of the counts and recounts disagreed on the presence or absence of eggs, but less than a third of these were negatives that were found positive on recount. False negatives dropped to 1.3% if duplicate counts were considered. The performance of the Kenyan microscopists was remarkably high and consistent throughout the 12-year period. This form of quality control is suitable for projects where limited funds preclude full-time supervisors using more sophisticated systems.

Introduction
Few would question the desirability of checking the performance of technical staff involved in the monotonous and repetitive task of counting *Schistosoma mansoni* eggs in faecal preparations. For all their imperfections, such counts remain the only means of estimating, however crudely, the intensity of schistosome infections and the magnitude of any changes following interventions such as treatment. Jordan (1) illustrated a method developed on Saint Lucia for calculating the “false-negative rate” (FNR), based on extra infections detected on re-examination of 10% of slides “negative” at the first examination. Later, the method was modified, but still re-examined only a fraction of slides declared to

---

¹ Senior Lecturer, Department of Medical Parasitology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, England. Correspondence should be addressed to this author.
² Division of Vector-borne Diseases, Ministry of Health, Nairobi, Kenya.
³ Kenya Medical Research Institute, Nairobi, Kenya.
⁴ Department of Pathology, University of Cambridge, Cambridge, England.

Reprint No. 5803

Materials and methods

Preparation and counting of Kato slides; selection and recounting of a random subsample

Studies, approved by the appropriate Kenyan authorities and conforming with the prevailing ethical requirements, involved numerous faecal surveys from 1981 onwards on whole communities, individual schools, or specified study populations or cohorts from the study areas in the Machakos and Makuene Districts (4–6). The following system of quality control audits was developed during the initial years of these studies.

At each survey, depending on the objectives of the study, one or more (to a maximum of five) stools were collected from each individual by mobile field teams comprising a driver/supervisor and two assistants, if possible within a 5–7-day period although sometimes longer. The field team plus an extra assistant prepared duplicate 50mg Kato smears from each faecal sample. The slides were labelled with the name and unique identity number (ID) of the subject, as well as the date and, if appropriate, the study code. Duplicate slides, labelled A and B, were stored in labelled, closable wooden or plastic slide boxes, together with a complete list (Form I) of the slides with columns for entering the observed counts. The A and B slides were placed sequentially in boxes holding 50 slides in one row, or side by side in adjacent rows in larger boxes holding 100 slides. The boxes were stored in a large plastic bag to minimize desiccation and, in later years, sprayed with a commercial aerosol insecticide before closing. With this system, a field team could collect and process 100 to 120 stool samples a day.

The boxes were transported to a local laboratory for microscopic examination by which time the slides were at least 24h old. The basic sampling unit was the slide box. Two microscopists, initially designated A and B, counted the A and B slides, respectively. To minimize fatigue, the microscopists were encouraged to take a short break every one-and-a-half to two hours. When the work load increased in 1989, an additional microscopist (designated C) was added and amended the appropriate column on Form I, depending on which duplicate he counted. Slides from any box were examined by only two of the three microscopists: all A or B slides were completed before passing the box to the other microscopist. Ideally, the second microscopist should not know the count of the first but this was impossible in practice: the second microscopist usually completed his count before looking at the first count.

When counting was completed, a random 10% sample of slides was chosen using a standard sampling table, initially by H.C.K. or F.W.T., but later by the microscopists. The selected slides were placed in another slide box together with a second list (Form II) which recorded the names of the two microscopists involved; the ID number, name of the subject and the count for each selected slide; and the duplicate slide count. These boxes were resprayed and stored in plastic bags until required. Copies of completed Forms I were sent to Nairobi for data entry on spreadsheets (at first, SuperCalc 3 for Apple; later, Excel for Macintosh). In periods of peak activity, each microscopist counted 80–100 slides a day, besides selecting the 10% subsample and completing the paperwork associated with the counting and subsampling.

Quality control audits were performed once or twice each year, when R.F.S. visited Kenya. All boxes of selected slides that had accumulated since the previous visit were brought to Nairobi and marked in numerical order — the maximum being 46 boxes in 1992. If there were too many slides to recount them all, a stratified subsample was chosen by selecting every 2nd, 3rd, 4th or 5th slide from box 1. The initial slide was chosen at random from slides 1, 2, 3, 4 or 5 as appropriate and the counting sequence was continued through box 2 onwards. Thus, at least a fifth of the initial 10% sample was selected for recounting (i.e., 2–10% of the original slides). The recounts were entered on a third form (Form III), together with the identity of the original microscopist, his counts, and the counts for the matching slides examined by the other DVBD microscopist as listed on Form II.

Data handling and analytical techniques

Initially, the raw data (counts and recounts) on Form III were processed manually with the aid of a pocket calculator. Later, the data were entered on statistical spreadsheets (at first, Nanostat; later, Minitab) for preparation of summary tables and for correlation and regression analyses of the original DVBD counts and the recounts. In addition, 2 × 2 tables were prepared for the individual observers, using the recount as the reference standard, showing the number of slides where the original counts and the recounts agreed and disagreed as $a = +, b = +, c = -+$ and $d = --$, where + and − denote eggs present and absent, respectively. Three indices were calculated from these tables (11): percent reproducibility calculated as 100$a / (a + b + c)$; percent test (observer) bias as 100 $(a + b) / (a + c)$; and trends in the discrepant counts $b$ and $c$ using a simplified sign test (12). The last-mentioned indicated the probabil-
ity of obtaining whichever was the larger, b or c, of the combined discrepant counts (b + c). Discrepant counts were further examined to determine whether the duplicate DVBD count agreed with the recount or the original count. The most serious fault, a positive case missed by both DVBD microscopists, was expressed as the percentage serious error (PSE), i.e. the percentage of such cases from all slides re-examined.

The main data set comprised all indices (except the PSE for the first four audits) for microscopists A and B for 19 sequential audits from 1984 to 1996 plus those for C for the last 10 audits from 1989 to 1996. In addition, two subsets were analysed: microscopists A and B for all 19 audits and all three microscopists for the last 10 audits. Analyses of these subsets are reported only if they differed from the main analyses.

### Results

In all, 10113 slides were recounted between 1984 and 1996, when it is estimated that the DVBD microscopists examined over 350000 Kato slides. Recounts were performed between 1 and 18 months after the slides had been prepared. On average, 211 slides per microscopist were re-examined at each audit, and the minimum usually exceeded 100 (Table 1). There was no significant difference between the number of slides per observer (F_{2,45} (2 against 45 degrees of freedom) = 0.27; P > 0.05) but the numbers recounted at the different audits (Fig. 1a) varied significantly (F_{10,29} = 9.70; P < 0.001). There was no significant linear trend with time (F_{1,46} = 0.28; P > 0.05). The overall proportion of discrepant counts (N, see Table 1) was 6.8% and was highest for microscopist A. Of the discrepant counts at almost every audit, more slides were recorded as positive by the DVBD microscopists (and negative by RFS) than were recorded as negative by the DVBD microscopists (and positive by RFS). A sign test showed this trend to be significant in 7/19 audits for microscopist A, in 1/19 for microscopist B, and never for microscopist C.

The performance of the DVBD microscopists over a period of time, especially their quantitative performance, was of interest. Both the mean counts

---

**Table 1: Number of slides examined with the number and percentage of discrepant counts by audit (time) and microscopist, and the percentage of serious errors (PSE) by audit**

<table>
<thead>
<tr>
<th>Microscopist A:</th>
<th>Microscopist B:</th>
<th>Microscopist C:</th>
<th>PSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>Year</td>
<td>n</td>
<td>x/N</td>
</tr>
<tr>
<td>1</td>
<td>1984</td>
<td>146</td>
<td>16/18*</td>
</tr>
<tr>
<td>2</td>
<td>1985</td>
<td>314</td>
<td>30/38*</td>
</tr>
<tr>
<td>3</td>
<td>1985</td>
<td>321</td>
<td>10/16</td>
</tr>
<tr>
<td>4</td>
<td>1986</td>
<td>147</td>
<td>6/9</td>
</tr>
<tr>
<td>5</td>
<td>1986</td>
<td>325</td>
<td>16/17*</td>
</tr>
<tr>
<td>6</td>
<td>1987</td>
<td>102</td>
<td>5/6</td>
</tr>
<tr>
<td>7</td>
<td>1988</td>
<td>194</td>
<td>7/11</td>
</tr>
<tr>
<td>8</td>
<td>1988</td>
<td>163</td>
<td>6/9</td>
</tr>
<tr>
<td>9</td>
<td>1989</td>
<td>289</td>
<td>5/10</td>
</tr>
<tr>
<td>10</td>
<td>1989</td>
<td>174</td>
<td>15/16*</td>
</tr>
<tr>
<td>11</td>
<td>1990</td>
<td>246</td>
<td>5/10</td>
</tr>
<tr>
<td>12</td>
<td>1990</td>
<td>176</td>
<td>13/15*</td>
</tr>
<tr>
<td>13</td>
<td>1991</td>
<td>175</td>
<td>8/12</td>
</tr>
<tr>
<td>14</td>
<td>1992</td>
<td>450</td>
<td>23/32*</td>
</tr>
<tr>
<td>15</td>
<td>1993</td>
<td>241</td>
<td>10/14</td>
</tr>
<tr>
<td>16</td>
<td>1994</td>
<td>274</td>
<td>17/20*</td>
</tr>
<tr>
<td>17</td>
<td>1995</td>
<td>206</td>
<td>8/12</td>
</tr>
<tr>
<td>18</td>
<td>1995</td>
<td>165</td>
<td>13/20</td>
</tr>
<tr>
<td>19</td>
<td>1996</td>
<td>215</td>
<td>11/16</td>
</tr>
</tbody>
</table>

* n = number of slides recounted.  
* x = larger of two discrepant counts (b = + or c = --, see text) and N = (b + c).  
* % = 100(N/n).  
* PSE % = percentage of serious errors (infected subjects missed by both DVBD microscopists).  
* P < 0.01 that values of x occurred by chance for the relevant N.  
* P < 0.05 that values of x occurred by chance for the relevant N.
and recounts (Fig. 1b) varied over time ($F_{18,29} = 15.81$ for DVBD and 17.03 for RFS, $P < 0.001$ in both cases). There was no significant regression of these counts on time for the main data set, but it was significant for the subset from all three microscopists between audits 10 and 19 (DVBD = 19.4–0.78 × time; $F_{19,29}$ for fitting the slope ($b$) = 8.39, $P < 0.01$ and RFS = 14.7–0.60 × time; $F_{19,29}$ for fitting $b$ = 9.03, $P < 0.001$). The percentage of variance ($\text{% var}$) accounted for by fitting the regression was 20% in each case. More importantly, the observer effect was not significant ($F_{7,245} = 0.06$ for DVBD and 0.10 for RFS; $P > 0.05$ in both cases). Overall, there was a significant positive regression of the recount on the count (RFS = 0.77 × DVBD−0.482, $F_{1,46}$ for $b$ = 511.6, $P < 0.001$; $\text{% var}$ = 92%).

The $\text{% var}$ of counts on recounts at each audit (Fig. 1c) did not vary over time ($F_{18,29} = 0.73, P > 0.05$), but there was a significant difference between observers ($F_{2,45} = 3.30, P < 0.05$). Overall, it was significantly higher for microscopists A and B than for C (89.2%, 90.7% and 84.5%, respectively) although there was no difference between microscopists A and B for all 19 audits ($F_{1,36} = 0.57, P > 0.05$), or among all three for audits 10 to 19 ($F_{2,27} = 2.52, P > 0.05$).

The mean correlation ($r$) and regression ($b$) coefficients for the counts and recounts at each audit (Fig. 1c) showed no observer effect ($F_{2,45} = 0.04$ for $r$ and 0.50 for $b$, $P > 0.05$ in both cases). The values of $r$ varied with time ($F_{18,29} = 2.68, P < 0.001$) but not linearly ($F_{1,46}$ for $b = 0.01, P > 0.05$). Although the time effect for $b$ was not significant overall ($F_{18,29} = 1.95, P > 0.05$), it was significant for the three microscopists between audits 10 and 19 ($F_{9,36} = 2.61, P < 0.05$), but the relationship was not linear ($F_{1,28}$ for $b = 1.29, P > 0.05$). The overall linear regression of $r$ on $b$ was significant ($r = 0.775 + 0.222b; F_{1,47}$ for $b = 7.40, P < 0.01; \text{% var} = 12\%$).

Fig. 1d shows test (observer) bias and repeatability. Test bias did not vary significantly over time ($F_{18,29} = 0.93, P > 0.05$) but varied significantly with observer ($F_{2,47} = 7.28, P < 0.001$): microscopist A consistently reported a significantly higher percentage (108.8%) of positive slides, compared with the recounts, than microscopist B (102.3%). Observer C (105.4%) fell between the two. Repeatability did not differ between observers ($F_{2,47} = 0.51, P > 0.05$) but did vary over time ($F_{18,29} = 5.24, P < 0.001$). However, the time trend was not linear ($F_{1,47}$ for $b = 0.14, P > 0.05$). There was a negative linear regression of repeatability on test bias, i.e. repeatability improved as the bias fell (Repeatability = 133–0.451× Test bias; $F_{1,47}$ for $b = 13.4, P < 0.001; \text{% var} = 23\%$).

**Discussion**

In view of the results presented above, it may seem unusual to use the R.F.S. recounts as the reference
standard for assessing the performance of the DVBD microscopists. Usually the reference or "gold" standard is the best, absolute test to which others are compared. At the start of these studies it was not possible to say who would be best at detecting *S. mansoni* eggs; nor was it certain that the same DVBD microscopists would be employed throughout the programme. By using R.F.S. as the reference standard, it was possible first to introduce additional microscopists if necessary, although, in fact, only one was needed; and secondly to monitor any changes in the performances of the microscopists over a period of time on the assumption that R.F.S.'s performance remained essentially constant with time.

With few exceptions, the DVBD microscopists detected more *S. mansoni* eggs than R.F.S. They might have been reporting artifacts as eggs, but the more likely explanation was undoubtedly the deterioration of the Kato preparations between the initial count and the recount. Deterioration was mainly due to the slides drying out, which was most noticeable among slides prepared in hot, dry weather, allowing drying before storage in the slide boxes. Similar climatic conditions during counting added to the problem when, additionally, unavoidable heat from the microscope lamps aggravated the problem.

To minimize drying, the microscopists were instructed to keep the boxes closed except when removing or replacing slides. Storing the boxes in plastic bags was an additional precaution. At worst, completely dried-out slides were black and uncountable but the proportion never exceeded 0.5% at any audit. Normally, drying occurred in only localized patches and sufficient of the preparation remained clear enough to detect *S. mansoni* eggs, even when subjective estimates at recount indicated that more than 75% of the area was uncountable. In a few cases, darkened slides were rehydrated using standard glycerol solution but this was logistically impractical for routine use in the procedure we were using; in hotter, drier conditions, rehydration may be needed even for the initial counts (13).

Another problem related to drying was the overclearing of *S. mansoni* eggs. Great care was required to ensure that they were not overlooked during recounts. Other minor problems included the presence of mites, fungus growth and dust. Mites ate portions of the Kato preparation but the damage was always restricted to the periphery of a preparation. Insecticides kept the problem to a minimum. Fungal growth was localized and rare but, at worst, obscured only a small portion of any preparation. Dust scratches on the surface of the cellophane cover-slips made recounting more difficult but never impossible.

The mean DVBD counts varied from audit to audit, but their significant decline with time between audits 10 to 19 is mirrored in the recounts and was thus not due to a deterioration in the ability of the microscopists to detect eggs: late 1989 and 1990 (audits 10 and 11) coincided with pretreatment examinations of heavily infected populations: the drop in subsequent counts reflects the effects of treatment. The correlation of the counts and recounts was high throughout for each DVBD observer. Significant variations over time did not indicate systematic improvement or deterioration in performance. The regression of recounts on counts did not vary significantly by DVBD microscopists and remained consistent over time. Test bias for the DVBD microscopists versus R.F.S. differed significantly between the three but was always greater than 100% and remained consistent over time. In contrast, repeatability was always below 100%, varying significantly (but not consistently) over time but not between microscopists. These two indices of performance were calculated differently: the inverse relationship between them, though significant, was relatively weak.

Variations between the performances of different microscopists is to be expected. The introduction of the third microscopist when the work load increased in 1989 also increased the variability in the data. This coincided with a period of maximum activity involving approximately 50,000 slides a year, many with high counts. Some slight drop in performance under this pressure is understandable. This period was when one subset showed a significant regression of *r* on *b*, but it was relatively weak and the performance of the microscopists remained acceptable. Conversely, it is important that the microscopists maintain an equally high standard to detect scarce eggs after treatment campaigns when a very high proportion of Kato slides contained no eggs. These data show no evidence of any serious variations in the microscopists’ performance or ability to detect eggs throughout the entire period.

The remaining question concerns the number of false negatives reported by the DVBD microscopists. The false-negative rate (FNR) calculated by Jordan (I) is equivalent to \( c / (a + b + c) \) in the 2 × 2 table format (see under Materials and methods). Calculated in this way, the FNR for the Kenyan data was 4.5%, certainly as good as, if not better than, the Saint Lucian figures. Only 2.18% of all the Kenyan slides re-examined were misclassified as "false negatives". Of these slides, 40% were diagnosed as positive by the DVBD microscopist examining the duplicate slide, reducing the percentage of serious errors (PSE, i.e. infected subjects misdiagnosed as
uninfected) to 1.3%. Most discrepant counts and recounts involved only 1 or 2 eggs on a Kato preparation, i.e. light infections less than 50 eggs/g. For the few higher counts, the identification numbers on the slides did not match those on the accompanying Form II, suggesting clerical errors (which proved to be the case where the original (Form I) records could be checked). Thus, the true PSE was probably less than 1.3%. Since three or more stool samples were examined in critical studies, the PSE was probably <1%, an acceptable misclassification rate for further, quantitative analyses of prevalence and intensity.

Intensity analyses are quantitative and use quantitative egg count estimates based on sub-samples drawn from the original stool samples. Egg counts are subject to numerous sources of bias and error so that excessive emphasis cannot be placed on their absolute value. However, changes in their relative values with time can still be informative, as long as they reflect the influence of a factor being investigated and are not merely artifactual variations in the accuracy of the counts. In this respect, it is of little consequence that the recounts were almost always lower than the original counts in these studies (probably for reasons already discussed) or that individual DVBD microscopists may have tended to count higher or lower than each other. The important point is that, within reasonable limits, these differences remained consistent. Quantitative proof of this consistency is not easy to obtain in the context of an ongoing field research programme such as the present one in Kenya. The high levels of correlation (r) of the counts and recounts are indicative, as is the consistency of the regression of counts on recounts (b). However, the consistency of the more easily obtained qualitative indices and the lack of serious divergences over time gives added confidence in the suitability of the original counts for quantitative analyses.

Acknowledgements
This work was funded by grants from the Edna McConnell Clark Foundation; the Medical Research Council; the Rockefeller/WHO Joint Funding Venture, and the European Community’s Science and Technology for Development Programmes (STD2/STD3); and received practical support from the Ministry of Health of the Government of Kenya. We gratefully acknowledge the excellent assistance of our field and technical staff and the communities where the studies were performed. We are especially indebted to the untiring and diligent technical assistance of our three microscopists, Edward W. Thindiv, Julius M. Mwakesi, and Sabiano O. Odero. This paper is published with the kind permission of the Director of Medical Services of the Government of Kenya.

Résumé
Contrôle de la qualité des numérations d’œufs de Schistosoma mansoni sur lame de Kato: 12 ans d’expérience passés au crible au Kenya

La numération des œufs de schistosomes, malgré ses limites, reste le seul moyen pratique d’estimer l’intensité de l’infection chez l’individu comme dans la communauté, que ce soit à des fins d’enquêtes épidémiologiques ou pour chercher à mesurer l’effet d’interventions de lutte en santé publique ou dans le cadre d’études expérimentales. Pour les infections à Schistosoma mansoni, la méthode d’examen la plus répandue est celle de Kato. On se trouve toutefois régulièrement confronté au problème du maintien des normes de qualité tout au long des études longitudinales de façon à assurer la comparabilité des données, notamment lorsque l’insuffisance des ressources limite le recours à un personnel spécialement chargé du contrôle de la qualité. Cet article décrit un système de contrôle de la qualité élaboré au cours d’études à long terme au Kenya, ainsi que les méthodes d’analyse utilisées pour vérifier que les normes sont maintenues ou que les résultats ne s’écartent pas d’une norme satisfaisante. En résumé, un échantillon aléatoire de 10% de la totalité des lames de Kato examinées a été conservé pour recomptage par un observateur indépendant, qui a vu au moins le cinquième des lames conservées. Les méthodes utilisées, et en particulier le système d’enregistrement des résultats, ont permis de contrôler la qualité du travail des différents microscopistes.

Au total, 19 contrôles annuels ou bisannuels ont été réalisés sur une période de 12 ans et ont porté sur plus de 100 000 lames de Kato sur les quelque 350 000 examinées par trois microscopistes kényens de la Division des maladies transmises par des vecteurs. Les recomptages étaient systématiquement plus faibles que les numérations initiales du fait d’une détérioration des préparations au cours du temps, mais les valeurs obtenues étaient fortement corrélées: une régression significative des deuxièmes numérations par rapport aux premières expliquait 80 à 90% de la variance observée. Le biais lié à l’observateur était sensiblement différent d’un microscopiste à l’autre mais restait stable au cours du temps. La reproductibilité des numérations variait peu en fonction du temps mais baissait légèrement pendant les périodes où la charge de travail était maximale. Environ 7% des recomptages étaient en désaccord avec la numération initiale quant à la présence ou l’absence d’œufs mais dans moins d’un tiers des
cas il s'agissait de "faux négatifs" trouvés positifs au recointage. Ce taux tombait à 1,3% lorsque les numérations étaient faites en double, ce qui est tout à fait comparable aux taux rapportés avec des techniques plus complexes et plus coûteuses. Cette forme de contrôle de la qualité, qui a démontré que la qualité du travail des microscopistes kényens était restée remarquablement élevée, et constante, au cours des 12 années considérées, convient pour des projets où la maigre des ressources empêche d'employer des responsables du contrôle de la qualité à temps plein.

References