Journal of Andrology, Vol. 28, No. 3, May/June 2007

Validation of Sperm Counting Methods Using Limits of Agreement

Andrology Lab Corner*

ELEANOR BAILEY, NICK FENNING, SARAH CHAMBERLAIN, LINDSEY DEVLIN, JAMES HOPKISSON, AND MATHEW TOMLINSON
From the School of Human Development, Fertility Unit, Nottingham University Hospital, Queen's Campus, Nottingham, United Kingdom.

A variety of methods exist for counting sperm. Since the introduction of semen analysis, one of these methods, the hemocytometer, has been regarded as the gold standard by andrology laboratories and the World Health Organization (WHO, 1999). The flexible features of this approach, involving fixation and immobilization of sperm, dilution of highly concentrated samples, and the counting of sperm in a single plane, contribute to the accuracy of the Improved Neubauer hemocytometer and its relative ease of use. This method is strongly accepted within andrology clinics and has been clinically validated by a number of studies (Dunphy et al, 1989; Tomlinson et al, 1996, 1999; Guzick et al, 2001).

As technology and techniques improve, manufacturers are continually trying to develop newer, simpler, quicker and more accurate methods for determining sperm concentration. Busy assisted reproduction technology (ART) laboratories in particular would find a quicker yet comparatively accurate method highly desirable, since sperm counting is an essential part of the semen preparation process. Although modern methods may be faster, since unfixed, undiluted semen is used, some labs find these analyses more difficult to use and believe that the counting of motile sperm may produce erroneous results. Furthermore, the WHO states that the newly introduced methods "are convenient in that they can be used without dilution of the specimen, but that they may lack the accuracy of the hemocytometer technique especially for highly viscous and/or heterogeneous specimens. If such chambers are to be used, their adequate accuracy and precision must be established by comparison with hemocytometers" (WHO, 1999).

In addition, it is now a requirement of laboratory accreditation systems that laboratories provide clinical validation for all methods used (ie, demonstrate that they are "fit for purpose"). Currently, the only sperm counting method with a considerable body of evidence to support and clincally justify its use is the hemocytometer (Mortimer, 1994; WHO, 1999).

When using the hemocytometer, the sperm number is calculated using a fixed volume of semen under the coverslip and counting the sperm in a single plane. A significant association between pregnancy and the sperm concentration measured has consistently been shown for this method (Dunphy et al, 1989; Tomlinson et al, 1996, 1999; Guzick et al, 2001). Thus, 64% of laboratories involved in the analysis of semen use this method routinely (Keel et al, 2000).

However, despite its validity as a method, the use of hemocytometry is thought by many to be inconvenient, in that the hemocytometer must be cleaned and assembled prior to each counting event and it involves the use of dilution techniques that can introduce errors, either due to poor technique or the viscous nature of the semen itself. Mathematical mistakes can occur when applying the correction factor to determine the eventual counts, and the recommended dilution method uses fixatives, such as formal saline, which are often a reason for rejection of the method by embryologists in IVF laboratories.

Comparisons of other counting chambers with the hemocytometer, particularly those marketed as easy-to-use 1-step methods have generally not been favorable. Makler sperm counts have been shown to be generally higher than the corresponding counts obtained with the hemocytometer (Coetzee and Menkveld, 2001; Sukcharoen et al, 1994). Indeed, Ginsburg and Armant (1990) have found the Makler chamber counts to be 62% higher than those obtained with the hemocytometer when using latex beads. Other methods, such as the Leja slide (Gynotec Malden, Nieuw-Vennep, The Netherlands) or the Microcell (Conception Technologies, San Diego, Calif) have been shown to produce significantly lower average sperm counts as compared to the hemocytometer. In particular, marked differences were seen at high

Correspondence to: Dr Mathew Tomlinson, Fertility Unit, East Block B Floor, Queen's Medical Centre, Derby Rd, Nottingham NG7 2UH, United Kingdom (e-mail: mathew.tomlinson@nuh.nhs.uk).

Received for publication November 16, 2006; accepted for publication January 9, 2007.

* *Andrology Lab Corner* welcomes the submission of unsolicited manuscripts, requested reviews, and articles in a debate format. Manuscripts will be reviewed and edited by the Section Editor. All submissions should be sent to the *Journal of Andrology* Editorial Office. Letters to the editor in response to articles as well as suggested topics for future issues are encouraged.

DOI: 10.2164/jandrol.106.002188

concentrations (Tomlinson et al, 2001). A number of potential sources of error have been suggested to cause these discrepancies when using 1-step methods. First, the recommendation that motile sperm are counted, which may mean a single spermatozoon is counted more than once or not counted at all. Second and perhaps more significantly is the phenomenum that affects capillary-loaded chambers, such as the Leja and Microcell slides, which is known as the Segre-Silberberg (SS) effect (Segre and Silberberg, 1961). The SS effect results in high-gradient fluid flow in thin capillary-loaded slides, which results in the sperm suspension being forced transversely towards the walls, causing uneven cell dispersion throughout the chamber. New measures have recently been introduced to compensate for this phenomenon. A correction factor is used, which makes allowances for changes in sample viscosity and appears to improve the performance of there slides in terms of agreement with the hemocytometer (Douglas-Hamilton et al, 2005). Third, many of the studies that have compared the performances of various methods with the gold standard of the hemocytometer are problematic in terms of the choice of statistics. Unfortunately, many of these studies have focused on analysing differences across a range of sperm counts, whereas a proper and more detailed analysis is one that compares each and every individual count and measures the agreement between these 2 parameters. Therefore, in the present study, we employ limits of agreement (LoA) to analyse comparative data. This approach, which is based on graphical techniques and simple calculations, allows comparisons between new measurement techniques (ie, the Leja and Makler chambers) and an established method (ie, the hemocytometer). Thus, we can evaluate whether the methods agree sufficiently for the new method to replace the old method and we can decide if the differences between the 2 methods are sufficiently small for the methods to be used interchangeably (Bland and Altman, 1986).

The objectives of the present study were: 1) to determine whether the sperm counts obtained using the Leja slide and Makler chamber compare favorably with the counts obtained using the hemocytometer; 2) to determine whether the accuracy and reliability of sperm counts obtained using the Leja slide and the Makler chamber are improved by prior fixation and dilution of the specimen; 3) to determine whether the sperm counts obtained using the Leja slide compared more favorably with the counts obtained using the hemocytometer when correction is made for the SS effect; and 4) to determine by the use of LOA whether these methods can be considered to be interchangeable, thereby providing validation for their routine clinical application.

Materials and Methods

Sample Collection—In all, 163 semen samples were obtained from patients who attended the NHS Fertility Clinic and Nurture Assisted Conception Unit at the University Hospital, Nottingham, United Kingdom. Patients were referred to the laboratory either by their local GP after a period of suspected infertility or by the Fertility and Nurture Clinics themselves. Samples were produced by masturbation and ejaculation into a sterile, wide-mouthed, nontoxic 60-mL plastic container (Alpha Laboratories, Eastleigh, United Kingdom).

On arrival at the laboratory, the samples were allowed to liquefy for up to 1 hour at room temperature. The samples were then well mixed in their original container prior to analysis. Specimen acceptance criteria were defined as follows: the sample had to be less than 1 hour old, complete, and produced in the recommended container after a period of abstinence of between 2 and 5 days.

Sperm Concentration Measurements Using the Hemacytometer—The double Improved Neubauer hemocytometer used for analysis was prepared for counting according to the WHO guidelines. The coverslip was applied by horizontal sliding. Correct positioning of the coverslip was confirmed by the presence of Newtons rings on both sides of the chamber (WHO, 1999). Before loading, the sample was diluted in 3% formal saline. Throughout the procedure, positive displacement pipettes were used for the transfer of semen (Gilson Microman; Anachem, Luton, United Kingdom). For most semen samples, a standard dilution of 1:20 was suitable. The dilution was adjusted appropriately according to the estimated concentration, to allow the enumeration of a mimium of 200 sperm. Each side of the counting chamber was then loaded until it was full (8–10 μL). Care was taken to ensure that the chamber was not underfilled and that the sample was not allowed to spill into the central trough. The hemocytometer was left for 5–10 minutes in a humid chamber, to allow the cells to sediment before counting and to prevent drying out.

Sperm Counting Using the Leja Slide—Duplicate counts using the same 163 specimen samples were made on the Leja slide according to the manufacturers instructions. Initially, the microscope was calibrated using an eyepiece reticule and stage micrometer, in order to obtain the objective correction factor. Positive displacement pipettes were used to fill the Leja chamber and filling time was recorded in seconds immediately after dispensing the semen. The number of sperm counted within the 10 × 10 grid was noted, with a minimum of 200 spermatozoa counted for each evaluation. The concentration was calculated according to the manufacturers' instructions and the formula (average

number of sperm per small square × microscope factor × F), where F is the factor compensating for the SS effect. The counts were repeated using specimens fixed and diluted in 1:1 formal saline (3.5%).

Sperm Counting Using the Makler Chamber—The Makler chamber was used according to the manufacturers instruction. Well-mixed semen (5 μL) was transferred to the Makler chamber. The cover was applied promptly, since a delay has been shown to be a potential source of error that results in a higher sperm concentration (Matson et al, 1999). When applying the cover, care was taken to avoid the formation of bubbles. Once again, the counts were repeated using specimens fixed and diluted in 1:1 formal saline.

Sperm Motility—Each sample received for sperm concentration analysis was also assessed for sperm motility. The scepticism surrounding counts obtained from the Leja slide and Makler chamber is thought to be associated with the use of motile sperm, yet there are very few publications on this topic. Therefore, by determining the motility of each sample in an accurate manner, it is possible to elucidate whether increased motility leads to greater discrepancies or inaccuracies.

The motility of each sample was assessed in the Improved Neubauer hemocytometer using the WHO classification system for a minimum of 200 spermatozoa (WHO, 1999). For consistency, all readings were carried out at 37°C using a heated microscope stage (Linkham Scientific, Tadworth, United Kingdom).

Reliability/Repeatability of the 3 Chambers—It is essential to show that a single observer can obtain the same results when repeated measurements are made using the same method under identical circumstances (Petrie and Sabin, 2005). Thus, any detectable variation in counts between chambers can be attributed to the inherent characteristics of the chambers rather than poor precision.

Ten of the 163 samples obtained from the clinics were used in a repeatability investigation. Each sample was thoroughly vortexed and aliquots were counted 5 times in the hemocytometer, 5 times on the Leja slide and then 5 times in the Makler chamber. This process was repeated after the sample was diluted 1:1 in diluent, with the counts being repeated on the Makler and Leja slides. This gave a total of 50 counts for each method.

Data Analysis and Statistics—Sperm counts obtained using the different methods were compared using LOA according to the methods of Bland and Altman (1986). In this analysis, the differences between the two methods being compared (ie, the Leja vs hemocytometer, and Makler vs hemocytometer) are plotted against the mean of the 2 methods, since the true value is not known and this is the best estimate of this true figure (ie, the real sperm concentration). It would be a mistake to plot the difference between each value separately, as the differences will be related to each, which generated a well-known statistical artifact (Gill et al, 1985).

The limits within which 95% of the differences are expected to lie are also plotted. These limits are estimated from the mean and standard deviation of the differences, and are plotted as horizontal lines 2 standard deviations either side of the mean. These limits, like the confidence intervals, give an idea of the spread of variation between the methods.

This method was used to compare the Leja undiluted, Leja 1:1, Makler undiluted, and Makler 1:1 counts to the corresponding counts obtained using the Improved Neubauer hemocytometer. The uncorrected SS counts were then compared to the hemocytometer, to investigate whether applying the advised SS correction factor increased the association between the Leja slide and hemocytometer.

As the data exhibited a slightly skewed distribution, they were normalized using a square-root transformation. The nonparametric raw data for the 3 chambers were tested for significance using the Wilcoxon Signed Rank test. Each method was compared in turn with the hemocytometer to evaluate any significant differences.

Statistical Analysis of Repeatability—The repeat results were analysed using SPSS 13.0 for Windows (SPSS Inc, Chicago, Ill). The Leja undiluted, Leja 1:1, Makler undiluted, Makler 1:1, and hemocytometer repeat counts were tested systematically for absolute agreement using Intraclass Correlation Coefficients (ICC) and the 2-way mixed effects model. This model was chosen as it is appropriate to situations in which people effects are random and measures effects are fixed (ie, the sample concentrations were random and the measures in the three different chambers should all be the same). The repeat counts for each sample should be identical, so that the closer the ICC value is to 1, the more reliable is the counting method.

Coefficients of Variation (CV) were also calculated, to determine the spread of the repeat counts relative to the mean for each method. The lower the CV, the smaller the variation between the repeat counts, and thus the higher the repeatability. The 95% confidence intervals (CI) were also measured for each method, and plotted in a chart to visualise the spread and thus, the variability of the repeated counts.

Results

Repeatability/Reliability of the Chambers—Intraclass correlation coefficients and 95% CI showed that the hemocytometer gave the highest degree of repeatability (intraclass correlation coefficient of 0.981), followed closely by the Leja slide (0.98) and the Makler chamber

Table 1. *Intraclass correlation coefficients and 95% confidence intervals for the different sperm counting methods*

Counting Method	Intraclass Correlation	95% Confidence Interval	
		Lower Bound	Upper Bound
Improved Neubauer hemocytometer	0.981	0.953	0.994
Leja slide (undiluted, unfixed)	0.980	0.952	0.994
Leja slide (1:1, fixed)	0.968	0.925	0.991
Makler chamber (1:1, fixed)	0.930	0.842	0.979
Makler chamber (undiluted, unfixed)	0.920	0.82	0.976

(0.92). This finding was confirmed by the CV of the repeat data, with the hemocytometer demonstrating the lowest CV and the Makler the highest CV. Fixing and diluting the sample increased the CV values for both the Makler and Leja slides. Therefore, the hemocytometer had the narrowest CI values, with the Makler chamber having the widest CI, it being greatest for the diluted method (Tables 1 and 2).

Comparisons of the Hemocytometer, Leja, and Makler Counts—The 163 semen samples measured using the three different chambers are displayed graphically. Since the hemocytometer is accepted as the gold standard, the counts have been ranked in order of concentration determined by method. Figure 1 shows a clear correlation between the three methods but also shows that the Makler and Leja counts fluctuated about the concentration obtained using the Improved Neubauer hemocytometer. The Leja slide produced counts that were on average 17% lower than those obtained using the hemocytometer, while the Makler counts were on average 32% higher. The differences between the three chambers appeared to increase as the sperm concentration increases, particularly for the Makler chamber, although in general, the same pattern was observed for sperm counts at the oligozoospermic end of the range and those in the normal range.

Significance—When the samples were diluted 1:1, the hemocytometer counts were significantly different for every chamber, with the exception of the Leja slide (Table 3). The median Leja (undiluted) count was significantly lower than that of the hemocytometer ($P < .05$), while the median Makler value was higher ($P < .0001$). The introduction of dilutions to both methods improved their agreement with the gold standard, to the point that the Leja counts were no longer significantly different from those of the hemocytometer. This test illustrates an overall significant difference between the methods tested, although a more complex statistical method is required to analyze individual differences.

Limits of Agreement Analysis—The Leja slide. The mean difference between the Leja and hemocytometer was −0.2. This denotes that a sample determined to have a concentration of 5×10^6 sperm/mL on the Improved Neubauer would be underestimated as containing an average of 4.1×10^6 sperm/mL on the Leja. A sample with 20×10^6 sperm/mL would be determined to contain an average of 18.2×10^6 sperm/mL using this method (Figure 2). The standard deviation was calculated as 0.9, so the limits of agreement lie between -2.04 and 1.7. This means that in a worst case scenario, specimens that contain 5×10^6/mL on the hemocytometer should have counts that lie between zero and 16×10^6 sperm/mL. In the same way, the Leja slide would determine the concentration of a sample with 20×10^6 sperm/mL to lie between 6×10^6 and 38×10^6 sperm/mL.

Table 2. *Comparing the Coefficients of Variation (CV) for repeated counts made using the different counting methods*

Counting Method	Coefficient of Variation (%)
Hemocytometer (n = 50)	8.5
Leja 1:1 (n = 50)	9.8
Leja undiluted (n = 50)	10.2
Makler undiluted (n = 50)	13.1
Makler 1:1 (n = 50)	17.9

From the plot, it is evident that 16/163 (9.8%) counts are outliers of these limits. Thus, almost twice as many of the Leja counts strongly disagree with the hemocytometer counts than would be statistically expected. However, there is no obvious relationship between the difference noted between the 2 methods and their means. Therefore, increasing the sperm concentration of the semen does not affect the accuracy of the Leja method.

Dilution of the semen with 3% formal saline followed by analysis with the Leja method produced agreement between the 2 methods (Figure 3). The Leja 1:1 method and hemocytometer had a mean difference of 0, which suggests that on average, their counts are in total agreement, provided that the SS factor is applied during their use.

The limits of agreement were calculated to be between −2.06 and 2.1. These are slightly wider than for the analysis of undiluted samples, so the Leja 1:1 method would determine a sample of 5×10^6 sperm/mL as being between zero and 19×10^6/mL, and a sample of 20×10^6/mL as being between 6×10^6 and 43×10^6 sperm/mL in a worse-case scenario. There are still however 10 outliers which lie outside the limits of agreement.

The Makler chamber. From the plot in Figure 4, it can be seen that the mean difference is 0.9. The limits of agreement are between −1.42 and 3.18. These are considerably wider limits than for the Leja method, so in a worst case scenario, a sample of 5×10^6/mL would

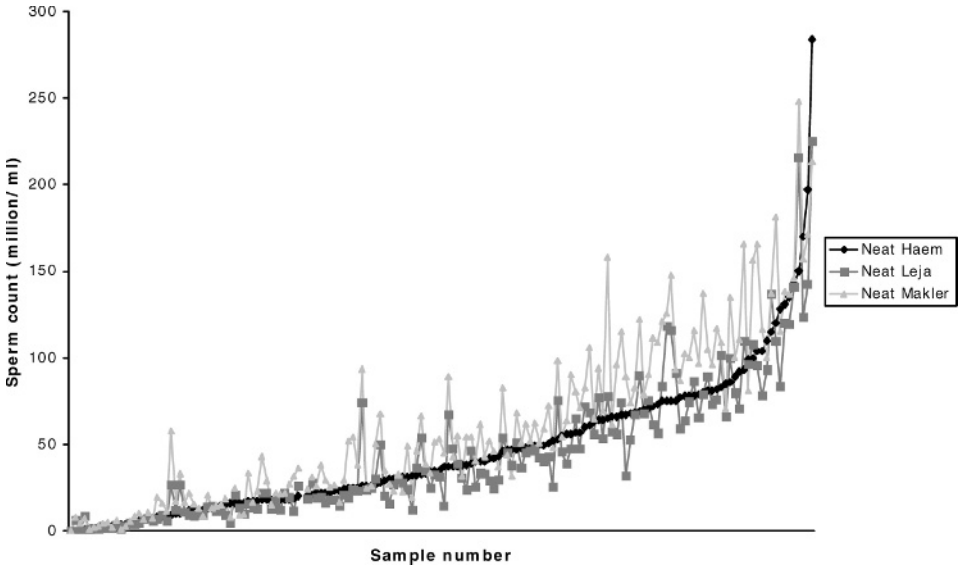

Figure 1. Graph showing sperm counts obtained on the Improved Neubauer hemocytometer, Leja, and Makler chambers using undiluted semen samples (n = 163).

be determined as being between 0.4×10^6 and 30×10^6 sperm/mL and a sample of 20×10^6/mL as being between 9.2×10^6 and 58.9×10^6 sperm/mL. Average concentrations of 5×10^6 and 20×10^6 sperm/mL were determined by the Makler to be 10×10^6 and 29×10^6 sperm/mL, respectively. There were 7 outliers of these limits, all of them above the upper limit, which further demonstrates the tendencies of the Makler method to overestimate sperm counts.

However, it can be seen by the trendline in Figure 5 that there is a change in bias as the subject mean (ie, semen concentration, increases). This suggests that the Makler overestimates to a greater degree as the semen concentration increases. Therefore, LOA analyses were repeated for the samples once they were subdivided into low (lower one third), medium (middle one third), and high (upper one third) counts. The hemocytometer and Makler showed the closest agreement (when the mean difference was 0) at very low sperm concentrations.

Table 3. *Median sperm counts determined by each counting method and the degree of significance between each method and the hemocytometer using the Wilcoxon Signed Rank test*

	N	Median Sperm Count ($\times 10^6$/mL)
Hemocytometer	163	37.0
Leja undiluted	163	30.6*
Makler undiluted	163	49.0**
Leja 1:1	163	34.8
Makler 1:1	163	42.0**

* Significantly different at $P < .05$ from the hemocytometer.
** Significantly different at $P < .0001$ from the hemocytometer.

The mean differences obtained with the hemocytometer for the lower, middle, and upper groups were 0.6, 0.8, and 1.3, respectively. This demonstrates clearly that the Makler overestimates to a greater degree as the sperm concentration increases.

As the semen concentration increased, the difference between the upper and lower limits of agreement increased as follows: lower concentration, −1.3 to 2.5; middle concentration, −1.5 to 3.1; upper concentration, −1.1 to 3.7.

Fixing and diluting semen on the Makler chamber. The Makler 1:1 method showed a mean difference of 0.7, which is a smaller difference than that obtained using undiluted semen. Transforming the data demonstrates that for diluted semen, the Makler overestimates semen samples of 5×10^6/mL as being 8.8×10^6/mL, and 20×10^6/mL as being 26.8×10^6/L on average, which is a small improvement in terms of agreement as compared to using undiluted semen (Figure 5).

Effect of counting method on patient diagnosis. From Table 4 it can be seen that using different counting methods has the potential to affect the overall diagnosis. In 28 men out of 163 (17%) in this population, diagnosis was clearly dependent upon the counting method used.

Using the Leja slide, 20 of the men would be diagnosed as oligospermic and possibly offered fertility treatment, while only 15 would be diagnosed as oligospermic based on their hemocytometer counts. In contrast, only 1/28 men was found to be oligospermic using the Makler slide because of its tendency to overestimate sperm counts. Diluting the specimens and using the Leja slide slightly improved the situation,

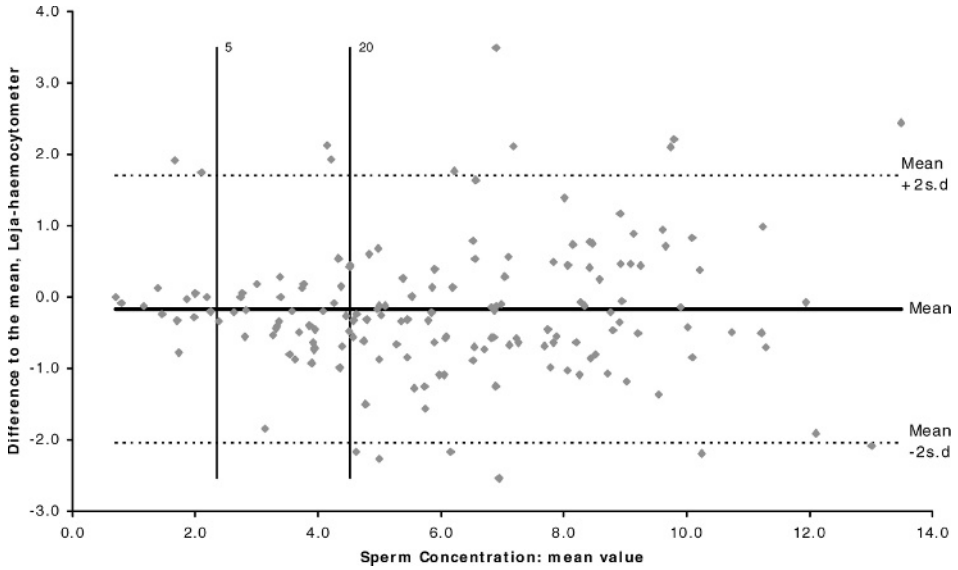

Figure 2. LOA plot for the Leja slide using undiluted semen vs the Improved Neubauer hemocytometer (n = 163). The bold horizontal line represents the mean difference, the horizontal dotted lines represent the LOA, 2 standard deviations either side of the mean, and the vertical lines illustrate the cutoffs at 5 × 10^6 and 20 × 10^6 sperm/mL for severe oligospermia and oligospermia diagnoses, respectively.

reducing the number diagnosed as oligospermic to 17, while using the Makler method this number increased to 7 men.

Discussion

The aim of the present investigation was to determine whether three different methods used for analyzing sperm counts in a fertility laboratory setting could be considered to be interchangable. All 3 methods are currently in use in laboratories throughout the United Kingdom and worlwide.

The ICC, CV, and CI values show that all the counting methods have a particularly high degree of repeatability/reliability. The hemocytometer was the most reliable method, with the highest ICC, lowest CV, and smallest 95% CI. This finding partially supports the hemocytometer as the gold standard method, although its precision is clearly not high. The Leja was the next most reliable

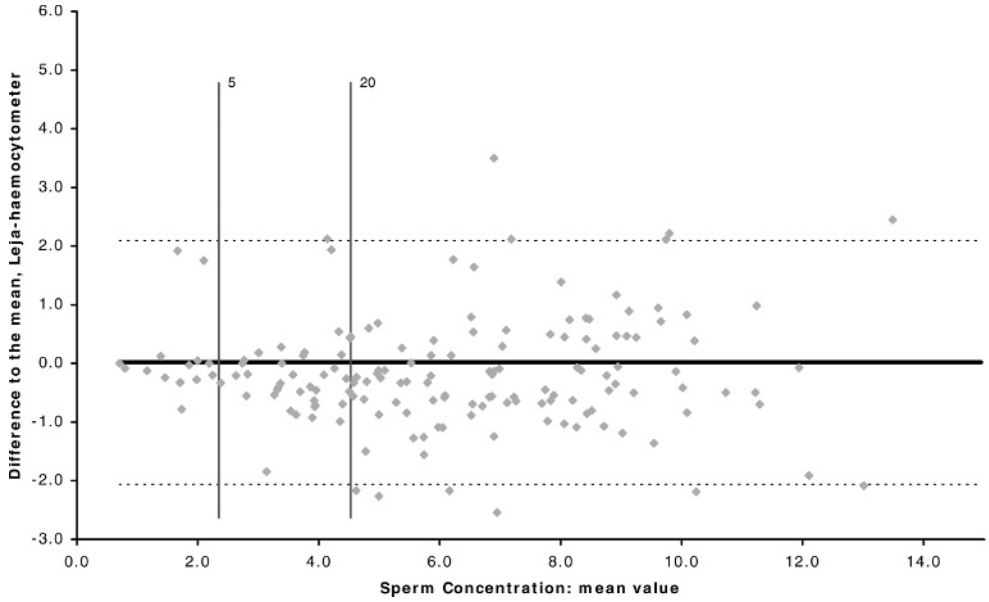

Figure 3. LOA plot for Leja 1:1 compared to the hemocytometer (n = 163). The bold horizontal line represents the mean difference, the horizontal dotted lines represent the LOA, and the vertical lines illustrate the cutoffs at 5 × 10^6 and 20 × 10^6 sperm/mL for severe oligospermia and oligospermia diagnoses, respectively.

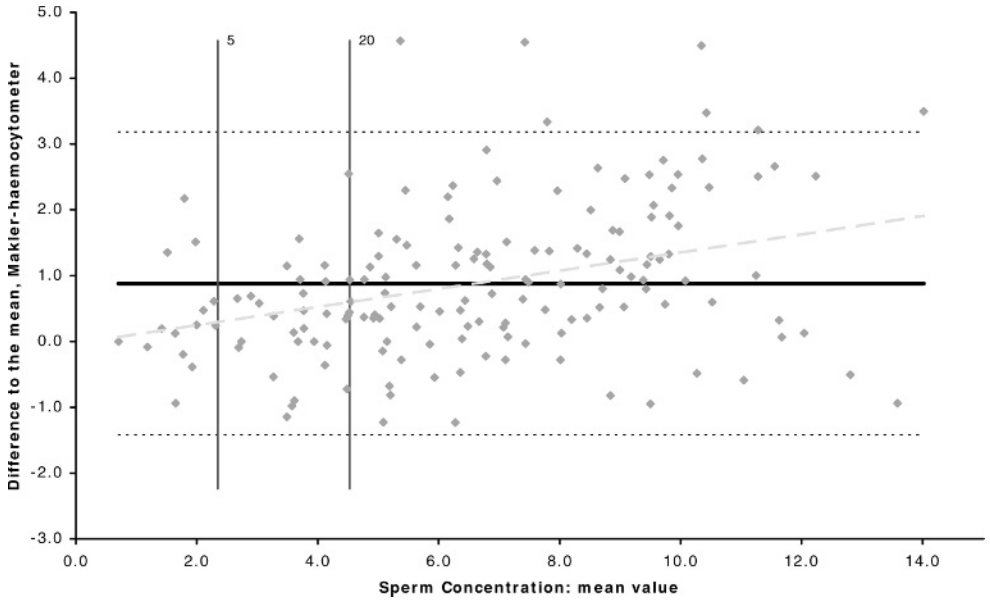

Figure 4. LOA plot comparing the Makler chamber and Improved Neubauer hemocytometer (n = 163). The bold horizontal line represents the mean difference, the horizontal dotted lines represent the LOA, and the vertical lines illustrate the cutoffs at 5×10^6 and 20×10^6 sperm/mL for severe oligospermia and oligospermia diagnoses, respectively. The dashed line is a trendline that shows the change in bias with increasing mean.

method, although it was more reliable without the fixation/dilution step, which suggests that this step may introduce a degree of error. The analyses showed the Makler chamber to have significantly poorer reliability than the hemocytometer or Leja. The ICC of the Makler chamber was much lower for both undiluted and diluted samples, and the CV and CI showed great variability between the repeated counts.

The LOA analysis showed that the Leja slide could be compared favourably with the hemocytometer, particularly for the 1:1 dilutions. However, despite showing that the Leja 1:1 and hemocytometer have a mean

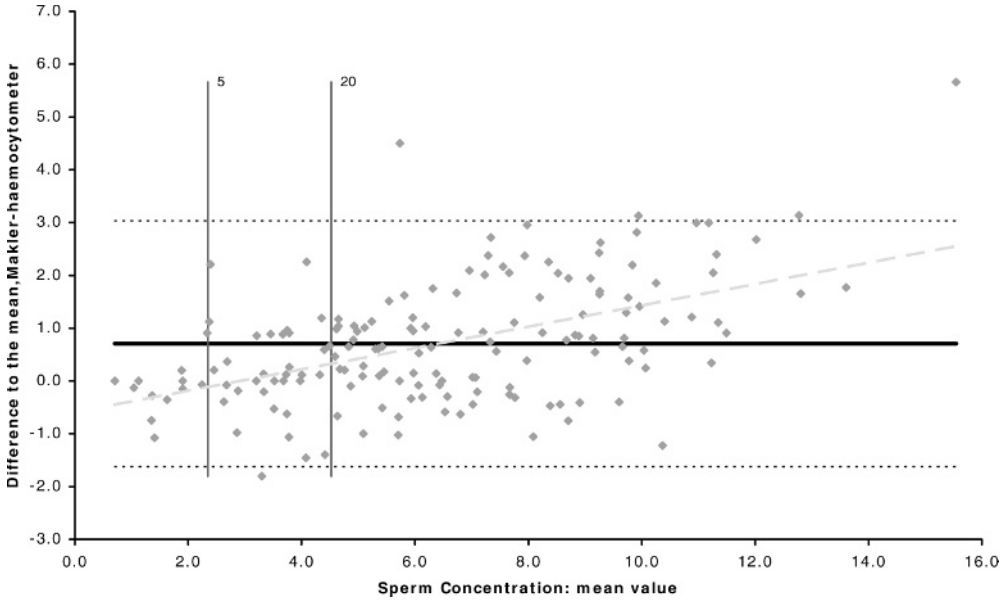

Figure 5. LOA plot for Makler 1:1 compared to the hemocytometer (n = 163). The bold horizontal line represents the mean difference, the horizontal dotted lines represent the limits of agreement, and the vertical lines illustrate the cutoffs at 5×10^6 and 20×10^6 sperm/mL for severe oligospermia and oligospermia diagnoses, respectively. The dashed trendline illustrates the change in bias with increasing mean.

Table 4. *Differential oligospermia diagnoses of 28 men who had their semen samples counted in different chambers*

Lab No.	Sperm count ($\times 10^6$ sperm/mL) Hemocytometer	Leja Undiluted	Makler Undiluted	Leja 1:1 Dilution	Makler 1:1 Dilution
4030	**9**	26.6	58	26.1	57
2350	**10**	26.3	33	20.9	**19**
2186	**12**	**9.07**	21.5	**10.56**	33
2781	**13**	**14**	20.5	**12.5**	**8.8**
2514	**13**	**11.6**	**14.5**	**9.0**	21
3069	**16**	20.7	24.5	**18.6**	25
4810	**17**	**15.4**	33.5	**12.6**	23
2761	**17.5**	**12.50**	23	**12.71**	29
3992	**18**	**12.2**	21	**9.4**	**10**
2501	**18**	**17.3**	22	**14.9**	**19**
2473	**18**	22.0	29	**18.7**	20
2611	**18**	**19.3**	27	26.7	24
2383	**18**	21.9	43	**15.8**	27
2684	**18**	22.1	21.5	**17**	28
4871	**18.5**	**11.3**	31.5	**12.5**	24
1679	20.5	**18.2**	24	**15.9**	25
2438	22	**17.7**	25.5	22.0	24
2263	22	**19.0**	26	21.9	30
2771	22	**15.9**	29.5	20.5	33
2377	22	**19.8**	38	23.2	34
2579	23	**14.4**	16.5	**16.7**	**11**
2279	23	**17.9**	26.5	21.1	**17**
2224	25	**19.21**	52	24.06	26
2447	26	23.5	24.5	**17.4**	25
2277	30	**15.7**	27	24.2	31
2308	32	**12**	19.5	**15.9**	**18**
2359	37	**14.4**	45	22.3	31
2507	40	33.0	43	**19.1**	28

Bold numbers indicate subjects found to be oligospermic (fewer than 20 $\times$ 10^6/mL), while light typeface numbers are subjects determined to be normospermic (more than 20 $\times$ 10^6/mL).

difference of zero, the standard deviation was calculated as 0.9. Thus, the LOA lie between -2.04 and 1.7, which means that in a worst case scenario, specimens that contain 5 $\times$ 10^6/mL in the hemocytometer could have counts between zero and 16 $\times$ 10^6 sperm/mL. Similarly, the Leja method could determine the concentration of a sample with 20 $\times$ 10^6 sperm/mL to be between 6 $\times$ 10^6 and 38 $\times$ 10^6 sperm/mL. Although the latter form of analysis may at first glance appear to lack precision, the same principle can be applied to repeat counts using the same chamber. In other words, in most cases, the difference between the chambers is no worse than repeated measures using the same chamber. There will always be occasional outliers and perhaps more outliers when semen is the test fluid in question and this could be due to any one of a number of factors (eg, hyperviscosity, sampling error, mathematical error, diluting error or chamber flaws). When assessing the agreement and interchangeability of a method, it has to be put in the context of the level of expectation. As there is innate variability in hemocytometer counts using either repeated measures or between individuals, we will undoubtedly detect differences when comparing this with other chambers. In this context, the Leja slide could be viewed as being interchangeable with the hemocytometer, particularly if the sample is first fixed and diluted.

All of the results show very clearly that applying the SS factor increases the level of agreement between the Leja counts and hemocytometer counts, as expected. This improvement was more pronounced for the 1:1 diluted samples, for which the mean difference with the hemocytometer became 0 for the modified Leja counts. This is reassuring, since the Leja Instruction Manual states that the correction factors were initially calculated by calibrating the Leja with the Improved Neubauer chamber.

The Makler chamber showed extremely poor agreement with all the other methods and had a tendency to overestimate sperm counts. In the worst case, a sample of 20 $\times$ 10^6/mL could be found have between 9.2 $\times$ 10^6 and 58.9 $\times$ 10^6 sperm/mL, which could have serious consequences for the diagnostic or treatment laboratory. This finding supports previous studies that have found that the Makler produces high sperm counts (Ginsburg and Armant, 1990; Sukcharoen et al, 1994; Seaman et al, 1996; Coetzee and Menkveld, 2001; Lu et al, 2004). Laboratories provided with a latex bead solution with a known, fixed concentration of 35 $\times$ 10^6 sperm/mL determined the concentration on the Makler to be 53.5 $\times$ 10^6/mL on average. A similar study conducted by Seaman et al (1996) found that the Makler chamber overestimated the known bead concentrations by as much as 50%.

The counts determined using the Leja slide were significantly different from the counts obtained using the hemocytometer ($P < .05$), unless they specimens were first fixed and diluted. The Leja median was significantly lower than that of the hemocytometer, and this finding was supported by the LOA analysis, which showed that the Leja on average underestimated the sperm count when undiluted semen was used. This underestimation was consistent throughout the data distribution and was not related to the sperm concentrations of the samples. It was expected that the Leja would underestimate the sperm count when compared to the hemocytometer, since a study by Tomlinson et al (2001) has found Leja counts to be significantly lower ($P < .0001$) than the corresponding hemocytometer readings, although these counts were performed using a previous version of the Leja chamber and without knowledge of the SS factor.

Fixing and diluting the samples for analysis by the Makler method reduced the mean difference and increased the level of agreement with the hemocytometer counts. However the counts produced in the Makler

chamber were significantly higher than those in the hemocytometer, regardless of whether they were performed on undiluted or diluted specimens.

In some instances, laboratories are clearly willing to compromise on the accuracy of sperm parameter measurements in order to maximise speed and convenience. The preference for methods of sperm counting and the selection of counting chambers may in many cases reflect the type of clinical service and the level of expertise provided by a particular laboratory. It is known that many IVF labs favor the 1-step methods owing to their convenience, even though this could indirectly compromise practice, in that future treatment decisions based on sperm thresholds could be derived from erroneous data.

It has been suggested previously that using different methodologies to analyze sperm concentrations is the major cause of variation in sperm counts between laboratories (Auger et al, 2000). Owing to disagreements between laboratories, a patient could be classified as normal by one and as infertile by another (Neuwinger et al, 1990). This has been known for some time and is supported by the findings of the present study. Sperm concentration discrepancies between the Leja and hemocytometer affected the diagnosis of 21 patients in the present study. Such erroneous diagnoses could have serious consequences for patients, some of whom may be misidentified as infertile (according to the WHO criteria), which could in turn have negative psychosocial implications and mean that some individuals would not receive the appropriate assistance in achieving pregnancy.

We know from previous studies that many of the current methods in use carry a degree of error, which is high in comparison with many other types of diagnostic testing, many of which are fully automated. Simple factors, such as the inherent viscosity and heterogeneity of semen, incomplete mixing, errors in pipetting, mathematics, and transcribing, are all contributory. It is likely that the general lack of consensus and lack of strong evidence linking sperm concentration with either natural or assisted conception are due to either poor practice or the intrinsic inaccuracies of many of the methods used.

Another source of error, and possibly the most significant, is sampling error. By its very nature, 1 small microlitre aliquot of semen may have very different qualities when compared to another. Therefore, only by better mixing, increasing the volume of seminal fluid, and increasing the number of sperm counted per analysis can we hope to improve the current degrees of accuracy and precision. However, this has to balanced against whether there is a need for such a degree of accuracy and also whether it is a good use of resources.

The relatively high risk of error during the analysis, be it due to the method used, sampling error, transcription or a case of mistaken identity, raises the question as to whether a diagnosis should ever be based upon the findings of a single semen sample. Adding in other variables, such as whether the sample is complete, length of sexual abstinence, illness, stress, and medications, all of which are known to affect sperm count, ensures that the relationship between male fertility and sperm count is weaker than perhaps it should be. The unfortunate consequence of this is that semen analysis continues to demonstrate poor predictive value in terms of either natural or assisted conception (eg, IUI, IVF, and ICSI).

The hemocytometer method has long been the accepted standard for the assessment of sperm concentration. Indeed, the WHO recommends in its guide to semen analysis (1999) that a hemocytometer, such as the Improved Neubauer, should be used for sperm concentration measurements. Despite this, an estimated 64% of laboratories used the hemocytometer, while 26% used the Makler chamber, and the remainder used alternative disposable methods, such as the Leja slide (Keel et al, 2000). One-step disposable methods, such as the Leja slide, and re-usable glass chambers, such as the Makler, have been put forward as convenient alternatives. The present study suggests that there is limited agreement between all of these methods, with the widest discrepancy between hemocytometry and the Makler method.

If the standard Leja method was modified by fixing and diluting the sample, the overall agreement with hemocytometery was very good. Therefore, if it can be shown that the 2 methods are likely to give the same counts, the 2 methods could be said to be interchangble, providing a degree of validation for the Leja method. The hemocytometer is marginally better in terms of precison but, like all of the methods, has intrinsic weaknesses in terms of reliability and repeatability, which may be related to the test fluid in question.

In conclusion, there is little to choose between the hemocytometer and diluted Leja method in terms of average sperm count, which suggests that the Leja slide may be suitable for use in clinical pactice. In contrast, the Makler method showed very poor agreement throughout the entire data range and cannot be recommended. The limited agreement seen between all 3 methods and their apparent lack of precision is a cause for concern. If the WHO continues to recommend the hemocytometer as the gold standard method, then there should at least be a reappraisal of the method. Testing inaccuracies are often blamed on poor technique or technical training, yet it is clear that all methods for sperm counting have inherent weaknesses. Laboratories must be better equipped in order to provide users of their service (eg, Ob/Gyn clinicians). with proper information about the test, including its limitations and the expected level of accuracy and precision, in

order to allow efficient management of their male patients.

References

Bland MJ, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1: 307–310.

Coetzee K, Menkveld R. Validation of a new disposable counting chamber. *Arch Androl*. 2001;47:153–156.

Douglas-Hamilton DH, Smith N, Kuster C, Vermeiden J, Althouse GC. Particle Distribution in low-volume capillary-loaded chambers. *J Androl*. 2005;26:107–114.

Dunphy BC, Neal LM, Cooke ID. The clinical value of conventional semen analysis. *Fertil Steril*. 1989;51:324–329.

Gill JS, Zezulka AV, Beevers DG, Davies P. Relationship between initial blood pressure and its fall with treatment. *Lancet*. 1985;1: 567–569.

Ginsburg KA, Armant DR. The influence of chamber characteristics on the reliability of sperm concentration and movement measurements obtained by manual and videomicrographic analysis. *Fertil Steril*. 1990;5:882–887.

Guzick DS, Overstreet JW, Factor-Litvak P, Brazil CK, Nakajima ST, Coutifaris C, Carson SA, Cisnero P, Steinkampf MP, Hill JA, Xu D, Vogel DL. National Cooperative Reproductive Medicine Network. Sperm morphology, motility and concentration in fertile and infertile men. *New Eng J Med*. 2001;345:1388–1393.

Keel BA, Quinn P, Schmidt CF, Serafy NT Jr, Serafy NT Sr, Schalue TK. Results of the American association of bioanalysts national proficiency testing programme in andrology. *Hum Reprod*. 2000;15: 680–686.

Lu J, Lu N, Huang Y, Li PS, Fisch H. Quality evaluation of three different sperm counting chambers (Translated from Chinese). *Zhong Hua Nan Ke Xue*. 2004;10:755–757.

Matson P, Irving J, Zuvela E, Hughes R. Delay in the application of the cover glass is a potential source of error with the Makler Counting Chamber. *Fertil Steril*. 1999;72:559–561.

Mortimer D. *Practical Laboratory Andrology*. Oxford United Kingdom: Oxford University Press; 1994.

Neuwinger J, Behre HM, Nieschlag E. External quality control in the andrology laboratory: an experimental multicentre trial. *Fertil Steril*. 1990;54:308–314.

Seaman EK, Goluboff E, BarChama N, Fisch H. Accuracy of semen counting chambers as determined by the use of latex beads. *Fertil Steril*. 1996;66:662–665.

Segre G, Silberberg A. Behaviour of macroscopic rigid spheres in Poiseuille flow. *J Fluid Mech*. 1961;14:115–135.

Sukcharoen N, Ngeamjirawat J, Chanprasit Y, Aribarg A. A comparison of Makler counting chamber and improved Neubauer hemocytometer in sperm concentration measurement. *J Medical Ass Thai*. 1994;77:471–476.

Tomlinson M, Turner J, Powell G, Sakkas D. One-step disposable chambers for sperm concentration and motility assessment: how do they compare with the World Health Organisation's recommended methods? *Hum Reprod*. 2001;16:101–104.

Tomlinson MJ, Kessopoulou E, Barratt CLR. The diagnostic and prognostic value of traditional semen parameters. *J Androl*. 1999;20:588–593.

Tomlinson MJ, Amissah-Arthur JB, Thompson KA, Kasraie JL, Bentick B. Prognostic indicators for intrauterine insemination (IUI): statistical model for IUI success. *Hum Reprod*. 1996;11: 1892–1896.

World Health Organization. *WHO Laboratory Manual for the Examination of Human Semen and Sperm-Cervical Mucus Interactions*. 3rd ed. Cambridge, United Kingdom: Cambridge University Press; 1999.