

**NOTICE: This document is a TAPPI Standard in ballot. Although available for public viewing, it is still under TAPPI's copyright and may not be reproduced or distributed without permission of TAPPI. This document is current under review to be maintained as a TAPPI Standard.**

WI \_\_\_\_\_ 230800.01 \_\_\_\_\_

T \_\_\_\_\_ 1200 \_\_\_\_\_

BALLOT NO. \_\_\_\_\_ 02 - SARG \_\_\_\_\_

DRAFT NO. \_\_\_\_\_ 01 \_\_\_\_\_

DATE \_\_\_\_\_ May 16, 2024 \_\_\_\_\_

WORKING GROUP  
CHAIR \_\_\_\_\_ N/A \_\_\_\_\_

SUBJECT  
CATEGORY \_\_\_\_\_ P&PQ General \_\_\_\_\_

RELATED  
METHODS \_\_\_\_\_ See "Additional Information" \_\_\_\_\_

**CAUTION:**

This Test Method may include safety precautions which are believed to be appropriate at the time of publication of the method. The intent of these is to alert the user of the method to safety issues related to such use. The user is responsible for determining that the safety precautions are complete and are appropriate to their use of the method, and for ensuring that suitable safety practices have not changed since publication of the method. This method may require the use, disposal, or both, of chemicals which may present serious health hazards to humans. Procedures for the handling of such substances are set forth on Safety Data Sheets which must be developed by all manufacturers and importers of potentially hazardous chemicals and maintained by all distributors of potentially hazardous chemicals. Prior to the use of this method, the user must determine whether any of the chemicals to be used or disposed of are potentially hazardous and, if so, must follow strictly the procedures specified by both the manufacturer, as well as local, state, and federal authorities for safe use and disposal of these chemicals.

## **Interlaboratory Evaluation of Test Methods to Determine TAPPI Repeatability and Reproducibility**

*(Five-year review of Standard Practice T 1200 sp-19)*

## 1. Scope

1.1 This practice describes techniques for conducting and analyzing the results of intralaboratory and interlaboratory studies. The steps described here will result in a good statistical design that provides sound data for formulating a broadly applicable precision statement regarding the performance of a TAPPI test method.

1.2 Two values are considered: (a) repeatability, which is defined as comparison of test results within a laboratory (same material, operator, apparatus, environmental conditions, making tests in the shortest reasonable timeframe); and (b) reproducibility, which is defined as comparison of test results among laboratories (same material, but different operator, apparatus and perhaps environmental conditions).

1.2.1 In the data chain leading to test results there are many possible sources of variation, and one can conduct studies to isolate these other sources, e.g. same laboratory and operator using different instruments or different laboratories using a shared calibration standard etc. For the purposes of TAPPI test methods, all of these sources of variation are to be captured in a reproducibility value.

## 2. Summary of practice

2.1 Reliable data may come from any of the following sources: existing data from a known source with sufficient supporting information, intralaboratory study, or interlaboratory study.

2.2 To conduct an interlaboratory study, randomized test specimens of each of several materials are distributed to the participating laboratories, along with appropriate instructions and data sheets. Data returned by the laboratories are analyzed one material at a time. (An assumption that different materials will yield similar precision data is NOT made.)

2.3 Tabular, statistical, and graphical diagnostic procedures are used to estimate averages and variance of the method under question and to show unusual features of the data and patterns among materials that may affect the validity or form of the precision statement.

2.4 The information is appropriately summarized in a form from which the precision statement may be readily understood and applied by users of the test method.

## 3. Significance

3.1 The precision statement gives an indication of the level of variability users of a TAPPI test method can expect. Examples of applications of these values are shown in Appendix A.3.

3.2 *TAPPI Standards: Regulations and Style Guide* requires precision statements in TAPPI Official and Provisional Test Methods. Official Test Methods must have both repeatability and reproducibility estimates. Provisional Test Methods must have repeatability estimates.

## **4. Definitions**

4.1 *Test determination* – the value obtained by carrying out the series of operations specified in the test method whereby readings are made on a test specimen.

4.2 *Test result* – the value obtained for one test unit of the sample. The result is either a single test determination or a specified combination of a number of test determinations (e.g., an average). The method must describe the manner in which each test determination is to be made, the number of test determinations to be made, and how these determinations must be combined to give the test results. For example, the test method for paper thickness, T 551 “Thickness of paper and paperboard (soft platen method)” requires one determination for each of 10 specimen strips; the test result for thickness is defined as the average of these 10 test determinations.

4.3 *Repeatability* – an estimated limit below which the difference between two test results is expected to fall 95% of the time, when the results are obtained at a single laboratory. Repeatability reflects the variability inherent in the experimental system used to determine its value(s).

4.4 *Repeatability conditions* – conditions that reflect the most controlled experimental circumstances possible. Major measurement biasing factors should remain constant including: equipment, operator, test procedure, environmental conditions, sample preparation and in a single testing interval using presumably homogenous samples.

4.5 *Reproducibility* – an estimated limit below which the difference between two test results is expected to fall 95% of the time, when the results are obtained in different qualified laboratories for the same homogeneous source of material. While each laboratory tests samples taken at random from a single source of material, all of the other factors influencing measurements not considered for repeatability contribute to the variability in the results obtained in different laboratories.

4.6 *Reproducibility conditions* – conditions that reflect only the most basic control over measurement-influencing factors. Reproducibility conditions should include qualified personnel in qualified laboratories carrying out the same procedure (test method) on different samples of the same presumably homogeneous material. Qualified laboratories are those having the required properly calibrated apparatus and environmental conditions and trained personnel who conscientiously follow the prescribed test method (see Section 5.3 for further information). Many of the influences that are considered constant in repeatability are not controlled for reproducibility conditions including: equipment, operator, environmental conditions etc.

4.7 The 95% probability level for both repeatability and reproducibility (4.4 and 4.6) is generally accepted for precision estimations and is also chosen by other standard writing organizations (e.g. ASTM and ISO). Thus, TAPPI repeatability and reproducibility can be considered compatible with the precision statistics in methods currently published by ASTM and ISO. Repeatability and reproducibility are different from, but related to, the 95% confidence interval for a particular test result. Confidence intervals are often estimated by performing replicates and correcting for the number of replicates using Student’s *t* test. For more information see:

<http://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm>.

4.8 *Intralaboratory study* – a study designed to estimate the precision (repeatability) of a test method that meets the requirements of Section 5 and is conducted in a single laboratory.

4.9 *Interlaboratory study* – a study designed to estimate the precision (repeatability and reproducibility) of a test method that meets the requirements of Section 5 and is conducted at multiple laboratories.

## **5. Planning the intra- or interlaboratory study**

5.1 *Basic design.* Keep the design as simple as possible in order to obtain estimates that are an accurate reflection of the method in common usage. The basic design is represented by a two-way classification table in which the rows represent the laboratories, the columns represent the materials, and each cell (i.e., the intersection of a row with a column) contains the measurements made by a particular laboratory on a particular material (Table 1 and Table 3).

### *5.2 Test method.*

5.2.1 The test method should be reviewed to assure that it has been well-written and subjected to careful investigation at one or more competent laboratories to determine clarity and stability. Furthermore, it should be easy to follow by any properly equipped laboratory with competent personnel having knowledge of the materials and of the property to be tested.

5.2.2 The method should have a clearly defined numerical result which can be derived from a single determination or a combination of multiple determinations. The results are assumed to follow a normal distribution.

5.2.3 The method should specify the care of test specimens, including prominent instructions for preconditioning and conditioning when required (see Section 6.2.4).

5.2.4 The method should specify the need for and frequency of adjustment and calibration of the test apparatus.

### *5.3 Laboratories.*

5.3.1 Any laboratory that would be considered qualified to run the test in routine testing situations is permitted and encouraged to participate in the interlaboratory study, but the number of laboratories may be restricted if there is a large amount of work involved in preparing samples or there is a significant increase in sampling variability due to the larger amount of material required.

5.3.2 Laboratories must be properly equipped to follow all details of the procedure and be willing to assign the work to a skilled operator on a timely basis. The decision of a laboratory to participate should be based on information provided by the working group, including information as to the required time for calibrating the apparatus and for testing all of the materials.

5.3.3 Subject to the conditions in 5.3.1 and 5.3.2, use as many laboratories as practical, preferably 20 to 30. Because of the possible loss of data from some laboratories for some materials, it is highly desirable to include at least eight laboratories to obtain a usable estimate of the reproducibility of the test method. In consideration of the referee nature of TAPPI test methods, more than a mere usable estimate of the reproducibility between laboratories is highly desirable. Because the data required from each laboratory are held to a minimum in this practice, to encourage the

participation of an adequate number of laboratories. Under no circumstances should a precision statement be based on an interlaboratory trial with fewer than five laboratories.

5.3.4 Avoid restricting the study to a group of laboratories judged to be exceptionally qualified and equipped.—Precision estimates for inclusion in a TAPPI test method should be obtained through the efforts of laboratories and personnel operating under conditions that are representative of the situations in which the test method will be used in practice.

5.4 *Materials.*

5.4.1 *Number and types.* The number and types of materials to be included in the interlaboratory study should (1) span the range of the values of the property being measured, (2) span the number of scales to cover the range if the instrument has more than one scale, and (3) reflect the number of types or classes of materials to which the test method is to be applied. If the precision trial is restricted in any of these areas, the omitted information must be noted in the precision statement.

5.4.2 Without previous knowledge concerning the behavior of different classes of material subject to the method, it should be assumed that different materials will exhibit different performance characteristics. When a method covers different classes of materials, it may be desirable to conduct separate trials to suit the expertise of the laboratories involved. Thus, the selected materials within each class differ from each other only in level and should not pose difficulties to the participating laboratories.

5.5 *Replicates.*

5.5.1 If the test method defines a test result as the average of a set number of test determinations, the study should consist of at least three times that number of replicate determinations.

5.5.2 If the test method requires a recalibration or some other special procedure between each test result but not between each test determination, each replicate in the study must be a complete test result, and at least three replicate results should be obtained.

## **6. Conducting an intra- or interlaboratory study**

6.1 The working group should prepare a protocol for the participants in the study, which should include detailed instructions for the following:

6.1.1 Any exception to or selection of the test procedure as specified in the test method. Expected date of the testing, especially if the aging of samples may be important (see 6.2.3);

6.1.3 Any special circumstances that must be addressed in implementing the repeatability conditions (one operator, one apparatus, shortest time span);

6.1.4 The recording of measurements, including the number of decimal places to be used, which, when possible, should be one more than is required to be reported by the test method under study in order to allow reliable calculation of the precision of the method;

6.1.5 The keeping of a log of unusual events that arise during any phase of the testing;

6.1.6 Notifying the study coordinator when an error in test procedure arises so that replacement test specimens may be sent.

6.2 *Sampling and sample preparation and distribution.*

6.2.1 *Sampling of materials.*

6.2.1.1 Sample each material so that the variability among the specimens of that material will be minimized.

6.2.1.2 Prepared in accordance with the sampling section of the method, if applicable.

6.2.1.3 For samples of paper, take all specimens from a small area of a single roll, avoiding the edges of the roll. Usually, specimens taken adjacent to each other in the machine direction are more similar than those adjacent in the cross direction. Similarly, composite specimens, e.g., corrugated board, should be sample from sections of the roll or stack that are expected to be homogenous. As a further refinement for physical tests, each sheet or specimen may be weighed individually to weigh each sheet or specimen and remove any that exhibit weights more than two standard deviations from the mean. Specimens for nondestructive optical tests may be individually measured in one laboratory. Those specimens with test values that exhibit values more than two standard deviations from the mean can be removed from the study.

6.2.1.4 For a granular solid, liquid or similar material representing any given test level, make the material as homogeneous as possible by mixing or other appropriate means prior to its subdivision into portions or specimens that are allocated to the laboratories.

6.2.2 *Preparation of specimens.*

6.2.2.1 Prepare enough specimens from each sample to provide the required test material for all participating laboratories and a significant number of additional specimens for replacement of any lost or spoiled specimens. Label each specimen by means of a code symbol and record the identification of the specimens for future reference. Completely randomize the specimens of a particular test material before dividing them into the required number of groups for assignment to the laboratories.

6.2.2.2 The complete randomization of specimens, as specified above, ensures that between-specimen variability is the same within laboratories as between laboratories and greatly simplifies the statistical analysis and interpretation. When each laboratory is given a single sheet of each material from which it is expected to obtain all specimens, intralaboratory variability is confounded with between-sheet variability. On the other hand, when blocks of adjacent specimens (which are presumably more alike) are assigned one from each block to each laboratory, the interlaboratory replication error is confounded with block-to-block variability and cannot be compared with the intralaboratory variability, which does not include the block-to-block variability.

6.2.3 *Aging of samples.* If the samples are such that their properties may significantly change in the course of days or a few weeks, coordinate the dates for the testing among the participating laboratories so that the effect of aging is not confounded with the differences among laboratories.

6.2.4 *Conditioning of samples.* If required by the test method under investigation, preconditioning (see TAPPI T 402 “Standard Conditioning and Testing Atmospheres for Paper, Board, Pulp Handsheets, and Related Products”) of the samples at low relative humidity (RH) prior to conditioning and testing at 50% RH will avoid confounding hysteresis effects with the differences among laboratories. The preconditioning and conditioning should

be left to each participating laboratory, although a preliminary preconditioning could be done while the test specimens are being randomized.

6.2.5 Allocation and shipment of specimens. Allocate and distribute the specimens from a single place. Ensure that the specimens are packaged properly for arriving without meaningful changes in the property under consideration, and addressed as specified by the participant for prompt delivery to the testing laboratory.

6.3 *Report form and questionnaire.* Supply each participating laboratory with a form for reporting the data to ensure that all pertinent data and information are reported in a uniform manner by all participants. In addition, provide space for other relevant information, such as relative humidity, temperature, type of instrument, and log of unusual events.

**7. Calculating only repeatability from an intralaboratory study (for Provisional Methods)**

7.1 *Data display.* Place the data to be used for the calculation of repeatability into a table. Display the materials as columns. Display the determination or result values under the respective materials. Place the calculation results at the bottom of the table. See Table 1 for a generalized example, and Table A1 for a specific example. To assist in conducting these calculations an Excel spreadsheet is available for downloading from the general Standards page on the TAPPI website (go to any TAPPI web page, click on the “Standards” button at the top of the page, then scroll down to the “General Information Document list.”).

7.2 *Data inspection.* The working group reviews the data to ensure that sufficient data have been gathered to form a reasonable precision statement. Because the calculation results will not be combined, there is not a requirement that the data be balanced between the materials, provided that sufficient data are present to adequately fill the requirements of Section 5. For example, the cells for Materials A, B and C may consist of 5 test results, whereas the cell for Material D may consist of only 4 test results due to a testing error.

**Table 1.** General example of the data display for an intralaboratory trial

<i>Material A</i>	<i>Material B</i>	<i>Material C</i>	<i>Material D</i>
Data 1, Data, 2, Data 3, etc.	Data 1, Data, 2, Data 3, etc.	Data 1, Data, 2, Data 3, etc.	Data 1, Data, 2, Data 3, etc.
calculation results	calculation results	calculation results	calculation results

7.3 *Calculations for repeatability.* Prior to performing the repeatability calculations, the working group members should evaluate the quality of the data (see section 9). For each material, use the equations and definitions shown in Table 2 to calculate means, standard deviations, repeatability, and repeatability ratio. Repeatability is related to the repeatability standard deviation ( $s_r$ ) of results obtained in accordance with the limitations outlined in Section 4.5. TAPPI repeatability ( $r$ ), is chosen to be 2.77 times the repeatability standard deviation ( $s_r$ ) and is an estimate of the maximum difference between two test results at a 95% confidence level. The factor  $2.77 = 1.96 \times \text{SQRT}(2)$ , where 1.96 is the factor for symmetric 95% limits for a two-sided normal distribution and  $\text{SQRT}(2)$  results from the fact that one is calculating the difference between two normally distributed test results.

## **8. Calculating repeatability and reproducibility from an interlaboratory study (for Official Methods)**

8.1 *Data display.* Place the data to be used for the calculation of repeatability into a table. The materials shall be displayed as columns. The measurements from the laboratories shall be displayed as rows under the respective materials for each laboratory. Group the data by laboratory, and reserve the rows following each laboratory's data for calculation results. Place the summary results at the bottom of the table. See Table 3 for a generalized example, and Table A2 for a specific example. To assist in conducting these calculations an Excel spreadsheet is available for downloading from the general Standards page on the TAPPI website (go to any TAPPI web page, click on the "Standards" button at the top of the page, then scroll down to the "General Information Document list").

8.2 *Data inspection.* The working group must review the data to ensure that sufficient data have been gathered to form a reasonable precision statement. Because the calculation results will not be combined, there is not a requirement that the data be balanced between the materials provided that sufficient data is present to adequately fill the requirements of Section 5. For example, a laboratory's submitted data for Materials A,B and C may consist of 5 test results, whereas the cell for Material D contains only 4 test results due to a testing error. Another example would be if a laboratory does not have the capability to test a particular material, so that the precision of Materials A, B, D and E are based on 15 laboratories and the precision for Material C is based on 14 laboratories. The calculations described below are based on the assumption that each laboratory provides the same number of test results, i.e., a balanced design. Small deviations, one or two omissions, from this assumption do not have a large impact on the results. If, however, the differences in the number of results are greater than two or if several labs have missing data, then these equations should not be used. In this case of unbalanced designs, the working group should have the calculations performed by a statistician.



Table 2. Equations for calculating repeatability

<p><b>Definitions</b></p> <p><math>x</math>'s are test determinations or results</p> <p><math>n</math> is the number of determinations or results</p> <p><math>q</math> is the number of test determinations averaged to calculate a test result (note <math>q</math> is set to 1 if a single determination is a result or if results are being used for repeatability calculation)</p>	
Material mean	$\bar{x} = \sum x/n$
Material standard deviation	$s_x = \sqrt{\sum(x - \bar{x})^2 / (n - 1)}$
Repeatability standard deviation	$s_r = \frac{s_x}{\sqrt{q}}$
Repeatability	$r = (2.77) s_r$
Repeatability ratio	$\%r = 100( r / \bar{x} )$

**Table 3.** Specific example of the data display for an interlaboratory trial

<i>Material A</i>	<i>Material B</i>	<i>Material C</i>	<i>Material D</i>
Lab #1 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>	Lab #1 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>	Lab #1 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>	Lab #1 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>
calculation results	calculation results	calculation results	calculation results
Lab #2 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>	Lab #2 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>	Lab #2 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>	Lab #2 <u>Data 1, Data 2,</u> <u>Data 3, etc.</u>
calculation results	calculation results	calculation results	calculation results
Continue with data from the rest of the labs	Continue with data from the rest of the labs	Continue with data from the rest of the labs	Continue with data from the rest of the labs
Summary calculation results	Summary calculation results	Summary calculation results	Summary calculation results

8.3 *Calculations for repeatability and reproducibility.* Prior to performing the repeatability and reproducibility calculations, the working group members should evaluate the quality of the data (see section 9). For each material, use the equations and definitions shown in Table 4 to calculate means, standard deviations, repeatability, repeatability ratio, reproducibility, and reproducibility ratio. Repeatability and reproducibility are related to the pooled standard deviation ( $s_p$  or  $s_R$ ) of results obtained in accordance with the limitations outlined in Section 4.5. See Section 7.3 for an explanation of the origin of the factor 2.77.

**Table 4.** Equations and definitions to calculate means, standard deviations, repeatability, repeatability ratio, reproducibility, and reproducibility ratio

<b>Definitions</b>	
x's are test determinations	
p is the number of laboratories	
n is the number of determinations at each laboratory	
q is the number of test determinations averaged to calculate a test result (note: q is set to 1 if a single determination is a result or if results are being used for repeatability and reproducibility calculation)	
Laboratory material mean (there are p of these)	$\bar{x} = \sum x/n$
Laboratory material standard deviation	$s_x = \sqrt{\sum(x - \bar{x})^2 / (n - 1)}$
Grand material mean	$\bar{\bar{x}} = \sum \bar{x} / p$
Grand material standard deviation	$s_{\bar{x}} = \sqrt{\sum (\bar{x} - \bar{\bar{x}})^2 / (p - 1)}$
Pooled standard deviation	$s_p = \sqrt{\sum s_x^2 / p}$
Repeatability standard deviation	$s_r = \frac{s_p}{\sqrt{q}}$
Reproducibility standard deviation	$s_R = \sqrt{s_x^2 + s_p^2(n - q)/nq}$
Repeatability	$r = (2.77)s_r$
Repeatability ratio	$\%r = 100(r / \bar{\bar{x}})$
Reproducibility	$R = (2.77)s_R$
Reproducibility Ratio	$\%R = 100(R / \bar{\bar{x}})$
h consistency statistic (there are p of these)	$h = (\bar{x} - \bar{\bar{x}}) / s_{\bar{x}}$
k consistency statistic (there are p of these)	$k = s_x / s_p$

## 9. Consistency evaluation

9.1 The working group members' common sense is the most important aspect in evaluating the data and implies the importance of having committee members with substantial experience and knowledge. It is the responsibility of each member of the working group to look carefully at the data and apply their common sense in evaluating the data. A quick glance by an informed scientist or engineer may often detect a problem with data that a fixed set of formal or graphical statistical procedures might miss. For example, a real interlaboratory study led to the following laboratory averages: 655, 1057, 1249, 1353, 1489, 1625, 1682, and 1825. (These laboratory averages were each an average of five test determinations.) The associated within-laboratory standard deviation associated with these test results was approximately 260 so the standard deviations of the laboratory averages were approximately  $260/\text{SQRT}(5) = 116$ . The associated coefficient of variation of the laboratory averages was 28%. Neither graphical procedures nor formal statistical tests detected a problem with these data. However, the working group knew that a 3 to 1 ratio of the largest to the smallest average measurement on the same material was simply too extreme for a well-defined measurement process of the material under consideration. The data discrepancy led the working group to decide that either the measurement procedure was not sufficiently well-defined or that one or more of the laboratories was having problems with their equipment or their operators.

9.2 Graphical procedures such as side-by-side box plots of laboratory results might aid committee members in evaluating the test results. (Note the National Institute of Standards and Technology (NIST) web page discusses checking for outliers via box plots:

<http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>. Most modern statistical packages can produce such graphs.

9.3 An assumption that data are normally distributed is required to use that procedures described in this standard. A Shapiro-Wilk test of normality, normal probability plots (<http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm>), or chi squared tests can all be used to determine if the data appear to be normally distributed. Working group members are encouraged to seek professional statistical support in evaluating the data distribution.

9.4 Consistency statistics  $h$  and  $k$  can be used to flag laboratories that are providing data that need investigation. A single outlying laboratory average will likely be flagged by an  $h$  test. However, if more than one laboratory average is an outlier, then the  $h$  test might no longer perform well. When multiple outliers are present, there is a tendency to mask the presence of each other in formal (non-visual) tests. If the graphical outlier checks indicate the possible presence of multiple outliers, a professional statistician should be consulted. The calculation of  $h$  and  $k$  applies only to data derived from interlaboratory studies.

9.5 It is often helpful to study the consistency statistics when displayed grouped by laboratory and also by material. This display of consistency statistics may be done numerically or graphically (bar charts of values). An example of the recommended data display for a study with 15 laboratories and 4 materials is depicted in Figure 1 and Figure 2. The goal of a displaying the data in this matter is to detect unusual features in the study data that may bias the calculated precision parameters.

9.6 Table 5 lists the critical values of  $h$  and  $k$  at the 0.5% significance level. To obtain the particular critical values from this table, search down the rows to find the number of laboratories ( $p$ ) and then across the columns to find the correct number of replicates ( $n$ ) in the study. When a calculated consistency statistic exceeds this value, the laboratory data shows excessive variation.  $h$  values in excess of the critical value indicate excessive variation from the Grand Mean.  $k$  values in excess of the critical value indicate excessive variation from the pooled material standard deviation. Mark the values in the display of consistency statistics that exceed the critical value. Data that exceed critical values should be studied by the working group to determine if the data adversely affect the calculation of precision parameters. It is important to note that all data that exceed the critical value are not automatically “bad,” because by definition 1 of every 200 results may exceed the critical value. If the working group determines that the data are not truly indicative of method performance, then the data should be treated according to 9.2.

9.7 *Handling of outlying data.* When a laboratory consistency statistic suggests that data are not indicative of the study, but are in error, there are two possible actions: omit or retest. A retest will be the preferred method of data adjustment.

9.7.1 *Data adjustment, retest and replacement.* If it is determined that the material should not have changed significantly since the conduct of the original study, laboratories with erroneous data may be allowed to retest in order to replace their data. Review the test method and testing instructions with the laboratory to correct any procedural variations that may have caused the error.

9.7.2 *Data adjustment, omit.* If it is suspected that the material has changed since the conduct of the original study, it is deemed that a retest is not possible, too burdensome, or a laboratory has non-correctable procedural flaws; laboratory data may need to be omitted. If sufficient data exists, the remaining data may be used as part of the precision statement.

## **10. Statement of the precision of the test method**

10.1 *Combination of repeatability and reproducibility estimates.* After reasonable values for TAPPI repeatability and reproducibility have been determined from the analysis for individual materials these estimates may be combined to provide a statement of the precision of the test method. Resist the temptation to combine  $r$  and  $R$  estimates. Unless it is determined that the estimates are indeed the same, precision results should be listed separately for different materials. Expect that different materials, instrumentation or testing protocols will yield different estimates of precision. For example, the brightness of fluorescent papers may be expected to vary differently than primarily non-fluorescent papers or two types of instrumentation may yield similar results but show different variations.

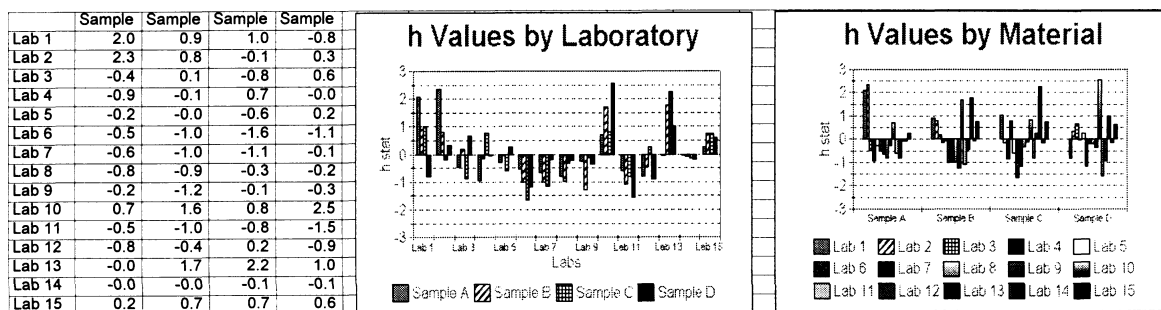


Fig. 1. Display of *h* stats

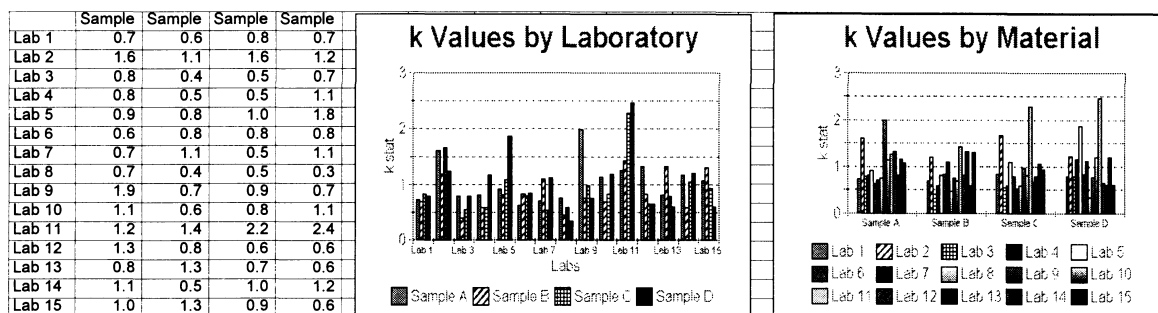


Fig. 2. Display of *k* stats

Table 5. Critical values of  $h$  and  $k$  at the 0.5% significance level

# labs	$h$ critical	$k$ critical								
		$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
$p = 3$	1.15	1.72	1.67	1.61	1.56	1.52	1.49	1.47	1.44	1.42
$p = 4$	1.49	1.95	1.82	1.73	1.66	1.60	1.56	1.53	1.50	1.47
$p = 5$	1.74	2.11	1.92	1.79	1.71	1.65	1.60	1.56	1.53	1.50
$p = 6$	1.92	2.22	1.98	1.84	1.75	1.68	1.63	1.59	1.55	1.52
$p = 7$	2.05	2.30	2.03	1.87	1.77	1.70	1.65	1.60	1.57	1.54
$p = 8$	2.15	2.36	2.06	1.90	1.79	1.72	1.66	1.62	1.58	1.55
$p = 9$	2.23	2.41	2.09	1.92	1.81	1.73	1.67	1.62	1.59	1.56
$p = 10$	2.29	2.45	2.11	1.93	1.82	1.74	1.68	1.63	1.59	1.56
$p = 11$	2.34	2.49	2.13	1.94	1.83	1.75	1.69	1.64	1.60	1.57
$p = 12$	2.38	2.51	2.14	1.96	1.84	1.76	1.69	1.64	1.60	1.57
$p = 13$	2.41	2.54	2.15	1.96	1.84	1.76	1.70	1.65	1.61	1.58
$p = 14$	2.44	2.56	2.16	1.97	1.85	1.77	1.70	1.65	1.61	1.58
$p = 15$	2.47	2.57	2.17	1.98	1.86	1.77	1.71	1.66	1.62	1.58
$p = 16$	2.49	2.59	2.18	1.98	1.86	1.77	1.71	1.66	1.62	1.58
$p = 17$	2.51	2.60	2.19	1.99	1.86	1.78	1.71	1.66	1.62	1.59
$p = 18$	2.53	2.61	2.20	1.99	1.87	1.78	1.72	1.66	1.62	1.59
$p = 19$	2.54	2.62	2.20	2.00	1.87	1.78	1.72	1.67	1.62	1.59
$p = 20$	2.56	2.63	2.21	2.00	1.87	1.79	1.72	1.67	1.63	1.59
$p = 21$	2.57	2.64	2.21	2.00	1.88	1.79	1.72	1.67	1.63	1.59
$p = 22$	2.58	2.65	2.21	2.01	1.88	1.79	1.72	1.67	1.63	1.59
$p = 23$	2.59	2.66	2.22	2.01	1.88	1.79	1.72	1.67	1.63	1.59
$p = 24$	2.60	2.66	2.22	2.01	1.88	1.79	1.73	1.67	1.63	1.60
$p = 25$	2.61	2.67	2.23	2.01	1.88	1.79	1.73	1.67	1.63	1.60
$p = 26$	2.62	2.67	2.23	2.02	1.89	1.80	1.73	1.68	1.63	1.60
$p = 27$	2.62	2.68	2.23	2.02	1.89	1.80	1.73	1.68	1.63	1.60
$p = 28$	2.63	2.68	2.23	2.02	1.89	1.80	1.73	1.68	1.63	1.60
$p = 29$	2.64	2.69	2.24	2.02	1.89	1.80	1.73	1.68	1.64	1.60
$p = 30$	2.64	2.69	2.24	2.02	1.89	1.80	1.73	1.68	1.64	1.60

The above critical values for the  $h$  and  $k$  consistency statistics were calculated from Student  $t$  and the  $F$ -ratio using the following relationships:

$p$  = number of participating laboratories

$n$  = number of results at each laboratory

$$h = (p - 1)t / \sqrt{p(t^2 + p - 2)} \quad t \text{ with } p - 2 \text{ degrees of freedom and a significance level of } 0.5\%$$

$F$  with  $n - 1$  and  $(p - 1)(n - 1)$  degrees of freedom and a significance level of 0.5%

$$k = \sqrt{p/[1 + (p - 1)/F]}$$

10.1.1 *Combinations of repeatability or reproducibility.* If the repeatability and/or reproducibility estimates for different levels of similar materials or for different materials are approximately the same, as determined by the working group, the estimates may be combined as an average. In doing this, the working group is confirming the additive model of variation was primary in the data for either different levels of a single type of material or across all materials. That is, there is a single numeric (non-proportional) variability associated with all results on the material in question, or across all materials (all values for  $r$  or for  $R$  are approximately equal).

10.1.2 *Combinations of repeatability or reproducibility ratios (%).* If the repeatability and/or reproducibility estimates derived as ratios for different levels of similar materials or for different materials are approximately the same, as determined by the working group, combine the estimates as an average. In doing this, the working group is confirming the concurrent model of variation was primary in the data for either different levels of a single type of material or across all materials. That is, there is a single proportional variability (numeric variability changes as the test result) associated with all results on the material in question, or across all materials (all values for % $r$  or for % $R$  are approximately equal).

10.2 *Suggested information for precision statements.* Because the precision listed in a method is an estimate of the variability of the method dependant on the data on which it was based, it is critical that users of the method be given sufficient supporting information to determine if the estimates apply to their testing parameters. For example, if the data were derived from a single type of material, the estimates may not be applicable to wholly different materials.

10.2.1 *Data background information.* Include any information that may limit the applicability of the precision estimates. This information includes, but is not limited to the following:

- 10.2.1 Source of the data, e.g. precision interlaboratory study, company round robin, commercial testing;
- 10.2.2 Date of the study;
- 10.2.3 Revision of the method, significant additional instructions or alternate protocols;
- 10.2.4 Number of laboratories (interlaboratory only);
- 10.2.5 Number of determinations, if determinations are used instead of results;



10.2.6 Material descriptions;

10.2.7 Repeatability and reproducibility standard deviations;

10.2.8 Repeatability and reproducibility estimates;

10.2.10 Additional statistics may be included such as grand material means, maximum and minimum test results for each material, and repeatability or reproducibility standard deviations.

10.3 *Example precision statement.* The following precision statement is provided as an **example only** (Table 6), and is not necessarily the recommended format for all precision statements.

The following estimates of repeatability and reproducibility are based on data from an interlaboratory trial conducted in 1999 using the “pm-94” revision of the test method and included test results from 19 laboratories. Additionally, laboratories were asked to clean and calibrate the apparatus between test materials. Five test results were obtained from each sample.

**Table 6 .** Property measurement results

**Example**

<b>Material Type</b>	<b>Sample</b>	<b>Repeatability</b>	<b>Reproducibility</b>
Material A	Newsprint	4%	13%
	Offset		
	Cotton bond		
Material B	Hardwood pulp	5%	19%
	Softwood pulp	7%	22%
Material C	Corrugated board	8%	15%

The repeatability and reproducibility reported above are estimates of the maximum difference that should be expected in 19 of 20 instances, when comparing two test results for materials similar to those described above under similar test conditions. These estimates may not be valid for different materials or testing conditions.

## 11. Keywords

Interlaboratory evaluation, Repeatability, Reproducibility, Precision, Testing

## 12. Additional information

12.1 Effective date of issue: To be assigned.

12.2 The 2000 revision of T 1200 reflects a combination of standard practices T 1200 sp-91 and T 1206 sp-91 in order to describe the complete formulation of precision statements in a single practice. This revision also reflects the removal of information not directly related to the formulation of precision and a simplification of the procedures. The requirement for a balanced study (same number of replicates for each cell) has been removed.

12.3 The 2000 version was a major revision. The two-way linear-model analysis of variance of TAPPI T 1200 os-69 has been replaced by a “one-material-at-a-time” analysis, which is much simpler to execute but loses some information in the process. This revision provides the information needed for the preparation of TAPPI precision statements (see TAPPI T 1206), including TAPPI repeatability and TAPPI reproducibility, but the very much more complex calculations for TAPPI comparability are no longer referenced. A numerical example is provided as a guide in following the calculations, with increased use of tables and graphs to help in understanding patterns or unusual data.

12.4 The revisions in the 2007 version include corrections for accuracy, inclusion of references to the Excel spreadsheet for accomplishing the numerical calculations available on the TAPPI standards web site, improved descriptions of the definitions of repeatability and reproducibility, corrections of the results of the calculations shown in Table A2, and addition of Appendix A.3. Application of repeatability and reproducibility.

12.5 The revisions in the 2014 version include simplification of the description of equations and the removal of a redundant section.

### Appendix A.1 Example of intralaboratory trial and repeatability calculations

A.1.1. An intralaboratory study was conducted to form an estimate of repeatability for TAPPI T 650 “Solids Content of Black Liquor.” Four samples were collected to be representative of the applicable range of the method. The test method calls for samples with a high solids concentration to be, diluted; therefore two samples that required dilution and two that did not require dilution were chosen. The test method requires three test determinations be averaged to get a test result (triplicate determinations in separate containers). The test method was carried out five times for each sample, yielding five test results, each of which was comprised of three determinations. Table A1 is a specific example of the data display for an intralaboratory trial, from 7.1 and Table 1. Please note that the units for laboratory material mean, laboratory material standard deviation, repeatability standard deviation, and repeatability are in the same units as the test results (% solids), whereas the repeatability ratio is unitless but uses (%) to denote the ratio. The test results below are reported to the nearest .01%, one digit more than called for by the test method. This “extra” significant digit was carried through the calculations until the calculation of repeatability.

A.1.2 The working group decided that the data and repeatability estimates for each of the materials are reasonable and reflect adequate estimates of the precision of the method. Furthermore, the working group decided that the estimates of repeatability for each material agree to a sufficient extent that it is appropriate to combine (by averaging) the four 4 estimates into a single estimate of repeatability. In choosing to combine the estimates, the working group indicated that users of the method, testing under similar parameters, can use the repeatability estimate for both black liquors that require dilution and those that do not require dilution. Based on the trial information, data, and analysis considerations above, the statement in A.1.3 reflects a reasonable precision statement for this method.

A.1.3 The following estimate of repeatability is based on data from an intralaboratory trial conducted in 1998 using the “om-89” revision of the method. The trial used two samples that required dilution and two that did not. Repeatability estimates listed below were derived from five test results for four samples; each result is the mean of three determinations.

**Table A1.** Specific example of the data display for an intralaboratory trial

<i>Black liquor samples</i>	<i>Sample A</i>	<i>Sample B</i>	<i>Sample C</i>	<i>Sample D</i>
Test results	43.17 % 44.20 % 44.15 % 44.23 % 43.80 %	46.82 % 46.66 % 45.90 % 45.57 % 47.23 %	69.91 % 70.67 % 70.55 % 70.20 % 69.38 %	75.41 % 76.73 % 76.58 % 75.80 % 75.71 %
Calculation results				
Laboratory material mean, $\bar{x}$	43.91 %	46.44 %	70.14 %	76.05 %
Laboratory mat stnd dev, $s_x$	0.45 %	0.68 %	0.52 %	0.58 %
Repeatability stnd dev, $s_r$	0.45 %	0.68 %	0.52 %	0.58 %
Repeatability, $r$	1.2 %	1.9 %	1.4 %	1.6 %
Repeatability ratio, % $r$	2.8 %	4.1 %	2.1 %	2.1 %

**Solids content of black liquor**

<i>Material Type</i>	<i>Repeatability, r</i>
Black liquors	1.5 % solids

The repeatability reported above is an estimate of the maximum difference to be expected in 19 of 20 instances, when comparing two test results for materials similar to those described above under similar test conditions. These estimates may not be valid for different materials or testing conditions; users of this method must determine the suitability of this estimate to their application.

**Appendix A.2 Example of interlaboratory trial and repeatability and reproducibility calculations**

A.2.1 An interlaboratory study was conducted to form estimates of repeatability and reproducibility for TAPPI, T807 “Bursting Strength of Paperboard and Linerboard.” Linerboard in the weights of 35 lb, 42 lb, and 69 lb were distributed to be representative of the applicable range of the method. The test method limits the range of applicability only to paperboard (as opposed to paper and corrugated or solid fiberboard), so the method may not truly represent all of the possible materials subject to the method. Users must determine whether the precision estimates are appropriate to their testing. For example, coated or multilayer paperboard may call for separate estimates of precision. The test method requires 10 test determinations be averaged to get a test result. The test method was carried out four times for each sample yielding four test results, each of which was comprised of 10 determinations. Table A2 is a specific example of the data display for an interlaboratory trial, from 9.1 and Table 3. An important aspect of this example is that it was is an unbalanced study. Laboratory 5 did not report data for the 69-lb sample, and Laboratory 2 had only three test results in the cell for 69-lb sample. Laboratory and summary calculations are still performed in the same manner as the other samples; these missing items were not replaced with calculated results. It is the responsibility of the working group chair to determine if missing data have adversely affected the estimates of precision calculated from the trial. Note: these data are also included as example in the Excel spreadsheet that is available from TAPPI.

**Table A2.** Specific example of the data display for an interlaboratory trial

Laboratory		35-lb Linerboard	42-lb Linerboard	69-lb Linerboard
1	Test Results	87.8 86.6 81.4 86.0	122.5 121.7 124.6 116.3	147.7 133.0 136.3 139.8
Lab calculation results	Lab material mean = Lab material stnd. dev. =	85.4 2.8	121.3 3.5	139.2 6.3
2	Test results	87.6 86.4 87.3 83.4	118.7 123.5 122.8 120.0	149.5 137.5 N/A 141.6
Lab calculation results	Lab material mean = Lab material stnd. dev. =	86.2 1.9	121.3 2.3	142.9 6.1
3	Test Results	91.3 91.3 89.2 88.7	120.6 121.2 117.8 117.5	134.5 139.9 141.3 147.1
Lab calculation results	Lab material mean = Lab material stnd. dev. =	90.1 1.4	119.2 1.9	140.7 5.2
4	Test Results	86.8 88.7	123.0 120.0	140.5 139.7

		86.1 87.2	121.9 118.0	139.7 143.9
Lab calculation results	Lab material mean = Lab material std. dev. =	87.2 1.1	120.7 2.2	140.9 2.0
5	Test Results	95.3 86.9 94.4 87.4	116.0 117.6 122.4 116.5	Not tested
Lab calculation results	Lab material mean = Lab material std. dev. =	91.0 4.5	118.1 2.9	Not tested

6	Test Results	90.5 92.8 91.3 86.0	132.9 122.8 123.2 127.0	147.0 141.7 143.8 145.2
Lab calculation results	Lab material mean = Lab material std. dev. =	90.2 2.9	126.5 4.7	144.4 2.2
7	Test Results	91.2 89.0 91.3 N/A	123.0 124.1 118.0 121.0	148.7 149.2 142.3 154.3
Lab calculation results	Lab material mean = Lab material std. dev. =	90.5 1.3	121.5 2.7	148.6 4.9
8	Test Results	89.1 88.7 87.0 88.8	120.0 113.6 113.8 118.9	141.0 144.8 139.6 139.0
Lab calculation results	Lab material mean = Lab material std. dev. =	88.4 0.9	116.6 3.4	141.1 2.6
9	Test Results	86.5 89.8 82.1 92.0	123.7 125.7 121.7 127.6	134.1 140.8 142.0 130.6
Lab calculation results	Lab material mean = Lab material std. dev. =	87.6 4.3	124.7 2.5	136.9 5.5

Summary calculation results	Grand material mean, $\bar{x}$	88.5	121.1	141.8
	Grand material stnd dev, $s_x$	2.0	3.1	3.5
	Repeatability stnd dev, $s_r$	2.7	3.0	4.7
	Repeatability, $r$	7.4	8.3	12.9
	Repeatability ratio, % $r$	8.4%	6.9%	9.1%
	Reproducibility stnd dev, $s_R$	3.1	4.0	5.4
	Reproducibility, $R$	8.5	11.2	14.9
	Reproducibility ratio, % $R$	9.6%	9.2%	10.5%

A.2.2 The working group decided that the repeatability and reproducibility estimates for each of the materials are reasonable and reflect adequate estimates of the precision of the method. Furthermore, the working group decided that it is advantageous to list the estimates of precision for each material separately due to relatively small number of materials. Based on the trial information, data, and analysis considerations above, the statement in A.2.3 reflects a reasonable precision statement for this method.

A.2.3 The following estimate of precision is based on data from an interlaboratory trial conducted in 1999 using the TAPPI T 807 om-94 revision of the method. The trial involved nine laboratories and employed three samples of linerboard with weights of 35, 42, and 69 lb. The repeatability and reproducibility estimates below were derived from four test results for each sample each of which was the mean of 10 determinations.

#### Bursting strength of paperboard

<i>Sample</i>	<i>Repeatability ratio, %<math>r</math></i>	<i>Reproducibility ratio, %<math>R</math></i>
35-lb linerboard	8.4%	9.6%
42-lb linerboard	6.9%	9.2%
69-lb linerboard	9.1%	10.5%

The precision reported above is an estimate of the maximum difference to be expected in 19 of 20 instances, when comparing two test results for materials similar to those described above under similar test conditions. These estimates may not be valid for different materials or testing conditions; users of this method must determine the suitability of this estimate to their application.

#### Appendix A.3 Applications of repeatability and reproducibility



A.3.1 For these examples, the data for T 804 “Compression Test of Fiberboard Shipping Containers” are used. The standard indicates that the repeatability, based on five determinations, is 7.0% and the reproducibility is 10.6%. A typical value for this test result is 800 lbs.

A.3.2 *Repeatability: Is a laboratory giving good results?* Ten box samples were submitted to a single lab for testing. The first set of five has an average value of 800 lb, and the second set has a value of 850 lb. Is this difference consistent with the repeatability value shown in the standard? The standard gives a repeatability value of 7.0%, so in this case, one would expect the difference between the two values to be less than  $0.070 \times ((800 + 850) / 2) = 58$  lb. Because 50 lb is less than 58 lb, the laboratory is giving results with a variance similar to the laboratories who participated in the original testing of the method.

A.3.3 *Repeatability: Does a particular sample meet a specification?* A customer had set a specification of 800 lb. As part of a QA/QC program, five samples were submitted for testing, and a result of 780 lb was found. Should this shipment be rejected? Because repeatability assumes that one is comparing two test results and this case involves comparing one test result with a fixed value, the relevant limit is reduced by a factor of  $\text{SQRT}(2)$ . Specifically, the laboratories in the round robin study exhibited a variation of  $7\% / \text{SQRT}(2) = 4.95\%$ , so if the true value was 800 lb, one would expect the test results to be between 760 and 840 lb,  $(1 - 0.0495) \times 800 = 760$  lb and  $(1 + 0.0495) \times 800 = 840$  lb. Because the test result was within the range of expected values, the shipment should not be rejected.

A.3.4 *Reproducibility: Are two laboratories giving different results?* Two different laboratories are each given five box samples from the same production run. Lab A returned a value of 800 lb and Lab B returned a value of 950 lb. Is this difference consistent with the reproducibility value shown in the standard? The standard gives a reproducibility value of 10.6%, so in this case, one would expect the difference between the two values to be less than  $0.106 \times [(800 + 950)/2] = 93$  lb. Because the difference is 150 lb, the difference between these two results was more than what was observed during the standard testing, and closer inspection of the testing procedures at the two laboratories was required. In contrast, if Lab A returned a value of 800 lb, and Lab B returned a value of 850 lb, the difference was less than what has been observed during the standard testing, specifically  $0.106 \times [(800 + 850) / 2] = 87 > 50$ . One could conclude that these results were not significantly different.

A.3.5 *Reproducibility: Are three laboratories giving different results?* Three different laboratories are each given five box samples from the same production run. Lab A returned a value of 800 lb, Lab B returned a value of 900 lb, and Lab C returned a value of 950 lb. Were these differences consistent with the reproducibility value shown in the standard? The standard gives a reproducibility value of 10.6%, so in this case, one would expect the difference between values to be less than  $0.106 \times [(800 + 875 + 950) / 3] = 93$  lb. Reproducibility is defined for pairwise comparisons only. In this case there were three differences, Lab B – Lab A = 75 lb, Lab C – Lab B = 75 lb, and Lab C – Lab A = 150 lb. While two differences were below the limit of 93 lb, the third was above, so these data were not consistent with the standard, and further investigation was required.

*Your comments and suggestions on this procedure are earnestly requested and should be sent to the TAPPI Standards Department.*

