

WSLH PT Scoring and Evaluation

Descriptions of the scoring criteria and evaluation processes utilized by WSLH PT for CLIA '88 regulated qualitative and quantitative analytes follow. Waived methods are evaluated following similar processes. For unregulated analytes, scoring criteria is established using knowledge of customer data, rules generally accepted by the laboratory community (e.g. Westgard multi-rule) and/or other statistically relevant processes (e.g. Grubbs).

Overview: Participants receive PT samples, test them and return results to WSLH PT using the result forms provided or via online entry. After a series of check-in and verification steps, the results are entered into the scoring database. The results are evaluated using predefined criteria. Scores are assigned to an analyte/sample/procedure. A score is a numeric value reported as a percent (zero to 100%) for a specific event. If the event score satisfies the criteria, it is termed “satisfactory” or “unsatisfactory” if it doesn’t satisfy the criteria. An evaluation report is sent to each participant and any designated consultant.

Scoring of the results is accomplished by applying the defined processes from two major categories: enumerated and quantitative. Each of these major categories has subtypes depending on the method used to determine the target value (referee or peer group).

Referee laboratories are PT participants whose scores over the previous three events are satisfactory (for most analytes ≥ 80 percent). Candidate referee laboratories are selected by the database and presented to the PT coordinator for final approval. Care is taken to select a slate of referees (minimum of ten) that represents a cross-section of the participants reporting a particular analyte.

A **peer group** is a group of participants using any of the following: a specific test system, a specific analytical principle, a composite test system, or all participants (inclusive or “All Methods” group). A peer group is not used in the scoring process until consensus testing validates it.

A **scoring group** is the peer group/referee laboratory result group used to calculate the acceptable result/target value and acceptable range for scoring purposes.

Calculating the percent **consensus (agreement) of grouped results** and comparing it to the specified consensus criteria tests the validity of grouping the results for referee/peer group scoring. Currently, CLIA regulations use the concept that 80% (for all analytes except Immunohematology which is scored using 95% criteria) of the results must match the accepted result or fall within the acceptable range in order for the scoring to be considered valid.

Scoring exceptions are handled as follows:

- Excuse requested status (requested prior to due date) – receives a score of 100% per result, analyte or sample.
- Late status (results received after the due date) – receives a score of 0% per result, analyte or sample.
- Not scored status – receives a score of 100%. Assigned only if N <10 and data does not fit the “All Methods” group or there is a documented sample quality problem.
- Non-consensus status – receives a score of 100%.
- Missing results for a regulatory analyte (one or more of five results left blank) – receives a score of 0% per result.
- Missing results for an unregulated or waived method or analyte may be listed as “not reported” and may not receive a score.
- Less than or greater than – if the number following the less than or greater than sign falls within the acceptable range, the result receives a score of 100% and if the number falls outside of the acceptable range, the result receives a score of 0%.
- Peer groups with less than 10 participants – promote to the next highest level within the method (e.g. all Beckman instruments) or “All Methods” group if data falls within the acceptable ranges and consensus is maintained.

There are six different scoring processes utilized by WSLH PT; each is explained below.

The **quantitative scoring process** is used when the results are numeric.

- A. Quantitative scoring by referee laboratory results (e.g. Blood Lead).
- B. Quantitative scoring by peer group results (e.g. Chemistry).

The **enumerated scoring process** is used when the results are non-numeric and require translation interaction with glossaries/tables.

- C. Enumerated scoring by referee laboratory results - General (e.g. Hematology Cell ID).
- D. Enumerated scoring by referee laboratory results - Microbiology (ex: Urine Culture).
- E. Enumerated scoring by peer group results (ex: Immunoserology).
- F. Enumerated scoring for antibiotic susceptibility.

A - Quantitative Scoring by Referee

This scoring type uses referee labs with acceptable performance over the past three events. The **referee target** (mean) and **acceptable range determination process** is as follows:

1. Calculate the arithmetic mean of the referee laboratories for each sample.
2. Calculate the acceptable range for each sample by applying a specified tolerance factor (stored in a tolerance table) to the referee mean.
3. Calculate the consensus of the referee labs for each sample by counting the number of results that fall within the acceptable range and dividing by the total number of referee labs.
4. If consensus is $\geq 80\%$, the mean and acceptable range is valid for scoring.
5. If consensus is $< 80\%$, all participants receive a score of 100% for the analyte.
6. Score all participant results for each sample based on the referee acceptable range. If the result falls within the acceptable range, the result receives a score of 100%. If the result falls outside of the acceptable range, the result receives a score of 0%. The event score for the analyte is the summation of all sample scores divided by the number of samples.
7. **Standard deviation index (SDI):** Once scoring of results is completed, the standard deviation index is calculated for each result. The SDI is a comparison of a result to the scoring group mean, expressed in terms of the standard deviation and is calculated as follows:

$$\text{SDI} = \frac{\text{Individual result} - \text{Mean}}{\text{Standard deviation}}$$

B - Quantitative Scoring by Peer Group

1. Results are grouped according to method codes. A method code consists of 11 digits comprised of the following four parts:
 - 4 digits that represent the instrument or kit test system
 - 3 digits that represent the reagent system
 - 2 digits that represent the measurement principle
 - 2 digits that represent an internal code used for additional grouping purposes
2. A peer group is a group of participant results using any of the following:
 - A specific test system – instrument model or kit
 - A specific method principle – ISE, certain sensitivity level or cutoff value
 - Like instruments/kits from a single vendor – various models of Radiometer blood gas instruments
 - All results for an analyte

A peer group must have a minimum of 10 results and is not used for scoring until consensus testing validates it. WSLH PT coordinators work with vendors of kits and instruments to determine what can be scored together. We strive to score by the most specific groups to give our participants the best tools to use in evaluating their performance and help in troubleshooting problematic analytes or methods.

3. The following statistics are calculated using the results for each peer group:
 - Mean
 - Median
 - Standard deviation (SD)
 - Coefficient of variation (CV)
4. After initial calculation of statistics, **outlying results** are discarded (not included in calculations) using the following rules:
 - Discard any result that is greater than ± 3 SD from the mean.
 - Recalculate mean, median, SD and CV
 - Discard any result that is greater than ± 3 SD from the mean.
 - Calculate “fences” using the median and discard any result that is outside the fences as described below.

Fence calculation: First the 25th percentile (Q1) and the 75th percentile (Q3) are calculated from the reduced population for each analyte. Then a scaling factor (SF) is defined as $1.5(Q3-Q1)$. Finally, the outer fences of the population are set as $Q1-2(SF)$ and $Q3+2(SF)$. Any result that falls outside these outer fences will be designated as a statistical outlier and discarded from further calculation.

- Recalculate mean, median, SD and CV for determination of final target value (mean).
5. The **acceptable range** is calculated by applying a specific analyte tolerance factor to the mean determined for that analyte.
 6. Both the upper and lower limits of the result range will be rounded using standard scientific rounding procedure as follows:
 - If the digit to be dropped is less than 5, the preceding figure is not altered.
 - If the digit to be dropped is greater than 5, the preceding figure is increased by 1.
 - If the digit to be dropped is 5, the preceding figure is increased by 1 if it is an odd number and the preceding figure is not altered if it is an even number.

7. Consensus for each peer group is calculated after determination of the acceptable range. Consensus is determined by counting the number of results that fall within the acceptable range and dividing by the number of included results in the peer group. If consensus is $\geq 80\%$, the peer group is approved as a scoring group. If the scoring group does not meet 80% agreement, all results in that group are scored by the “All Methods” scoring group. If the “All Methods” group does not achieve consensus, all results are flagged as non-consensus and receive a score of 100 points.
8. Each result is compared to its appropriate scoring group. If the result falls within the acceptable range, the result receives a score of 100%. If the result falls outside of the acceptable range, the result receives a score of 0%.
9. The event score for the analyte is the summation of its sample scores divided by the number of samples.
10. **Standard deviation index (SDI):** Once scoring of results is completed, the standard deviation index is calculated for each result. The SDI is a comparison of a result to the scoring group mean, expressed in terms of the standard deviation and is calculated as follows:

$$\text{SDI} = \frac{\text{Individual result} - \text{Mean}}{\text{Standard deviation}}$$

C - Enumerated Scoring by Referee-General

This scoring type uses referee labs with acceptable performance over the past three events. The **referee target determination** is as follows:

1. For each analyte/procedure/sample, **define the acceptable results** using results reported by referee laboratories. Each sample will have only one expected result.
2. Calculate the frequency for each result code.
3. Count the total number of reported results.
4. Calculate the percent occurrence for each result code by dividing the frequency by the total number of results and multiplying by 100.
5. Select and save in a table of acceptable results based on agreement of results among referee laboratories (percent consensus).
6. If $\geq 80\%$ of the referee results are for the same code, it is valid to score participant results based on referee results. The particular code is selected as the correct answer.

7. If <80% of the referee results are for the same code, more than one result code may be accepted as correct based on the judgement of the program coordinator. The coordinator will either select additional acceptable result codes or decide that it is not valid to score participant results based on referee results. (See discussion under Cell Identification Scoring Criteria, pg. 7). If it is not valid to score participants based on referee consensus, all participants receive a score of 100% for the analyte.
8. Score all participant results based on the selected acceptable referee results. Each correct result is given 100%. If the participant result is not among the acceptable referee results, it is given 0%.
9. The event score for the analyte is the summation of all sample scores divided by the number of samples.

Cell Identification Scoring Criteria

Number of Challenges. The specialty will provide five morphologic challenges per testing event. There will be three testing events per year.

Types of Challenges. Each event participant will be sent a combination of morphologic challenges that include moderately complex and highly complex skill levels. Challenges may be selected from normal, healthy individuals or may focus on a particular disease state in which normal and abnormal cells may be present. In these cases cells characteristic to a particular disease state will be selected. All challenges for an event will come from the same sample or donor whenever possible.

Types of Responses. A glossary of acceptable responses will be provided to participants each event. This is a comprehensive list that will meet the needs of moderate and high complexity laboratories.

Moderate Complexity Laboratories: A general knowledge of cellular elements in normal peripheral blood is required. Common atypical or immature blood cells such as atypical lymphs, bands, and polychromatophilic erythrocytes should be identified. Common red blood cell morphology should also be identified. The presence of uncommon atypical or immature cells (precursor cells, large or abnormal platelets, or extensive abnormal RBC morphology) needs to be recognized and **referred**.

High Complexity Laboratories: A comprehensive knowledge of normal and abnormal/immature production in all cell lines is required. All distinctive morphological characteristics, both normal and abnormal, in all cell lines need to be identified. This category would include blast cells, prolymphocytes, plasma cells, red blood cells with Howell-Jolly bodies, or other distinguishable inclusion bodies.

Determination of Target Values. The criterion for satisfactory performance for cell identification is 80% or greater consensus on each morphologic challenge among 10 or more referee laboratories. Referee laboratories are selected based on satisfactory performance from the previous year and represent a cross section of our participant population. More than one correct identification for each morphologic challenge may be considered satisfactory if there is scientific justification.

- Normal cell types (neutrophils, monocytes, basophils, eosinophils, lymphocytes, erythroid cells, and platelets) will not be combined to determine an acceptable response.
- Mature and immature cells will not be combined to determine an acceptable response.
- Malignant and benign cells will not be combined to determine an acceptable response.
- Abnormal findings critical to the diagnosis of a certain disease state (blasts, malignant cells, infectious agents, or sickle cells) must be correctly identified as such and will not be grouped with other cell types.

Scoring. Scoring of participant responses will be based on referee consensus if 80% consensus is reached. All responses that match the referee response for a particular challenge are given “satisfactory” status and all responses that do not match are given “unsatisfactory” status. If referee consensus for a particular challenge does not reach 80%, all responses are not scored and are deemed satisfactory.

Determination of Analyte Score. Number of acceptable responses for analyte / Total number of challenges x 100 = Analyte score/testing event

D - Enumerated Scoring by Referee-Microbiology

This scoring type uses referee labs with acceptable performance over the past three events. The **referee target determination** is as follows:

1. For each procedure/sample, **the acceptable results** are defined using results reported by referee labs, referring to the Table of Equivalent Results as necessary. The Table of Equivalent Results contains all possible pathogenic organisms, with each organism having its own list of equivalent (acceptable) results. For example, acceptable responses (depending on the complexity of testing a participant performs) for *Pseudomonas aeruginosa* might include: Gram negative bacterium, Gram negative rod, *Pseudomonas* species or *Pseudomonas aeruginosa*.
2. The total number of reported results for each procedure/sample and the frequency of results are counted for each identification code (from an organism glossary).
3. The percentage is calculated for each identification code. This is the frequency

divided by the total number of results x 100.

4. A Referee Results Summary Table is created that includes: identification term, identification code from the glossary, frequency of results and percentage for each procedure and specific sample.
5. For each referee Results Summary Table, the identification codes are compared to the Table of Equivalent Results by organism. Results are acceptable (equivalent) when the reported results are contained in the table. For each identification code contained in the table, the percentages of the equivalent results are totaled.
6. If the percent agreement (consensus) among referee labs is $\geq 80\%$, it is acceptable to score participant results based on referee results. The target organism and its acceptable equivalent results are stored in a table.
7. Participant results are compared to the target organism and equivalent result table. If a participant result is found in the list of equivalent results, a 100% score is assigned. If a participant result is not one of the equivalent results, a 0% score is assigned.
8. If percent agreement (consensus) among referee labs is $< 80\%$, it is not acceptable to score participant results and all participants receive a score of 100% for the procedure/sample.
9. Participants are penalized for reporting extraneous organisms that are known to be absent in the sample.

E - Enumerated Scoring by Peer Group

1. For each analyte/sample/procedure, all participant results are grouped according to peer code (table of all possible method codes).
2. The total number of reported results and the frequency of results are counted for each result reported.
3. The percentage is calculated for each result by dividing the frequency by the total number of results and multiplying by 100.
4. For each peer group, the mode (most frequently occurring result) is calculated, selected and stored. The mode indicates the correct result.
5. If $\geq 80\%$ of the participant results equal the mode, it is valid to score all participant results based on peer group results.
6. If $< 80\%$ of participants results do not equal the mode, more than one result may be accepted as correct based on the judgement of the coordinator. The coordinator will select additional acceptable result codes or decide that it is not valid to score

participant results based on peer group results. In the case where responses are expressed as dilutions, if the target value falls between two dilutions, the range becomes the greater dilution plus one dilution and the smaller dilution minus one dilution. If 80% peer group consensus is not achieved, all results in that group are scored by the “All Methods” scoring group. If the “All Methods” group does not achieve consensus, all results are flagged as non-consensus and receive a score of 100 points.

7. All participant results are scored for each analyte/procedure based on the selected acceptable results determined by the peer results.
8. Each correct answer is given a score of 100%. If the participant result is not among the acceptable results, the result is given a score of 0%.
9. For each participant reporting results, the event score is calculated by dividing the summation of all sample scores by the number of scored samples and multiplying by 100.

F - Enumerated Scoring for Antibiotic Susceptibility

Appropriate antimicrobials are identified based on the target pathogen, the sample source, and the CLSI document M100. Because the M100 is reviewed and edited by the CLSI each year, all target pathogens are cross-checked against the most current version of the M100 when determining appropriate antibiotics.

Any reported antimicrobial selections identified as inappropriate for the organism and/or source are flagged as “Unsatisfactory” on the Evaluation Report. When the antimicrobial is inappropriate, the S/I/R interpretation is over-written with “**Inappropriate Antibiotic**” and is specifically identified in the event summary.

All antimicrobials not included in the M100 for the target pathogen are manually reviewed:

- Those selected based on incorrect identification of the target pathogen which are not also appropriate for the target pathogen are flagged as inappropriate. AST score is reduced.
- Inappropriate responses are flagged as unsatisfactory and the AST score is reduced. (i.e., reporting a Gram-positive agent for a Gram-negative pathogen)
- Newer antimicrobial agents not yet addressed by the M100 are manually passed if references/literature and in-house expertise support its selection as appropriate. If insufficient information is available, the agent is flagged as “Not scored”. There is no reduction of the AST score for either action.

Evaluation of the AST Interpretation:

Antimicrobials reported on the AST result form without a corresponding interpretation or

comment are considered incomplete and are processed as “Missing/Invalid result.” These are listed by analyte name with an interpretation code of “M” and flagged with an asterisk as “Unsatisfactory”. The total AST score is reduced even if the antimicrobial selection would have been appropriate.

Interpretations are evaluated for appropriate antimicrobials only when the peer group size is ≥ 10 and consensus has been achieved for a single interpretation. Because all participants are testing aliquots from a single stock culture/strain, it is reasonable to expect that all subcultures will exhibit similar growth and susceptibility testing characteristics. Therefore, if 80% or more of participants appropriately report the pathogen as “Susceptible” to a particular agent, the interpretation “Susceptible” is treated as the “most correct” response and participants reporting either “Intermediate” or “Resistant” are penalized. If these participants also provided MIC values and the coordinator is able to identify the source or error as a change in the M100’s listed breakpoint for that agent, then this will be included in the AST discussion of the Event Summary.

Participants reporting MIC values must give both a correct/appropriate MIC *and* a correct interpretation in order to receive a passing result for that antimicrobial. For example, according to the CLSI M100-S19 document, a MIC of 2 for Ciprofloxacin should be interpreted as “intermediate.” Therefore, if a participant reported a MIC of 2 and an interpretation of “susceptible” for Ciprofloxacin, the antimicrobial would receive a score of 0%.

When fewer than 10 participants report any single agent, or when consensus is not achieved on the interpretation, the agent will still be included in the Evaluation Report, but will not have an interpretation listed in the “Accepted Result” column.

To assist participants with their written post-event work, a table of participant responses for each/any Susceptibility Testing is included in the Event Summary. This table currently lists all reported antimicrobials and the reported interpretations.

- References.** Federal Register, Vol. 57, No. 40, Section 493.941, p. 7159.
Federal Register, Vol. 58, No. 141, Section, p. 39873.
Clinical and Laboratory Standards Institute; *Performance Standards for Antimicrobial Susceptibility Testing; Seventeenth Informational Supplement, M100-S19*, Vol. 29, No. 3, January 2009