United States
Department of
Agriculture

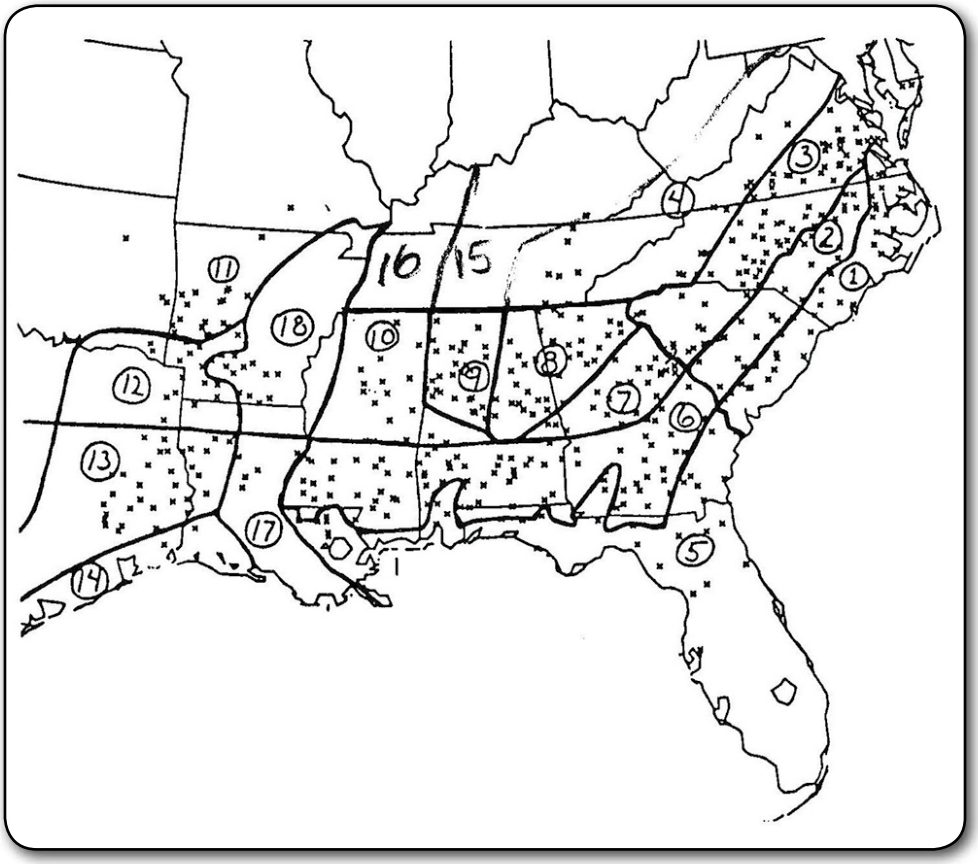
Forest Service

Forest
Products
Laboratory

General
Technical
Report
FPL–GTR–221

Analysis of the Part I Southern Pine In-Grade Program Data

James W. Evans
David E. Kretschmann
Cherilyn A. Hatfield
David W. Green



Abstract

It has been suggested that within- and between-mill variability of samples may have changed since the original North American In-Grade Program and a new sampling method may be required. The cooperative research "In-Grade Program" was conducted in the late 1970s to the late 1980s to establish allowable properties (modulus of elasticity (MOE), allowable bending stress (F_b), allowable compressive stress parallel to grain (F_c), and allowable tensile stress parallel to grain (F_t)) for dimension lumber based on full-size testing of graded lumber. The In-Grade Program was divided into three major administrative parts, each with specific objectives. Part I estimated bending strength and stiffness of 2 by 8 No. 2 and 2 by 4 STUD grades of Douglas-fir and Southern Pine. Part II considered other grades, sizes, and methods of loading (such as tension and compression parallel to the grain) for Douglas-fir and Southern Pine. Part III considered other species. The purpose of this document is to explain how Part I of the In-Grade Program, which looked at regional and mill variability, was designed and to re-examine the analysis that was conducted in 1980 to determine sampling methods. The present report discusses Southern Pine data, and sampling design issues related to regions, mills, and pieces sampled in a mill are presented. Also, other techniques not applied in 1980 are used to reanalyze the test data. The random sampling of mills within region based on production was based on a balancing of the significance of regional, mill-to-mill, and within-mill variability. We feel that the decisions that were made in setting up the stratified cluster sampling methods to minimize variability in sampling for Part II and Part III were justified and necessary given the limited resources available to conduct Part II and Part III of the program.

Keywords: sampling, In-Grade Testing, Southern Pine, Part I, variance component analysis

Contents

April 2013

Evans, James W.; Kretschmann, David E.; Hatfield, Cherilyn A.; Green, David W. 2013. Analysis of the part I southern pine in-grade program data. General Technical Report FPL-GTR-221. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 11 p.

A limited number of free copies of this publication are available to the public from the Forest Products Laboratory, One Gifford Pinchot Drive, Madison, WI 53726–2398. This publication is also available online at www.fpl.fs.fed.us. Laboratory publications are sent to hundreds of libraries in the United States and elsewhere.

The Forest Products Laboratory is maintained in cooperation with the University of Wisconsin.

Analysis of the Part I Southern Pine In-Grade Program Data

James W. Evans, Ph. D., Research Mathematical Statistician
David E. Kretschmann, PE, Research General Engineer
Cherilyn A. Hatfield, Statistician
David W. Green, Ph. D. (retired)
Forest Products Laboratory, Madison, Wisconsin

Executive Summary

Recently, concerns have been raised about the effect that changes in the lumber mill type (small log or large log mills) may have on deciding what stratified cluster sampling method to use for establishing a "global number." A global number is one number that covers the entire growth range for Southern Pine dimension lumber for each allowable property. It has been suggested that within- and between-mill variability of samples may have changed since the original North American In-Grade Program and a new sampling method may be required. This has led us to a re-examination of the methodology used to establish the sampling method that is currently used to gather data for the development of a global number for visually graded dimension lumber.

The cooperative research "In-Grade Program" was conducted in the late 1970s to the late 1980s to establish allowable properties (modulus of elasticity (MOE), allowable bending stress (F_b), allowable compressive stress parallel to grain (F_c), and allowable tensile stress parallel to grain (F_t)) for dimension lumber based on full-size testing of graded lumber. The In-Grade Program was divided into three major administrative parts, each with specific objectives. Part I estimated bending strength and stiffness of 2 by 8 No. 2 and 2 by 4 STUD grades of Douglas-fir and Southern Pine. Part II considered other grades, sizes, and methods of loading (such as tension and compression parallel to the grain) for Douglas-fir and Southern Pine. Part III considered other species. This report documents how the information gathered in Part I of the In-Grade Program, which looked at regional and mill variability, was designed and how the statistical analysis that was conducted by James Haskell in 1980 was used to determine sampling methods for Part II and Part III of the program. This report discusses Southern Pine data, and sampling design issues related to regions, mills, and pieces sampled in a mill are presented. Techniques not used by Haskell but commonly available at the time are also used to reanalyze the Southern Pine test data.

The re-examination of the past analysis and reanalysis of the 1980 data suggest that both the regions and the mills within regions were statistically significant in accounting for variation in lumber properties at the 0.05 level. Mill size based on production, which relates to mill type, was also a parameter investigated and accounted for in the Part I study. We also determined that a particular lot was highly significant in accounting for variation in MOE, suggesting that it would be beneficial to continue sampling by lots as opposed, say, to sampling more pieces from a single pack of lumber. The establishment of the sampling method that is currently used in determining global numbers for the major species groups in the North America took these statistical significances (for regions, mills, and lots) along with logistical and cost factors into account. The random sampling of mills within region based on production was based on a balancing of the significance of regional, mill-to-mill, and within-mill variability. We feel that the decisions that were made in setting up the stratified cluster sampling methods to minimize variability in sampling for Part II and Part III were justified and necessary given the limited resources available to conduct Part II and Part III of the program.

Introduction

As a result of destructive testing of 2 by 4 No. 2 lumber in bending and tension parallel to grain (ASTM 2011c), the Southern Pine Inspection Bureau (SPIB) in 2011 proposed an interim drop in dimension lumber values (American Forest and Paper Association (AFPA) National Design Specifications (NDS)) for Southern Pine by approximately 30%. Large parts of the lumber industry are concerned about the disruption and economic effect that these changes would have and have challenged the sampling method used by SPIB to conduct their follow up bending, tension, and compression testing of three size 2 grades (ASTM 2011 a,b). Using procedures similar to those developed and used in the U.S. In-Grade Program (FPRS 1989), SPIB used a stratified sampling procedure. The growth region for Southern Pine was divided into 18 regions that were believed to be relatively homogeneous regions based on previous research. Specimens were sampled from each region in proportion to production for the region by randomly sampling mills in the region. In each mill, a limit of two lots was sampled with 10 on-grade pieces taken sequentially from the lot.

A recent challenge to this sampling methodology has raised the issues of whether it would be better to (1) randomly

sample mills generally or to first stratify by regions or (2) (in reducing variation due to sampling) sample fewer mills with more specimens per mill. It is being advocated to repeat the preliminary work that was done in the U.S. In-Grade Program that led to the adoption of the current sampling method. Before this preliminary work is repeated, the authors feel it is important to review the work done in Part I of the In-Grade Testing Program used to establish the current sampling methodology.

This paper has four main parts. First, background will be provided on the reasons for initiating the In-Grade Program in the United States. Second, the preliminary work, including sampling method and analysis of data that led to the method used today, will be discussed. Third, alternatives to the original analysis commonly available at the time of the preliminary work will be explored. Fourth, an evaluation of the whole sampling method used will be presented. The intent is to provide useful guidance to anyone hoping to repeat a similar study.

Because this preliminary work (Haskell 1980) on sampling was never made widely available, it is desirable to recreate it and look at it under the present day concerns. This means reviewing the sampling method and analysis used, reassessing the conclusions drawn from the data, and looking at additional ways that could be used to analyze the preliminary data. This document contains a very thorough summary of the sampling portion of the unpublished Haskell report and a description of the proper interpretation of the results as it relates to current Southern Pine sampling. We have only evaluated the regional and mill size analysis sections of Haskell's unpublished report thoroughly but have not investigated the analysis covered in the rest of the sections. We therefore do not address the analysis conducted or any errors or inconsistencies that may exist in that portion of the report.

The current challenge involves Southern Pine dimension lumber; therefore, only the Southern Pine data from the In-Grade Program will be evaluated. Presently, other commercial species are being looked at to determine if the published design values are appropriate, as those species may be facing these same issues in the near future.

Background on Initiation of the In-Grade Program in the United States

Prior to 1980, a concern arose that despite excellent in-service performance of wood-frame houses, some single pieces of full-size, visually graded structural may have strength values less than those published for use in design (Bodig 1977; Madsen 1978; Galligan and Green 1980; Galligan and others 1980). However, tests of full-scale floor and wall components and of a full-scale house designed by current accepted methods, lumber grades, and design stresses indicated that houses were generally over-designed (Goodman and others 1974; Polensek and Atherton 1976; Tuomi and McCutcheon 1975), resulting in an anomaly that needed to be resolved if wood was to be used in the most efficient manner.

In 1977, a comprehensive research program was organized by the Forest Products Laboratory (FPL) to develop and implement light-frame construction research and to transfer the related technical information to designers and builders (Hans and others 1977). A by-product of this program was to look at material property research.

In May 1977, a meeting of U.S. and Canadian rules-writing agencies was held. This meeting resulted in a request that a detailed plan be formulated for a more extensive study of the mechanical properties of lumber that related to design methods and use. Subsequent discussions with the U.S. Forest Service resulted in an offer by FPL to prepare an overview of research needs and to design a plan that could be considered for country-wide application.

In July 1977, representatives of four U.S. rules-writing agencies (Redwood Inspection Service (RIS), Southern Pine Inspection Bureau (SPIB), West Coast Lumber Inspection Bureau (WCLIB), and Western Wood Products Association (WWPA)), met to review a "Five-Year Testing Program" and a rough draft of a lumber sampling method, both prepared by FPL. The agencies agreed that the FPL sampling method had considerable merit but that some modifications were necessary in order to devise a workable schedule based on knowledge of mill production and similar considerations. A cooperative research agreement was reached between FPL and three of the agencies (SPIB, WCLIB, and WWPA).

This cooperative research agreement became known as the "In-Grade Program." It was designed to be divided into three major administrative parts, each with specific objectives. Part I looked at bending strength and stiffness of 2 by 8 No. 2 and 2 by 4 STUD grades of Douglas-fir and Southern Pine. Part II was to consider other grades, sizes, and methods of loading (such as tension and compression parallel to the grain) for Douglas-fir and Southern Pine. Part III was to consider other species.

Part I of the In-Grade Program

Part I attempted to meet three major objectives:

1. Characterize the bending strength and stiffness of 2 by 8 No. 2 and 2 by 4 STUD grades of Douglas-fir and Southern Pine lumber determined to be "on-grade" by agency quality supervisors.
2. Characterize the bending strength and stiffness of lumber "as-graded" by the mill grader.
3. Characterize wall and floor performance by modeling these light-frame components with the results of (1) and (2).

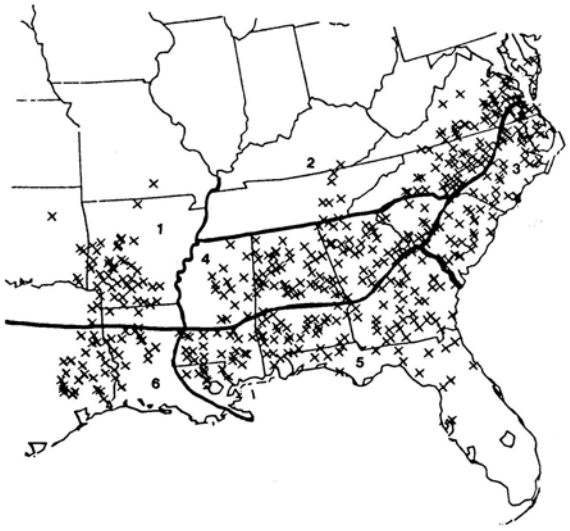


Figure 1. Original six regions for Part I.

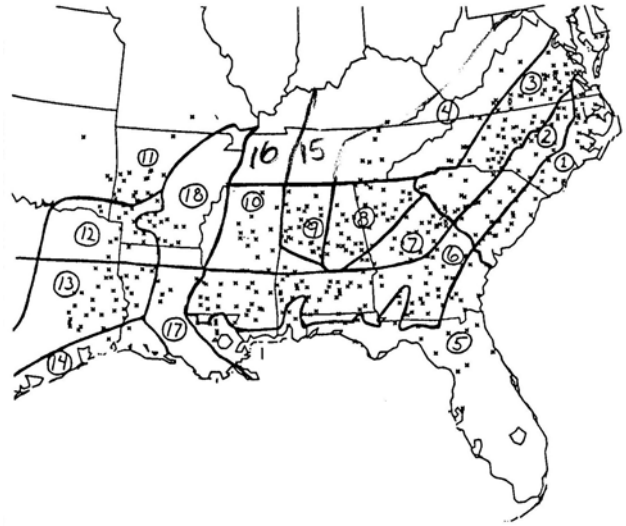


Figure 2. The current Southern Pine growth region boundaries.

A secondary objective was to provide information that would help in the design and analysis of the other parts of the In-Grade Program. Part I also looked at issues that included how we effectively and efficiently look at the population of material in a particular grade and size that is actually being produced at the time. This meant an emphasis on sampling material that reflected a "production" population and not a "forest" population. This means that conclusions from this research relate to the properties of the mill produced product, not the properties of the standing timber.

Sampling Method for Part I

The first step in developing the sampling method of Part I was to compile a list of all the "potential" mills producing 2 by 8 No. 2 or 2 by 4 STUD. These mills were then subdivided into 6 strata (regions) based on growth patterns and geographical boundaries as is shown in Figure 1 for the 2 by 8 No. 2 Southern Pine data. In this way, the usual statistical philosophy of homogeneity of properties within a stratum (region) was approached. The result of this method was a stratified sampling method for mills. Note that in Part II the regions were further subdivided to produce 18 regions (Fig. 2, Jones 1989).

The next important question that needed to be answered was "How to best sample material at each mill so that it represents properties of lumber furnished to the designer/consumer?" It was decided to sample the material at mills either in lots (a stack of lumber bound together) from existing inventories or if no inventory was present, directly as it was produced by the mill. A lot became known as a sample from the mill.

At a sampled mill, 10-piece serial lots were selected for testing. Serial lots meant that once the outer layers of a pack of lumber were removed, pieces of lumber were selected for a given layer in the order in which they were stacked until 10 pieces were obtained. The outermost two layers were excluded because of possible damage from banding the pieces together in the stack. Lots relate to the manner in which lumber is produced and distributed. Ten was the number of pieces chosen as "typical" for historical wall and floor assemblies. This serial selection method was used to approximate a "lot," as what might be used by a carpenter in building a conventional wall or floor system. "After the first 10 pieces were selected, the lumber was re-graded by an agency quality supervisor and, if the re-grade did not agree with the desired structural grade, additional pieces were taken to make a total of 10 'on-grade' pieces available from each bundle" (Green 1983).

At a mill with an existing inventory, 10 pieces (a 10-piece lot) were taken in sequence from a pack. At a mill with no inventory, a 10-piece lot consisted of 10 successive pieces produced. Since observations from successive pieces from the sample mill were expected to have a higher correlation among themselves than randomly selected pieces, subsequent lots taken from the sample mill would each be taken from a separate bundle of lumber or after production of a number of intermediate specimens. If the mill inventory was large enough and there was a way to identify the production date, lots were to be sampled to span the production period. Statistically, this means we were now using a stratified cluster sampling method.

Part I work was broken into two phases called A and B. In some ways, Phase A was a trial for further work in Phase B and the two phases would be combined for the final analysis. Part I as a whole looked at several issues with either the 2 by 8 No. 2 data or the 2 by 4 STUD data or both. One issue explored was the comparison of "on-grade" samples and "as-graded" samples. Other issues considered included the effect of mill size on the 2 by 8 No. 2 data, the effect of mill type on the 2 by 4 STUD data, and the effect of

Table 1. Sources of Southern Pine lumber

Type of lumber	Region						
2 by 8, No, 2, S-Dry	1	2	3	4	5	6	Total
Number of mills producing this lumber	50	82	100	128	143	47	550[a]
Number of mills sampled in Phase A	2	0	2	2	2	2	10
Number of mills sampled in Phase B	4	6	4	4	4	4	26
Number of mills sampled in Phase A & B	6	6	6	6	6	6	36

[a]The 1981 report has 556 but the regional total based on our current data set adds to 550.

Table 2. Distribution of Part I, 2 by 8 No. 2 sample by mill size

Region	Small	Medium	Large	Total	Mills
1	30	59	60		6
	30	30		239	
		30			
2	30			180	6
	30				
	30				
	30				
	30				
	30				
3	59	30	60	239	6
	30	30	30		
4	30	30	59	240	6
		60	30		
		31			
5	31	30	60	241	6
	60	60	30		
		31			
6	18	60	60	228	6
	30	30	30		

kiln drying using a conventional schedule rather than the high-temperature schedule on both 2 by 8 No. 2 and 2 by 4 STUD Southern Pine. All work except the kiln-drying study was done separately for Southern Pine and Douglas-fir data. For Southern Pine 2 by 4 STUD, the interaction of mill type and kiln type was also studied. The focus of this paper will be those items that influenced the method of sampling.

Other Research Studies Conducted in Parallel with Part I Work

A great deal of other research and work was done on issues that were needed in order to accomplish the In-Grade Program. A brief listing of some of this work includes major deviations from normal laboratory testing procedures such as (1) faster rate of loading, (2) random location of the load with respect to the length of the piece, (3) random placement of the "worst edge" with respect to the direction of loading, (4) shorter span-to-depth ratio and (5) non-equilibrium moisture contents at time of test. This and other work have been well documented in published sources and often incorporated into ASTM standards (Green and Evans 1988; FPRS 1989).

Analysis of In-Grade Program Part I Data

As mentioned in the introduction, the Part I work of the In-Grade Program has not been widely available. Some significant quotes from early reports of that time are reproduced in the Appendix. This paper is limited to Southern Pine data only and the sampling issues related to regions, mills, and pieces sampled in a mill (which also raises the question of lots in a mill). Part I on the 2 by 8 No. 2 data for Southern Pine that addresses the effort to look at the sampling issues will be examined first.

Review of Part I, 2 by 8 No. 2 Sampling and the 1980 Analysis

Table 1 summarizes the sampling done in Part 1 for 2 by 8 No. 2, which was stratified by regions and mill size (Green and others 1981). Mill size was defined, based on production, as small ($<15 \times 10^6$ board feet), medium ($15–40 \times 10^6$ board feet), and large ($>40 \times 10^6$ board feet). This method of sampling indicates that there was recognition of potential differences in mills by amount of production. It should be noted the number of mills, overall, for each size mill and the number of pieces within a mill were not constant throughout the design (Table 2). In all but one mill, a minimum of three 10-piece lots were taken. For a few mills, six 10-piece lots were taken. The initial analysis looked only at the effect of regions and mills within regions on modulus of elasticity (MOE), modulus of rupture (MOR), stiffness (EI), and moment potential (RZ). This analysis was done separately for both on-grade and as-graded samples. A nested analysis of variance (ANOVA) was performed to look at the effect of regions and mills nested in regions. If regions were statistically significantly different at the 0.05 level, a Duncan multiple-range test was performed to determine which regions were different from each other (Steel and Torrie 1960).

When the data for Part I were used to conduct the same Duncan multiple-range test, they reproduced similar results as were reported in Haskell (1980). Tables 3 and 4 show the results for MOR and MOE, respectively.

A brief discussion of how to interpret these results may be useful to some readers. Table 3 will be used to illustrate how

Table 3. GLM Procedure for on-grade MOR

Source	DF	Mean square	Test[a]	*F*-value	Pr > *F*
Region	5	77064521.9		3.30	0.0170
Mill (region)	30	23320396.9		4.11	<0.0001
Error	1331	5676534.0			

Duncan grouping			Mean[b]	*N*	Region
	A		6815.9	228	6
	A		6790.1	241	5
B	A		6558.6	239	3
B	A	C	6083.7	239	1
B		C	5770.0	240	4
		C	5314.9	180	2

[a]Test shows the structure from ignoring an effect (start of arrow) to using an effect (point of and arrow) of the statistical test used in creating the significance results in the table.
[b]Means with the same letter in the Duncan grouping are not significantly different at the 0.05 level.

Table 4. GLM Procedure for on-grade MOE

Source	DF	Mean square	Test[a]	*F*-value	Pr > *F*
Region	5	5.16739		4.34	0.0044
Mill (region)	30	1.19175		7.09	<0.0001
Error	1,331	.16815			

Duncan grouping			Mean[b]	*N*	Region
	A		1.8086	241	5
B	A		1.7162	239	3
B	A	C	1.6766	228	6
B		C	1.5008	239	1
		C	1.4632	240	4
		C	1.4501	180	2

[a]Test shows the structure from ignoring an effect (start of arrow) to using an effect (point of and arrow) of the statistical test used in creating the significance results in the table.
[b]Means with the same letter in the Duncan grouping are not significantly different at the 0.05 level.

to interpret the results. GLM stands for general linear model. The first part of the output at the top provides a summary of the results of SAS ANOVA (SAS 2012). The first column of the table shows the sources of variability (regions, mills-within-a-region, and error, which is a combination in this case of lot variability within a mill in a region and piece variability within-a-region, mill, and lot). Column 2 gives degrees of freedom for each source of variability. The mean square error gives the variability of the source. Test is a visual display of the error term used to test each source of variability. Then comes the *F*-test and finally the significance level of the *F*-test. The table below the ANOVA gives a summary of the multiple comparison test used to see which regions are significantly different from each other. Regions connected with the same letter under the grouping column are not statistically different at the 0.05 level.

The results shown in Tables 3 and 4 indicate that both the regions and the mills within regions are statistically significant at the 0.05 level. The statement often quoted from Jones (1989) is the following:

"This phase (Part I) of the U.S. program had a sampling strategy intended to evaluate the strength properties for the most typical species-grade-size for a given application, as well as the effect of mill size, regions, and peeler core mills. The results were used to design the rest of the program. Of particular importance was the conclusion that between-mill variation was not statistically significant considering within-mill variation."

This statement is poorly phrased and is not meant to suggest that the differences between mills were not observed to be significant. It is intended to emphasize, as described in Green and Evans (1988), that the variation in properties within a mill was at least as large as that between mills. The equivalent variation within a mill and between a mill also meant that fewer mills could be taken with more samples per mill (Green and Evans 1988).

Alternatives to 1980 Analysis

Other ways of looking at the data are now readily available and should be investigated. Duncan's multiple range test has

Table 5. GLM procedure for on grade MOR

Source	DF	Mean square	Test[a]	*F*-value	Pr > *F*
Region	5	77064521.9		3.30	0.0170
Mill (region)	30	23320396.9		3.71	<0.0001
Lot (region mill)	101	6289237.1		1.12	0.2078
Error	1,230	5626223.0			

Tukey grouping		Mean[b]	*N*	Region
	A	6815.9	228	6
	A	6790.1	241	5
B	A	6558.6	239	3
B	A	6083.7	239	1
B	A	5770.0	240	4
B		5314.9	180	

[a]Test shows the structure from ignoring an effect (start of arrow) to using an effect (point of and arrow) of the statistical test used in creating the significance results in the table.
[b]Means with the same letter in the Tukey grouping are not significantly different at the 0.05 level.

Table 6. GLM Procedure for on-grade MOE

Source	DF	Mean square	Test[a]	*F*-value	Pr > *F*
Region	5	5.16739		4.34	0.0044
Mill (region)	30	1.19175		4.07	<0.0001
Lot (region mill)	101	.29274		1.85	<0.0001
Error	1,230	.15792			

Tukey grouping		Mean[b]	*N*	Region
	A	1.8086	241	5
B	A	1.7162	239	3
B	A	1.6766	228	6
B	A	1.5008	239	1
B		1.4632	240	4
B		1.4501	180	2

[a]Test shows the structure from ignoring an effect (start of arrow) to using an effect (point of and arrow) of the statistical test used in creating the significance results in the table.
[b]Means with the same letter in the Tukey grouping are not significantly different at the 0.05 level.

largely been replaced with other multiple comparison tests that have better properties such as a lower number of false significant results and a better way of handling different sample sizes for the treatments being compared. The Tukey–Kramer test, generally referred to as a Tukey test or Tukey hsd (honestly significant difference) test has been viewed by some statisticians as performing optimally or near optimally in a broad variety of circumstances (Steel and Torie 1960; Stoline 1981). So it will used in any further testing. Also, the lot variation within a mill in a region was combined with the piece-to-piece variation within a lot in a mill and region in the original analysis. In the interest of having a better understanding of the sources of variability, the nested lot variability will be added to the analysis. Finally, some other variables will be looked at in addition to the original MOR and MOE.

Tables 5 and 6 show the analysis of MOR and MOE, respectively, using the Tukey test and separating the error variance into lot and piece variability. The results show a decrease in significant differences among regions as would be expected from a more conservative test. Looking at lots versus pieces within a lot shows lots were not significant for MOR, but highly significant for MOE. This result would imply that there might be a benefit to sampling from multiple lots for MOE, but not necessarily for MOR. Because it is not known which mills came from an existing inventory and how well spaced across time lots were for mills without inventory, it is hard to determine a reason for this.

Another thing that can be done with the data that might prove beneficial is to adjust all specimens to a common moisture content. A standardized procedure for adjusting the properties of dimension lumber was not yet available in 1980 and was at the time a subject of ongoing research. Using present ASTM procedures, the MOR and MOE of all pieces were adjusted to 15% moisture content and labeled MOR15 and MOE15. For species that can be produced either green or dry, this could eliminate one source of variation. Tables 7 and 8 provide the results of looking at MOR15 and MOE15. For both MOR15 and MOE15, there is very little difference from the results for MOR and MOE. For MOR15, the regional effect is slightly less but still significant; and where region 2 results for MOR were significantly different from regions 6 and 5, for MOR15 region 2 is different only from region 6. For MOE15, the regional effect is also slightly less but still significant; and where region 5 results for MOE were significantly different from regions 2 and 4, now for MOE15 region 5 is significantly different from regions 2, 4, and 1.

Variance Component Analysis

In addition to testing the mean regional results, the original 1980 analysis looked at how much each source of variation contributed to the total variability of MOR and MOE through variance component analysis. The way this was done for the regional analysis in 1980 was to use the model that says each observation (Y) consists of an overall mean (μ) plus a regional effect (R) plus a mill within region effect (M) plus an error term (ϵ). Each effect is either fixed, random, or finite. Technically, region should be a fixed effect because the entire population of regions is sampled. Mills would be a finite effect because a subset of the finite number of mills in a region is sampled, and error is a random effect drawn from a population that is essentially infinite. Because only six mills were sampled from each region that has a very large number of mills that could be chosen, the finite population correction factor is essentially 1 as it would be for a random effect. So mills are often considered as a random effect. The original analysis chose to also label region as a random effect noting "It is recognized that this assumption is not strictly appropriate for the study. However, no statistical significance is attached to these estimates. They are only used to address the question of practical significance" (Haskell 1980).

Table 7. GLM Procedure for on-grade MOR15

Source	DF	Mean square	Test[a]	*F*-value	Pr > *F*
Region	5	60627810.7		2.92	0.0289
Mill (region)	30	20745198.4		3.49	<0.0001
Lot (region mill)	101	5950500		1.09	0.2682
Error	1,230	5473811			

Tukey Grouping		Mean[b]	*N*	Region
	A	6667.4	228	6
B	A	6605.0	241	5
B	A	6322.5	239	3
B	A	5872.6	239	1
B	A	5710.8	240	4
B		5325.6	180	2

[a]Test shows the structure from ignoring an effect (start of arrow) to using an effect (point and arrow) of the statistical test used in creating the significance results in the table.

[b]Means with the same letter in the Tukey grouping are not significantly different at the 0.05 level.

Table 8. GLM Procedure for on-grade MOE15

Source	DF	Mean square	Test[a]	*F*-value	Pr > *F*
Region	5	4.295939		4.02	0.0065
Mill (region)	30	1.068413		4.18	<0.0001
Lot (region mill)	101	.255489		1.70	<0.0001
Error	1,230	.150143			

Tukey Grouping		Mean[b]	*N*	Region
	A	1.7692	241	5
B	A	1.6665	239	3
B	A	1.6422	228	6
B		1.4607	239	1
B		1.4540	240	4
B		1.4516	180	2

[a]Test shows the structure from ignoring an effect (start of arrow) to using an effect (point and arrow) of the statistical test used in creating the significance results in the table.

[b]Means with the same letter in the Tukey grouping are not significantly different at the 0.05 level.

Using the ANOVA for each MOR and MOE, expected mean squares can be calculated for the data. These are given in Tables 9 and 10. From these, the actual mean squares for each factor can be used to estimate the size of each component of variation. The resulting variance component estimates are shown below the respective ANOVA tables. One way to estimate variance components is as follows: note that the error mean square is the estimate of the error term (ϵ), which can then be used to estimate the mill within region (M) variance component, etc. When finished, these estimates can be used to estimate the variability of an individual observation and decide how best to allocate the sampling.

Another interesting feature to point out is found in the expected mean squares. If you look at the regions and mills-within-region effects, it can be seen that the number in front of the variance of mill-within-region term is different. This is due to the slight unbalanced sample sizes taken in the study. In the study, six regions and six mills in a region were chosen. However some mills had three lots and some had six lots. In addition, an occasional lot only had nine pieces. This means the test using mill-within-regions to test regional effects is only approximate. However, the difference from being an exact test is not very much and should not make any practical difference in any conclusions.

In trying to reproduce the analysis that was used in Haskell (1980), we must note that the exact analysis is not given in detail in the paper. Haskell states that he used SAS to estimate variance components. The first step is to calculate the expected mean squares. If SAS was used as shown in Tables 9 and 10 in a way to estimate the expected mean squares that take into account the slight unbalance in the sample sizes, the results are as shown. However, the difficulty of calculating expected mean squares led many researchers to use a balanced design for calculations that approximated the unbalanced design. Because the paper does not actually show the expected mean squares, we are using

Table 9. Variance components estimation procedure for MOR

Source	DF	Mean square	Expected mean square
Region	5	77064522	variance(error) + 42.238 variance (mill(region)) + 227.42 variance(region)
Mill(region)	30	23320397	variance(error) + 37.098 variance (mill(region))
Error	1,331	5676534	variance(error)

Variance component	Estimate
Variance (region)	225578.5
Variance (mill(region))	475596.2
Variance (error)	5676534.0

Table 10. Variance components estimation procedure for MOE

Source	DF	Mean square	Expected mean square
Region	5	5.16739	variance(error) + 42.238 variance (mill(region)) + 227.42 variance(region)
Mill(region)	30	1.19175	variance(error) + 37.098 variance (mill(region))
Error	1,331	0.16815	variance(error)

Variance component	Estimate
Variance (region)	0.01686
Variance (mill(region))	0.02759
Variance (error)	0.16815

the unbalanced design. The actual design is so close to balanced, the results should be very similar either way.

In either case, the estimates of the variance components would proceed in a similar fashion. The estimate of the variance (error) is the error mean square in column 3. Thus, from Table 9, the variance (error) equals 5676534. The estimate of the mill(region) variance component comes from the mean square of 23320397 equal to the variance of the error of 5676534 plus 37.098 time variance (mill(region)). Similarly, the variance (region) can be estimated. This method of estimating variance components was still common at the time of the Haskell report. Now many new and better procedures are available.

Having the variance component estimates can help design more efficient studies. For example, consider the MOE results in Table 10. If the study had six regions, six mills in a region, and 30 pieces within a mill in a region, the mean MOE would have the following variance:

$$\mathrm{Var}(\text{mean MOE}) = \frac{\mathrm{Var}(\text{error})}{30\times 6\times 6} + \frac{\mathrm{Var}(\text{mill}(\text{region}))}{6\times 6} + \frac{\mathrm{Var}(\text{region})}{6}$$

Using the variance component estimates above as estimates of these variances

$$\begin{aligned}\mathrm{Var}(\text{mean MOE}) &= \frac{0.16815}{1080} + \frac{0.02759}{36} + \frac{0.01686}{6}\\ &= 0.00016 + 0.00077 + 0.00281\\ &= 0.00374\end{aligned}$$

Note that if the study design had six regions, nine mills per region, and 20 pieces per mill in a region, the variance would have been less (0.00348). If the study had 12 regions, three mills, and 30 pieces per mill in a region, the variance would have also been considerably less (0.00233). The within mill variation is much larger than the mill within a region variation and the regional variation, but that does not mean that the latter two are not important and thus need to be incorporated in the design to get the best estimate for the same number of specimens tested. It is a balancing act of cost and accuracy to determine the best sampling method.

Discussion and Conclusions

Any sampling method must be constructed using a balancing process to account for the various significant sources of variability. In the case of the In-Grade Program, those sources of variability were identified as being region, between mill-within-a-region, and pieces within-a-mill as influenced by lots. In the original analysis of 1980, the conclusion of the regional analysis stated, “Even though the variance component for regions is small, the effect is significant and unstable between dependent variables. The use of regions should be maintained for planning to guarantee representative sampling and because other grades and sizes may display different effects” (Haskell 1980). The reanalysis of the 1980 data also shows this. Between-mill variability was also observed to be a significant source of variability. Finally, the piece variability within a mill was identified as being the largest source of variability.

Application of the Results from Part I to Part II of the In-Grade Program

By the time that Part II of the In-Grade Program was started, in addition to the statistical significance of regions, another reason for maintaining regions had been recognized as valuable. In Part II other grade–size combinations were to be tested. It was planned to go to a mill and collect two lots for each of the grade–size combinations chosen for the Part II studies. Invariably, as was discovered in Part I, at the time of the visit, some mills did not have much in the way of inventory on hand or may not have been producing the particular grade-size combinations needed to collect a sample.

Two elements of the sampling method in Part I greatly reduced the effort and cost of the sampling in Part I and the design of Part II.

First, if a mill had been the primary unit of sampling and the region was ignored, a sampled mill that did not have all the material needed would have required another mill to be randomly selected. That might mean having to travel to any other mill in the population to try to obtain the rest of the missing sample. This could require more than one additional mill to be sampled. Sampling by regions and mills in a region limited the potential cost and time it might take to acquire the sample.

Second, the assumption that the regions divided the total growth region into homogeneous regions also played an important part into further reducing costs. If the region is homogeneous, any mill in the region would have material from the same population. That meant if a mill chosen randomly in a region and when sampled only had 2 by 4 and 2 by 10 for the select structural specimens, any mill in a homogeneous region could provide 2 by 8 specimens. This concept was so important to the cost of the Program that agencies subdivided the original six regions into sub-regions based on other studies (such as the density survey studies done in the 1960s: U.S. Forest Service 1965a, b) and on expert analysis by people from the grading agencies that had a long history of interacting with mills in a proposed sub-regions. Issues like elevation, soil quality and type, climate factors, and many other criteria were available to the experts that decided on 18 regions for Southern Pine as shown in Figure 2. When actual sampling in Part II occurred, two of the original regions (15, 16) had no mills that could be sampled and thus only 16 regions were used.

The assumption of homogeneity of a region was accepted by a broader group that reviewed the sampling method. Regions were accepted as a primary sampling unit because their variability, when compared to the mill within region variability, was significant.

The division of the regions into smaller geographic areas was based on the issues discussed above. It was felt that the multiple regions would enhance the homogeneity within a region. This assumption allowed sampling regions in proportion to production in order to facilitate a simpler estimation of population parameters.

Conclusions from the Reanalysis of Past Data

In the re-examination and re-analysis of the 1980 data, both the regions and the mills within regions are statistically significant at the 0.05 level. This information was available and used as part of the background decision making for setting up Part II of the In-Grade Program. It was also determined that a lot was highly significant for MOE, suggesting that it would be beneficial to continue sampling by lots.

Summary of What Was Learned from Looking Back at Part I

The development of the current sampling and testing methodology that is used to determine global numbers for major species groups involved a series of tradeoffs between capturing the significant sources of variability and what was practical to accomplish. At the time that the decisions were being made for sampling, it was known that regions and between-mill variation were a significant part of the overall variation that would describe a global number. The reexamination and reanalysis of the 1980 data suggest that both the regions and the mills within regions were statistically significant in accounting for variation at the 0.05 level. The establishment of the sampling method that is currently used in determining global numbers for the major species groups in the North America took this finding into account. The current random sampling of mills within a region in proportion to production was based on a balancing of the significant regional, mill-to-mill and within-mill piece variability. It was also affirmed that the lot variability for MOE was highly significant. The use of region in the sampling method had the added benefit of controlling costs for sampling and testing. Thus, the decision to continue with the random sampling of mills in proportion to production by region was justified and necessary to keep costs at a reasonable level to conduct the extensive sampling envisioned for Part II.

References

American Forest & Paper Association and American Wood Council, National Design Specification for Wood Construction. [current edition] Washington, D.C.: American Forest & Paper Association.

ASTM. 2011. Annual Book of Standards, Vol. 04.10. West Conshohocken, PA.: American Society for Testing and Materials.

a. ASTM D1990–07. Standard methods for establishing allowable properties for visually-graded dimension lumber from In-Grade tests of full-size specimens.

b. ASTM D2915–10. Standard method for evaluating properties for stress grades of structural lumber.

c. ASTM D4761–05. Standard methods for mechanical properties of lumber and wood-base structural materials.

Bodig, J. 1977. Bending properties of Douglas-fir-Larch and Hem-Fir dimension lumber. Special report No. 6888. Department of Forestry and Wood Science. Colorado State University, Fort Collins, CO. 59 p.

FPRS. 1989. Proceedings of the Workshop on In-Grade Testing of Structural Lumber, Madison, WI, April 25–26, 1988, Proceedings 47363, Forest Products Research Society: Madison, WI.

Galligan, W.L., and D.W. Green. 1980. Development of standardized concepts for assignment and assessment of the mechanical properties of lumber—a continuing challenge for the 80's. Forest Products Journal. 30(9):39–46.

Galligan, W.L., D.W. Green, D.S. Gromala, J.H. Haskell. 1980. Evaluation of lumber properties in the United States and their application to structural research. Forest Products Journal. 30(10):45–50.

Goodman, J.R., M.E. Criswell, J. Bodig. 1974. A rational analysis of wood joist floor systems. Final Report; National Science Foundation Grant GK-30853. Colorado State University, Fort Collins, CO.

Green, D., W.L.A. Marin, J.W. Evans. 1981. Flexural properties of 2 × 8, No. 2 and 2 × 4 stud grades of Douglas-fir Southern Pine dimension lumber. Un-numbered Buff Colored Report. USDA Forest Service, Forest Products Laboratory, Madison, WI. 171 p.

Green, D. W. 1983. In-grade testing impetus for change in the utilization of structural lumber. In: Proceedings, from stump thru mill: recent advances in spruce-fir utilization technology. Publication No. 83-13. Bethesda, MD: Society of American Foresters.

Green, D.W. and J.W. Evans. 1988. Evaluating lumber properties: practical concerns and theoretical restraints. Invited keynote address for the Wood Properties session. International Conference on Timber Engineering. American Society of Civil Engineers, Seattle, WA. Sept. 19–22, 1988.

Hans, G.E., W.L. Galligan, W.L. Lehmann, H.M. Montrey, R.C. Moody. 1977. Five-year action plan for light-frame construction research. U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, Madison, WI.

Haskell, J.H. 1980. Interim Statistical Report Part I. "For Official Use Only not for publication." USDA Forest Products Laboratory, Madison, WI. Internal Report, June 1980.

Jones, E. 1989. Sampling procedures used in the in-grade lumber testing program. In: Proceedings of workshop sponsored by In-grade Testing Committee and Forest Products Society. Proceedings 47363. Madison, WI: Forest Products Society.

Madsen, B. 1978. In-grade testing: Problem analysis. Forest Products Journal. 28(4):42-50.

Polensek, A., and Atherton G.H. 1975. Compression, bending strength, and stiffness of walls with utility grade studs. Forest Products Journal. 26(11):17–25.

SAS. 2012. Statistical Analysis System. Cary, NC: SAS Institute Inc.

Stoline, Michael R. 1981. The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. The American Statistician (American Statistical Association). 35(3): 134–141.

Steel, R.G., J.H. Torrie, D.A. Dickey. 1960. Principles and procedures of statistics. New York: McGraw-Hill Book Co. 481 p.

Tuomi, R.L., W.J. McCutcheon. 1975. Conventional house challenger simulated forces of nature. Forest Products Journal. 25(6):13–20.

U.S. Forest Service. 1965a. Southern wood density survey: 1965 status report. Research Paper FPL-RP-26. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 40 p.

U.S. Forest Service. 1965b. Western wood density survey; report no. 1. Research Paper FPL-RP-27. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory.

Appendix—Quotes of Interest

Quote from Haskell (1980)

Region Analysis Conclusion

"Even though the variance component for regions is small, the effect is significant and unstable between dependent variables. The use of regions should be maintained for planning to guarantee representative sampling and because other grades and sizes may display different effects."

Mill Size Analysis Conclusion

"From a practical view point mill size was not an important factor in 2 × 8's of No. 2 grade. Pieces within a mill was once again the dominant source of lumber property variation. This statement does not imply that weighting design values estimates by mill volume can be ignored."

Mill Type Analysis Discussion and Conclusion

"Variance component estimates for mill type in the case of on-grade Southern Pine 2 x 4's are greater than those for mills within mill type. This fact is contrary to the main effects discussed earlier. Thus, of the main effects studied in this report, mill type would appear to be a major factor of importance for future sampling considerations. Note, however, the variance component for pieces within mills is still much larger than the other components."

Quote from Green and Evans (1988)

"In the initial phase of Part I of the In-Grade program, the committee conducted a limited amount of testing to evaluate our procedures and to get an estimate of property variability. From this initial study the committee could detect no drastic mechanical property variations across geographic growth regions for the major volume species groups. Further, the variation in properties within a mill was at least as large as that between mills. This meant that we could divide the species growth regions into broad subregions to assure a representative sample. The equivalent variation within a mill and between a mill also meant that we could sample fewer mills with a larger sample per mill. These assumptions saved us a considerable amount of time and money in collecting our sample because we did not have to collect as many samples from as many mills. However, it also restricted our ability to precisely identify any low-strength pockets in our data. Had the committee's objective been to determine the existence of low pockets, or if the between-mill variation had been much larger than the within-mill variation, we would have sampled more mills and taken fewer samples per mill."