WCTE 2014 Quebec City Canada August 10-14
World Conference on Timber Engineering

MAINTENANCE PROCEDURES FOR NORTH AMERICAN VISUALLY-GRADED DIMENSION LUMBER DESIGN VALUES

David Kretschmann[1], Don DeVisser[2], Kevin Cheung[3], Bob Browder[4], Al Rozek[5]

ABSTRACT: ASTM International D1990 Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens, that governs the development of design values for dimension lumber in North America, was first adopted in 1991with recognition that the resource and manufacturing of lumber could change over time impacting design values. D1990 Section 13 instructed users of the standard to conduct a reassessment of property values derived by the practice if there is cause to believe that there has been a significant change in the raw material resource or product mix but no guidance was given for how to monitor or evaluate properties to determine if a reassessment of the design values developed according to D1990 is necessary. In 2014, guidance on design value maintenance has been included in D1990. This paper presents the multiple stage design value maintenance provisions and the thought process behind developing these provisions for monitoring, evaluation, and reassessment in D1990.

KEYWORDS: Visually-Graded, Dimension Lumber, Monitoring, Evaluation, Reassessment, Allowable Property Maintenance

1 INTRODUCTION

Dimension lumber visually graded in accordance with the National Grading Rule and assigned design values derived in accordance with the ASTM International D1990 *Standard Practice for Establishing Allowable Properties for Visually-graded Dimension Lumber from In-Grade Tests of Full-Size Specimens* (D1990) [1] has provided satisfactory performance in homes and other structural applications in North America for many years. In-grade tests are tests conducted on material selected to be in the size and grade in which their design values are claimed. Since D1990 was first adopted it recognized that there is a potential for the resource used to derive design values to change with time. Language was included in Section 13 of the original standard announcing the importance of reassessment of design values. No guidance, however, was given for how to monitor or evaluate properties to determine if a reassessment of the design values developed with D1990 is necessary.

No standard is a static document and changes and revisions to D1990, since its original approval in 1991, have been made to reflect the knowledge gained and the needs of the industry when deriving design values from in-grade full-size lumber testing programs. These include refinements to the Grade Quality Index (a measure of how lumber properties relate to clear wood) and standardizing the minimum requirements for monitoring, evaluation, and reassessment of lumber properties developed in accordance with D1990. This paper discusses the multiple stage design value maintenance provisions in D1990 for monitoring, evaluation, and reassessment for visually graded dimension lumber that have been adopted by ASTM International Committee D07 on Wood in February 2014 [2].

2 BACKGROUND

From the time that design values based on D1990 were first adopted in 1991 it was recognized that there was a potential that the resource used to change with time. D1990 Section 13 instructed users to conduct a reassessment of values derived by the practice if there is cause to believe that there has been a significant change in the raw material resource or product mix; but no guidance was given for how to monitor or evaluate properties to determine if a reassessment of the design values developed according to D1990 is necessary [1].

Derivation of in-grade design values are based on data from destructive testing of a matrix of bending, tension

[1] David E. Kretschmann, United States Department of Agriculture Forest Service Forest Products Laboratory (USDA FS FPL), 1 Gifford Pinchot Dr., Madison, WI 53726, USA. Email: dkretschmann@fs.fed.us
[2] Don DeVisser, West Coast Lumber Inspection Bureau, USA
[3] Kevin Cheung, Western Wood Products Association, USA
[4] Bob Browder, Southern Pine Inspection Bureau, USA
[5] Al Rozek, National Lumber Grades Authority, CA

parallel to grain and compression parallel to grain samples of commercially produced structural lumber. Samples are representative of the entire growth region of the species or species group population. This data, when processed following ASTM International standards, results in a "global characteristic value", which is a statistical estimate of an overall population property. Each size/grade sample was built from smaller samples of existing production or "production on the ground" [3]. Test cell data checks are applied to minimize the probability of developing non-conservative property estimates. This includes the use of lower tolerance limits and adjusting data further so that results are consistent with the test cell results.

Global characteristic values for bending stiffness are estimated at the mean level, while values for strength properties are determined at a "near minimum" value, or specifically the 5th percentile value. Data are adjusted to standardized conditions of temperature, moisture content and size to increase the sample size used to establish a 5th percentile lower tail estimate.

The following decisions made about sampling during the development of the original in-grade program also influenced the establishment of procedures for maintaining in-grade design values.

- The in-grade sample was a single cross-sectional sample which was maximally distributed throughout the producing regions by using small sampling lots.[6]
- An attempt was made to capture normal temporal variation in lumber properties (though the use of maximally distributed sampling).
- By adjusting all data to a standard size, the effective sample size was increased for improving the precision of the estimates of 5th percentile properties.[7]
- The sample size needed to estimate the 5th percentile values with a given statistical precision far exceeds that needed for estimating the mean bending stiffness values.

Methods for monitoring design values have been discussed since the 1990s [4]. The first resource monitoring program was initiated by Southern Pine Inspection Bureau (SPIB) in 1994. The SPIB resource monitoring program has resulted in the first full matrix reassessment of design values using D1990 which was implemented in June 2013. Over time other North American grading agencies responsible for visually-graded lumber have undertaken monitoring efforts. A consensus approach to maintenance of design values which specified the minimum requirements for monitoring, evaluation, and reassessment of lumber properties developed in accordance with D1990 was finally adopted in February 2014. The next section presents highlights of the discussions that resulted in the revision of D1990.

3 DEVELOPMENT OF PROGRAM FOR MAINTENANCE LUMBER DESIGN VALUES

3.1 OVERVIEW

It was a long process to get consensus on an approach for maintenance of visually-graded structural lumber design values derived in accordance with D1990. The initial work on design value maintenance was begun by the American Lumber Standard Committee (ALSC) within its Lumber Properties Task Group (LPTG) on October 12, 2010. This group worked for 1½ years to develop a multi-stage approach to maintenance of design values. The multi-stage approach involved stages that monitored one size-grade, evaluated additional sizes, and then if necessary reassessed the claimed design values for all grades and sizes. This multi-stage approach was brought to ASTM International and the first ballot for the revision of Section 13 of D1990 was issued in January 2012. The initial returns for this ballot affirmed that the multi-stage concept suggested by the ALSC LPTG was a suitable approach but two schools of thought on the approach for judging when a monitored property had changed were present.

One approach preferred to look at the monitored test data collected for a given size-grade test cell and compare this directly to the design value that is claimed for the tested size-grade cell. The other approach preferred to compare new size-grade cell information directly to previous size-grade cell information used to determine the currently claimed design values. A number of simulations were conducted to help facilitate the decision making process between the two approaches.

The discussions of the pros and cons of the two approaches revealed that the various options for comparison between old and new cell data from the monitoring portion of a maintenance program needed to be investigated further. An effort was initiated to determine what methods were practical for cell comparisons. The discussions also generated a list of characteristics that were important to a maintenance program. Table 1 summarizes characteristics that should be sought from each maintenance program, keeping in mind that the overall program should be both cost effective and practical.

[6] Ad hoc studies in 1978-80 examining the variability in lumber properties within a mill and between mills suggest that the two are comparable. This suggests that some (if not all) of the variation over time can be replicated by involving more mills in the sample.

[7] This assumes that there are no issues or bias with the size adjustment model.

Table 1: *Basic Components of Maintenance Program*

Element	Desirable/Required Characteristics
Sampling	Is representative of the underlying population Provides insight into the untested sizes or grades Collects enough information to make judgment on causes of cautionary variations
Tests	Are repeatable and reproducible and can directly be compared back to the original test data Can provide insight into untested cells or properties
Analysis	Will differentiate between basic and cautionary variations Has low error rate (minimize false positives, false negatives, and practical differences)
Identification and Prioritization	Will identify and prioritize potential causes for cautionary variation

Some latitude can be accommodated in the make-up of the characteristics of the elements depending on the objectives of the program. The elements can be selected to optimize how technical resources are deployed to minimize the overall costs but still meet the objective of providing evidence that the design value assignments are appropriate. The choices have a bearing on the amount of material to be tested and the efficiency at which the change of interest can be detected. The experience of agency personnel based on past experience, available resources, and available data should be taken advantage of in designing the program.[8] Where there needs to be a high degree of standardization, the elements may be prescribed such as in the D1990 standard.

3.2 OBSERVATIONS FROM INITIAL SIMULATIONS USING IN-GRADE DATA

Many simulations using existing in-grade test data have been conducted over the last several years. These simulations have provided insight into the relationship between properties, the role sample size, and statistical significance testing method plays in a monitoring program's ability to detect differences.

3.2.1 Relationships between Properties and Sizes

Most strength and stiffness properties are correlated, allowing one property to be estimated from another. Similarly, with an understanding of size effects, a change in the property of one size may be used to infer a change in the other sizes. Because of this, it may not be necessary to monitor all cells and sizes for a given species on an ongoing basis. On the other hand, one should be aware of how a unit change in one cell or property translates into change in another cell or property. Also the consensus method for determining design values for grades not tested once values for the anchor grades (such as No. 2 and SS) are established can be used to get some indication of the changes of other cells or properties.

For example, there is considerable variability about the MOE-MOR regression relationship for lumber but it is well-established that the correlation is positive [4]. For sizes and grades considered in this work, one rounding rule in MOR 0.344 MPa (50 psi) typically corresponds to, on average, a change in MOE of approximately 0.14 GPa (0.02 x 10^6 psi) and one rounding rule in MOE 0.69 GPa (0.1 x 10^6 psi) corresponds to, on average, about four rounding rule 1.376 MPa (200 psi) changes in MOR [5].

3.2.2 Role of Sample Size

The effect of sample size on the percentage of false positive readings, for one rounding rule in MOE and MOR and three targeted shifts in MOE 0.69, 0.17, 0.14 GPa (0.1, 0.025, 0.02 x 10^6 psi), were investigated using 100,000 simulated samples for each case [5]. The results of these simulations demonstrate the importance of sampling method to the sample size required. If an agency were dependent on the information gathered from just one year's sample (i.e. one step) then the sample size required to reduce the chances of detecting false positive below 1% for a change in MOE of 1.7 GPa (0.25 x 10^6 psi) would be in most cases well over 1000 specimens. This same level of confidence can be reached with a little more than 200 specimens after three separate samples (three steps) have been taken. For a random sample size, a repeated sampling (a multi-step approach) helps to ensure that a detected shift isn't a result of the natural variability of the random sample by reducing the chances of falsely detecting a shift. The original in-grade size/grade sample size of 413 pieces is very unlikely to give a false positive indication of a shift for MOE in a multi-step process. The sample size simulation results suggested that detection with three steps and a sample size of 200 is roughly equivalent to a two-step procedure with a sample size of 413. Therefore, a smaller sample size could be used with more steps to detect a targeted shift. However, to get a

[8] For example, in the SPIB resource monitoring program, SPIB chose to monitor No.2 2 by 4.

good representation of all the geographic regions, a sample size of at least 360 should be considered. Also it is worth noting that if a shift in MOE of one rounding rule is observed, a shift of several rounding rules for MOR may have occurred.

3.2.3 Testing for Statistically Significance Changes

Monitoring programs are essentially based on null hypothesis testing. It all starts with a claimed property that has been established by testing. The monitoring program sets up a testing procedure where we assume there has been no change in the property (the null hypothesis H_0: the claimed value is true or higher) and we look at either a statistical or practical difference or see if it indicates that that hypothesis should be rejected (H_A: the value is less than claimed). These tests are run at decided upon significance levels α or at agreed upon practical differences that indicate a practical need to change the claimed values. It was quite difficult to get agreement on the acceptable practical difference. Therefore, an α of 0.05 (the commonly used significance level utilized for wood properties) was selected for the significance level for our basis of comparison.

3.2.4 Power Curves Based on Statistical Difference

When looking for a statistical difference, in any one monitoring sample, an upper confidence limit is established to determine if this value is statistically below the claimed value. Power curves have been developed for simulations to illustrate the likelihood of picking up statistical significant differences. Power is the probability that if you are looking to find a certain difference in treatments you will be able to pick the difference up, if it exists, from the study. Power is expressed in percent. A power of 80% means that if the difference you are trying to pick up exists, in an infinite number of studies you will achieve statistical significance 80% of the time. The higher the power the more certain you will pick up the difference you are interested in if it exists.

Different tests are run for mean and fifth percentile properties. For mean properties the upper confidence level of the mean is estimated by Equation (1):

$$\bar{x} + z_{\alpha}\frac{s}{\sqrt{n}} \qquad (1)$$

where: x-bar is the average, Z_α is the standard normal variate dependent on confidence level α, S is standard deviation, and *n* is sample size.

For 5th percentile estimates of MOR the nonparametric percentile estimate for upper bound of the 5th percentile for large samples is calculated by Equation (2):

$$p_{0.05} + z_{\alpha}\sqrt{\frac{p_{0.05}(1-p_{0.05})}{n}} \qquad (2)$$

where: Z_α is the standard normal variate dependent on the confidence level α, $p_{0.05}$ is the fraction of the population that falls below the 5th percentile (0.05) and *n* is sample size [6]. The rank is then obtained by multiplying Equation (2) by the sample size *n*.

The values of MOR that correspond to this upper bound can be found by ranking the MOR values using ASTM D2915 [7]. In D2915, the ordered test values in ascending order are denoted as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. The order statistics are ranked test values from the lowest to the highest. For example, the first order statistic, $x_{(1)}$, is the lowest test value or the weakest piece in the sample, the second order statistic, $x_{(2)}$, is the second weakest, etc. Beginning with the lowest value (the 1st order statistic) we calculate $p_i = \frac{i}{n+1}$. This value is plotted on the y-axis with $x_{(i)}$ plotted on the x-axis and the confidence bound on the 5th percentile is the MOR value found through interpolation that corresponds to the percentile on the y-axis.

The results of one of the power curve simulations for MOE are shown in Figure 1. These simulations were based on determining, for 10,000 repetitions of random sampling from a large data set which had been adjusted to a set change in MOE 0.17, 0.34, 0.52, 0.69, 0.86, 1.0 GPa (0.025, 0.05, 0.075, 0.1, 0.125, 0.15 x 10^6 psi) for a given sample size (e. g. 120, 180, 240, etc.).

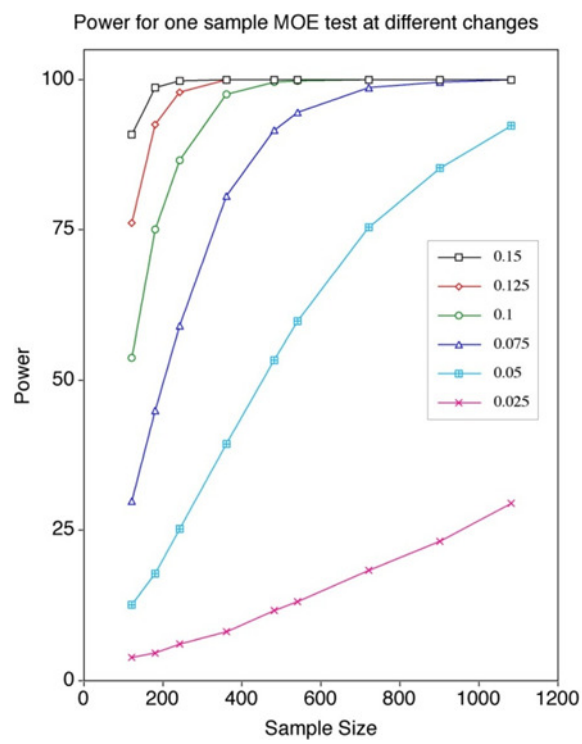

Figure 1: *Power curve examples. The different curves are shown for increasing increments of MOE (simulations run by Dr. James W. Evans USDA FS FPL shown for increments of 10^6 psi)*

For a given sample size and adjusted change, the percentage of the time the procedure shows a statistically significant change was recorded. In Figure 1 you can see that as the sample size increases the power increases. Also, as the difference you are trying to detect increases

the power dramatically increases at smaller and smaller samples.

3.2.5 Frequency Curves Based on Practical Differences

When checking for a change based on a practical difference, a sample's property is checked to see if the property is more than the practical difference below the claimed value. In a practical difference approach, frequency curves are used to determine the effectiveness of picking up changes that are greater than the practical claim. Like power curves, frequency curves count the percent of time something has occurred when a specified change has occurred in the treatments. In frequency curves, however, the frequency of a specified change is counted when a known change in the large population has occurred. In power curves, a change of statistical significance rather than an event above a threshold is counted when the known change has occurred. For a given sample size and forced change by different amounts, we record the percentage of times the procedure shows a practical change.

Figure 2 shows the percent of times the one random sample test for MOE shows a difference of 0.34 GPa (0.05 x 10^6 psi) when a large sample of MOE has actually been changed by a set amount 0.17, 0.34, 0.52, 0.69, 0.86, 1.0 GPa (0.025, 0.05, 0.075, 0.1, 0.125, 0.15 x 10^6 psi) for 10,000 simulations. In this simulation example, if you force a shift of 0.34 GPa (0.05x 10^6 psi) in the large population sample you are likely to see this shift roughly 50% of the time if you are looking of a 0.34 GPa (0.05 x10^6 psi) change for all sample sizes. For a sample size of 400 the number of times a change of 0.52 GPa (0.075 x 10^6 psi) in the data is detected when trying to detect a 0.34 GPa (0.05 x10^6 psi) difference is over 80 percent.

As the sample size increases you can do a much better job of picking up changes. From this work it is clear that a practical level should be selected that is below the change you really are concerned about.

The simulations conducted focused on and revealed considerable information about the power and expected frequency of detection of practical differences. This work was most pertinent to the design value comparison approach but it also provided insight for the cell comparison approach. In the end, however, after considerable discussion, debate and repeated balloting, the cell comparison approach was selected as the consensus approach for D1990.

3.3 TEST CELL COMPARISION METHODS CONSIDERED

A task group was setup to consider what option(s) were preferred for the cell comparison. Dr. Steve Verrill, Mathematical Statistician with the USDA FS FPL, conducted a number of simulations using the various options that had been proposed by the task group. Several methods, such as targeted action thresholds based on a one-sided lower 95% confidence bound on the 5th percentile calculated from the baseline data, nonparametric

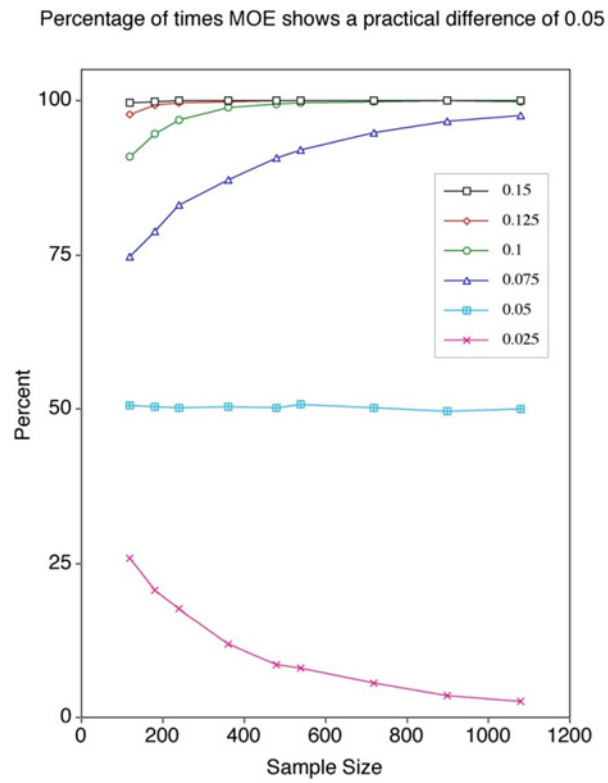

Figure 2: *Percentage of time Southern Pine MOE shows a practical difference of 0.05 x10^6 psi for a known mean difference in MOE (simulations run by Dr. James W. Evans USDA FS FPL shown for increments of 10^6 psi)*

Analysis of Variance (ANOVA/Tukey multiple comparison test), χ^2, Kolmogorov-Smirnov, and Wilcoxon, comparison tests for monitoring design values were considered as ways to monitor for resource change.

The results of the initial simulations with the various techniques were shared with the task group. These results suggested that the Wilcoxon nonparametric comparison technique was the preferred method for cell comparisons. The task group agreed that the Wilcoxon method would be applied to both the mean MOE and lower tail strength properties to determine when a significant change has occurred and a consensus of the voting members in ASTM International agreed.

During the discussion of the cell comparison evaluation methods two questions were raised requiring additional simulations, (1) the effect that smaller sample size might have on detected differences and (2) the impact of correlation of material within a mill. The Section task group reviewed the additional simulations and still concluded that the Wilcoxon method with an α level of 0.05 would provide adequate detection of a potential change in resource. The members of ASTM Committee D07 on Wood agreed and adopted a maintenance program that checks the cell data against the corresponding cell data

that is currently being used to establish design values for the species of concern.

4 PERIODIC CORROBORATION OF LUMBER DESIGN VALUE

The periodic corroboration of assigned design values program adopted for maintenance of lumber design values is divided into three stages: Stage 1) a monitoring program, with potentially two steps, to periodically check for changes in product performance; Stage 2) an evaluation program, upon detection of a statistically significant downward shift, to evaluate monitoring data and confirm effectiveness of remedial actions; and Stage 3) a reassessment program to re-establish design values. These stages rely on the application of a nonparametric Wilcoxon statistical test to delineate the states where on-going monitoring can continue, and where actions need to be taken to change the lumber sorting or grading process and/or re-establish the design values. A flow chart of a multiple stage periodic verification approach is presented in Figure 3. The sampling and testing as well as the results are closely monitored by the authority having jurisdiction (AHJ) over design values which is the ALSC.

4.1 MONITORING

Monitoring (Stage 1) is the on-going portion of the program. Its role is to determine if there is sound evidence to believe that there has been a change in the product performance sufficient to justify a further evaluation or a complete reassessment of current design values. A monitoring program can potentially have two steps. Figure 4 shows a flow chart for a monitoring program.

A nonparametric Wilcoxon test is used to determine whether the action level is reached and whether to proceed to Step 2 (an additional destructive sampling of a size-grade cell) of Stage 1. The action level is reached when a comparison of the cell property that was used to determine the current cell value is significantly different from the monitored cell value at an α level of 0.05. If the action level for a downward shift in Step 1 of Stage 1 is not reached, the original periodic testing shall be reinitiated. If the action level for a downward shift in Step 1 of Stage 1 is reached, then either another round of destructive testing, repeating the size and grade of material tested in Step 1 of Stage 1 is conducted in Step 2, or an evaluation of the currently claimed design values is started.

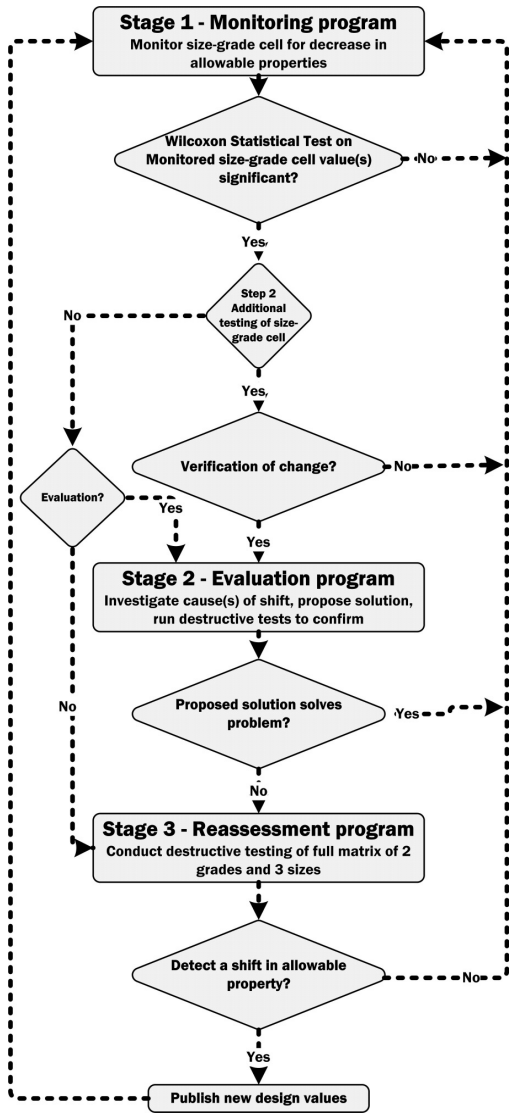

Figure 3: *Flow chart showing 3 stages for design value maintenance program*

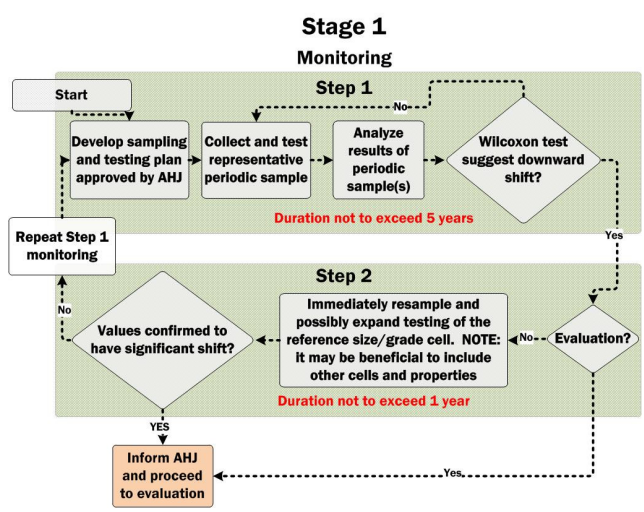

Figure 4: *Details of Stage 1 Monitoring (AHJ is authority having jurisdiction)*

Some important features of the Stage 1 monitoring program are:

- Appropriate sampling procedures, a sample size, and test methods to accomplish the objectives must be contained in a sampling and testing plan which has been approved by the AHJ.
- A Wilcoxon test with an α level of 0.05 is to be used to determine whether to proceed Step 2 of Stage 1.
- For major commercial species or species groups destructive testing of a size-grade cell must occur at least once every five years.
- The results of monitoring programs are to be submitted to the AHJ.
- If an action level is reached for a property, Step 2 of Stage 1 or the evaluation Stage 2 must be completed within one year.

4.2 EVALUATION

Evaluation (Stage 2) is initiated after a statistically significant downward shift in a monitored cell has been confirmed. Alternatively, if the shift detected in the monitoring stage is considered large enough the evaluation stage can be skipped and the reassessment (Stage 3) can be initiated. Figure 5 depicts a flow chart for Stage 2.

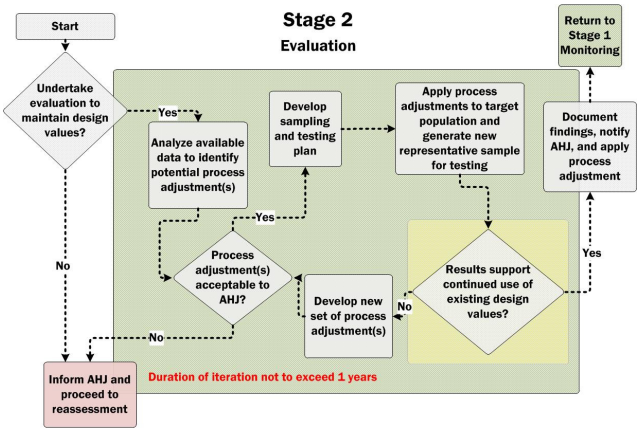

Figure 5: *Details of stage 2 Evaluation (AHJ is authority having jurisdiction)*

The evaluation stage is used to identify cause and scope of change and develop process adjustments that would support the continued use of current design value(s). Acceptable responses include altering the description of the visual grade, changing the method of processing, or restricting the resource that can be processed. Confirmation testing must be done to confirm the response brings the design values back within acceptable range of the published design values for all affected size-grades and properties. The result of the evaluation stage is either a change to the process which will subsequently need to be confirmed as sufficient to allow the current design values, or a need to reassess and establish new design values.

4.3 REASSESSMENT

Reassessment (Stage 3) is initiated if evaluation (Stage 2) is either not selected or is not successful after confirmation of a significant shift in Stage 1. Figure 6 shows a flow chart of Stage 3.

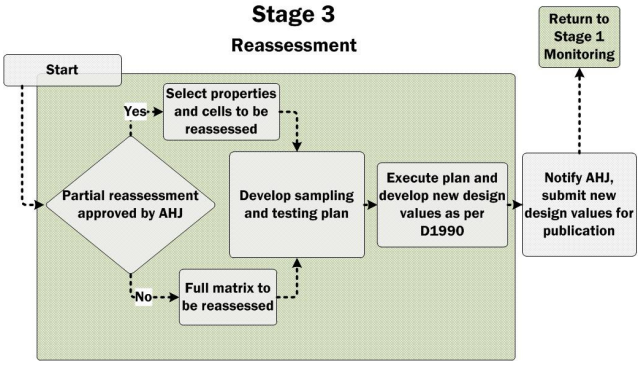

Figure 6: *Details of Stage 3 Reassessment (AHJ is authority having jurisdiction)*

A reassessment of design values shall be conducted if there is cause to believe that there has been a significant change in the raw material resource or product mix detected by the monitoring which has been unresolved by evaluation. This reassessment shall be conducted using the sampling matrix upon which the current design values are based with an awareness of changing production conditions. The reassessment may result in establishing new design values.

5 CONCLUDING REMARKS

Further guidance on design value maintenance in D1990 was needed. Many options were considered for periodic corroboration of design values. It was decided, through a consensus process, that multiple stage design value maintenance provisions be used for monitoring, evaluation, and reassessment of North American visually-graded structural lumber.

ACKNOWLEDGEMENT

The efforts of the American Lumber Standard Committee lumber properties task group in initial development of a draft maintenance criterion are acknowledged.

REFERENCES

[1] ASTM International: D1990-91 Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens. Annual Book of Standards, Section Four Construction, Vol. 04.10 Wood. West Conshohocken PA. 1991.

[2] ASTM International: D1990-14 Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens. Annual Book of Standards, Section Four Construction, Vol. 04.10 Wood. West Conshohocken PA. 2014.

[3] D. W. Green, B. E. Shelley, and H. P. Vokey. In-Grade Testing of Structural Lumber. FPS Proceedings 47363, Forest Products Society. Madison, WI. 110 pgs. 1989.
[4] D. W. Green and D. E. Kretschmann. Lumber Property Relationships for engineering Design Standards. Wood and Fiber Science, 23(3), 1991, pp. 436-456. 1991.
[5] D. E. Kretschmann, J. W. Evans, and L. Brown. Monitoring of Visually Graded Structural Lumber. *Research paper FPL-RP-576.* Madison, WI: U. S. Department of Agriculture, Forest Service, Forest Products Laboratory. Madison, WI 18p. 1999.
[6] W. J. Conover. Practical nonparametric statistics. Wiley series in probability and statistics: applied probability and statistics section. 3rd. ed. 1999.
[7] ASTM International: D 2915-10. Standard Practice for Sampling and Data-Analysis for Structural Wood and Wood-based Products. Annual Book of Standards, Section Four Construction, Vol. 04.10 Wood. West Conshohocken PA. 2013.

In: Proceedings, 2014 World Conference on Timber Engineering. Quebec City, Canada, August 10-14, 2014. 8 p.