Statistical Modeling of Discrete Percentage Measurements with Application to Construction of Acceptance Bounds for Wood Failure in Structural Adhesive Testing

Thomas M. Loughin[1], Nathaniel Payne[1,2], Romulo Casilla[3], Conroy Lum[3]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6, email: tloughin@sfu.ca

[2]Cymax Group, 4170 Still Creek Dr #310,Burnaby, BC V5C 6C6

[3]Advanced Building Systems Department, FPInnovations, 2665 East Mall, Vancouver, BC, Canada, V6T 1Z4

Abstract

The goals of this paper are (1) to provide a statistical analysis approach that is appropriate for data from an interlaboratory study where responses are measured in discrete percentages and are subject to multiple sources of random variability, and (2) to apply this model to data on wood-failure percentages from block-shear tests on structural wood adhesives. We treat percentage responses measured in 5-point intervals as having arisen from observing 20 independent binary responses on different parts of the observed wood blocks. The overdispersion that is likely to result from the practical inadequacy of this assumption is overcome empirically by the inclusion of a random effect for blocks. We propose an analysis based on a parametric bootstrap to provide sampling distributions for statistics that regulators might wish to use in setting standards for acceptance of wood adhesives. Similar computational methods are developed to assess the fit of the model. This model is shown to provide a reasonably good fit for actual data in many of the cases to which it was applied.

Keywords: Binomial, random effects, laboratories, overdispersion, wood, adhesive

Final version published as Loughin, T. M., Payne, N., Casilla, R., and Lum, C., "Statistical Modeling of Discrete Percentage Measurements With Application to Construction of Acceptance Bounds for Wood Failure in Structural Adhesive Testing," *Journal of Testing and Evaluation* 45(5), https://doi.org/10.1520/JTE20150189

Introduction

Structural wood adhesives (SWAs) are used to create engineered wood products such as laminated wood, plywood, I-joists, and finger-jointed lumber [1]. Before an adhesive can be used in such products, it must meet laboratory testing standards based on protocols that vary by country. For example, in Canada, SWAs are subject to testing according to the standards CSA O112.9-10 [2] or CSA O112.10-08 [3], while in the US, ASTM D2559-10a [4] applies. In both Canada and the US, it is not uncommon for product specifications to reference additional standards to assess other attributes of the adhesives. For example, ASTM D7247 [5] provides additional requirements for assessing the adhesive's high temperature performance in certain applications. These protocols, and several others mentioned in the Discussion and Conclusions section, specify several tests that are used to assess the suitability of an adhesive for use in a structural wood product (i.e. a wood product that can bear load in a building structure). Among these tests is a block shear test. ASTM D905 [6] provides a detailed description of the adhesive block shear test. In the block shear test, flat surfaces from pairs of wood blocks are glued together under pressure. A smaller sample (typically called a "shear block") is cut from this glued assembly in such a way that it can be placed in a specialized test jig that induces shear on the glued interface. The loading is such that the forces are applied parallel to the wood grain, which produces the highest shear forces before either the wood fails or the adhesive fails. An increasing load is applied to the shear block until it fails; at that point, it is possible to separate the specimen along the formerly glued interface. Optionally, the shear block may be exposed to some predefined conditions to simulate aging after the adhesive has cured or set and before the shear force is applied.

If the SWA is strong and durable, then it is likely that the separation occurs mainly cohesively within the wood. On the other hand, with a weak or nondurable SWA, the separation occurs cohesively within the adhesive and/or at the interface between the wood and the adhesive layer. The force required to achieve separation is recorded along with a trained technician's subjective assessment of the percentage of the block face that shows wood failure rather than adhesive failure. A practice for estimating the percentage of wood failure in wood-adhesive bonds is given in ASTM D5266 [7]. The wood failure percentage (WF%) is recorded in 5-unit increments; i.e., WF% takes a value 0, 5, 10, …, 95, or 100. A set of 30 blocks is tested in this way, and the median and first quartile of WF% are used to summarize the results. High values of these quartiles indicate a strong and durable SWA.

It is recognized that laboratory test measurements like WF% are subject to numerous sources of variability, including variability among trained evaluators and laboratories [8]. The latter source is important, because (a) manufacturers of SWAs may contract with any of a number of accredited laboratories to perform the tests on their adhesives, and (b) regulators need to be reasonably certain that results are representative of the true potential for both performing and non-performing SWAs, regardless of the laboratory undertaking the test. In order to determine an acceptance criterion for measures from tests that can be conducted in different laboratories, the *repeatability* and *reproducibility* of the test must be understood [8]. In essence, repeatability relates to the variability of a given test result upon repeated tests under the same conditions within the same laboratory, while reproducibility relates to the additional variability that is imparted when the same test is conducted in different laboratories. Because these effects are random and tend to be difficult to isolate and replicate, they are best quantified in an interlaboratory study by a representative sample of “qualified” laboratories all following their interpretation of the same test standard.

The standard practice for conducting interlaboratory studies is given in ASTM E691-09 [9], and specifies running some number of replicate tests at each of several laboratories. An analysis of the results should provide enough information to reliably interpret the outcome of a single test run at a single laboratory. Specifically, the goal of any acceptance criterion is to ensure that, across all qualified laboratories, poor-performing products are designated as not acceptable and good performing products are accepted, while still keeping the testing as practical and cost effective as possible. The analysis of the interlaboratory study should therefore be able to set limits such that a product with a given level of performance should equal or surpass the limit with probability that is easily computed. This implies that it is necessary to be able to estimate the sampling distribution of the statistic on which the criterion is based. The ASTM E691-09 [9] standard specifies that statistical analysis of the test results is conducted using a one-way random-effects model. The standard assumes that the summary measure for each laboratory is the mean response of all tests conducted there, and thus it provides calculation formulas that can be performed using a spreadsheet to provide estimates of the repeatability and reproducibility. From these, limits can be constructed within which acceptable products should fall with prescribed probability when they are tested at a random laboratory that follows the protocol correctly.

While the one-way random-effects model is appropriate for normally distributed data [10], it has several flaws when an interlaboratory test is done to set standards for WF% and similar percentage-based subjective assessments. First, measurements like WF% are discrete. In particular, for a strong and durable SWA, the WF% values often take on only a few of the largest possible values. It is not unusual that 30 separate block shear tests on a good adhesive results in a majority of measurements at 100%, with a few values of 95% and 90%. With such a skewed distribution and so few unique values in the data, the justification of a normal-based random-effects model is questionable.

Second, data from very good SWAs (means near 100%) or very poor SWAs (means near 0%) exhibit block-to-block variability that is much smaller than what is observed when the mean WF% is more intermediate. Furthermore, in standards such as CSA O112.9-10 [2] or CSA O112.10-08 [3], quartiles are used to summarize WF% data from a given laboratory, rather than the means that a standard 1-way random-effects model assumes are to be used. Although ASTM E691-09 [9] does indicate that caution is needed with discrete data, it offers no alternative analysis. Unfortunately, the sampling distributions of quartiles of data from skewed, discrete distributions with random effects are not known and are not easy to derive exactly.

The goal of this paper is to provide a statistical analysis approach that is appropriate for data from an interlaboratory study where responses are measured in discrete percentages. This analysis approach can then be used to create acceptance criteria for measurements like WF%. We first transform the data so that they may take on values from the consecutive integers 0, 1, 2, …, 20. We then argue that these data can be modeled approximately using an overdispersed binomial distribution, which is described in the next two sections. Accounting for the laboratory effects results in a model from the class of generalized linear mixed models [11], [12]. We show how this model can be fit and use a parametric bootstrap [13] to estimate limits of repeatability and reproducibility. We then describe a pilot interlaboratory study designed to examine the impacts of repeatability and reproducibility on the acceptance or rejection of performing or non-performing SWAs. We analyze these data and assess whether the proposed model provides a reasonable fit to the data using parametric bootstrap techniques.

Mixed Model Analysis of Discrete Percentage Data

Binomial Approximation to Wood Failure Percentage

To start, we fix some notation. Suppose that B blocks are tested in each of L laboratories. Let Y_{ik} represent the response measurement on block k in laboratory i, for $i = 1, \dots, B;\ k = 1, \dots, L$. These random variables are assumed to be supported on the equally spaced values 0, 5, 10,…,100.

Let $W_{ik} = Y_{ik}/5$, so that W_{ik} is supported on 0,1,2,...,20. Notice that this structure suggests that we might approximate the distribution of W_{ik} with a binomial distribution with 20 trials and probability of success π, denoted $Bin(20, \pi)$ [12]. In fact, this distribution would be correct if the face of each block were divided into 20 regions of equal size; if each region were assigned a 1 or a 0 according to whether the wood either did or did not fail, and the recorded response represented the sum of these indicators; if the probability of failure were constant in each region; and if the regions were independent of one another. Although this is not at all how the responses are obtained—they are merely discrete visual approximations of the proportion of the block face that has experienced wood failure—we nonetheless consider $Bin(20,\pi)$ to be the starting point for an empirical working model that may provide a reasonable fit for the data.

To develop this model further, refer to the 20 hypothetical regions on the block face as “pseudo-trials.” Note that neighboring pseudo-trials would be expected to respond more similarly to one another than to those on distant regions of the same block, because the tearing of wood fiber does not respect the hypothetical boundaries between pseudo-trials. This creates a positive spatial correlation among the binary responses on the pseudo-trials, thus violating the assumption of independent trials. Positive correlation among trials causes the counts to have more variability relative to what is expected under a binomial distribution [12]. That is, the counts are “overdispersed.”

Because the bond performance depends on a number of complex factors (wood, adhesive, and interphase regions of wood and adhesive) [1], it is possible that properties of wood strength and wood-adhesive bonding vary randomly within a block, causing different regions of the block to be more or less likely to experience wood failure. This variability in probabilities of success would also cause overdispersion of the resulting counts [12]. Thus, while the $Bin(20,\pi)$ working model does not perfectly represent WF% counts , the main consequences of its two primary defects both lead to the same result. Therefore, an overdispersed binomial model might provide a very reasonable approximation to the distribution of W_{ik}. We show in the section, "Assessing the Model Fit," that this model often provides a reasonable fit for real data from a pilot study.

Generalized Linear Mixed Model for Interlaboratory Study Data

Overdispersion can be incorporated into a binomial model in several ways [11], [12]. The model can be changed to one which allows extra-binomial variability, such as a beta-binomial; the likelihood function can be empirically adapted to allow extra variability using quasi-likelihood; or random effects can be added to the model, creating a generalized linear mixed model (GLMM). Since the ASTM E691-09 [9] standard uses a random-effects linear model to account for interlaboratory variability when data are assumed to arise from a normal distribution, we take the parallel approach here by using a GLMM.

To construct a model for WF% from a particular SWA, let π_{ik} be the probability of wood failure in each pseudo-trial from block k in laboratory i, for $k = 1, \ldots, B$ and $i = 1, \ldots, L$. We model the transformed responses as $W_{ik} \sim Bin(20, \pi_{ik})$, with

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \alpha + l_i + b_{ik}, \qquad (1)$$

where α is the unknown average log-odds (logit) of wood failure among all possible laboratories and blocks; $l_i \sim N(0, \sigma_L^2), i = 1, \ldots, L,$ are independent random effects representing the deviation of laboratory i from the average logit; and $b_{ik} \sim N(0, \sigma_B^2), k = 1, \ldots, B,$ are independent random effects representing the deviation of block k from the average logit among all possible blocks in laboratory i [11], [12]. The anticipated overdispersion of block responses within a laboratory is accounted for by block random effects, while the inter-laboratory variability is accounted for by the laboratory random effects. As is customary, it is assumed that all l_i and b_{ik} are independent. The mean logit, α, and the variance components, σ_L^2 and σ_B^2, are unknown and must be estimated from the data. The standard method is maximum likelihood (ML) estimation using computational approximations outlined in [12] and described in more detail in [11].

Analysis of the Model

Recall that the acceptance criteria for WF% are based on the first two sample quartiles, $\hat{Q}_1$ and $\hat{Q}_2$, from a set of 30 tested blocks at a single laboratory. Unfortunately, the sampling distributions of quartiles from a binomial GLMM are not known and not easy to derive exactly. A parametric bootstrap [13] is used instead to approximate the required sampling distributions.

The bootstrap is a computational statistical procedure that can be applied to many problems to estimate various properties of a statistic, such as its standard error. The essential idea is to mimic what one would like to do in an ideal world—take sample after sample from the population, compute the statistic on each sample, and use the distribution of these statistics to infer properties of the statistic on the original sample. However, because one cannot collect endless amounts of data in real life, a model of the population is constructed and the new samples (called “resamples”) are drawn from this model. In particular, a parametric bootstrap begins with a model for the distribution of the data. The model contains unknown parameters that are estimated by the data. The resampling process is then a simple computational process that draws a large number of resamples, each containing the same number of observations as in the original sample. The statistic is computed on each resample, and the resulting distribution of these statistics provides information regarding the properties of the original statistic.

In the present context, let $\hat{\alpha}, \hat{\sigma}_L^2$, and $\hat{\sigma}_B^2$ be the ML estimates of their respective parameters from model (**Error! Reference source not found.**). Alternatively, $\hat{\alpha}$ might be derived from a specific expected performance level of a product, $\hat{Y},$ via $\hat{\alpha} = \log(\hat{Y}/(100 - \hat{Y}))$. The parametric bootstrap simply substitutes these estimates for their parameters in model (1), and uses the estimated form of the model to simulate data for a test of a SWA at a single laboratory. This simulated data set is summarized into a sample first quartile and median. Repeating this process a large number of times provides a large number of simulated laboratory results for each statistic. The empirical distribution of estimates for each quartile approximates the sampling distribution of each quartile under model (1).

The step-by-step process for an adhesive with average logit $\hat{\alpha}$ is as follows:

1. Select a random laboratory effect l_r^* from $N(0, \hat{\sigma}_L^2)$.
2. Select 30 random block effects $b_{rk}^*, k = 1, \dots, 30$, independently from $N(0, \hat{\sigma}_B^2)$.
3. Compute 30 logits, according to model (1): $\gamma_{rk}^* = \hat{\alpha} + l_r^* + b_{rk}^*$
4. Transform the logits into probabilities, $\pi_{rk}^* = \exp((\gamma_{rk}^*)/(1 + \exp(\gamma_{rk}^*))$
5. For each probability, generate one observation $W_{rk}^* \sim Bin(20, \pi_{rk})$, and set $Y_{rk}^* = 5W_{rk}^*$.
6. Compute Q_1^{*r} and Q_2^{*r} from $Y_{r1}^* \dots Y_{r,30}^*$.
7. Repeat for $r = 1,2, \dots, R$ for some large number R.

Appropriate limits and acceptance criteria can be derived from the estimated sampling distributions of $\hat{Q}_1$ and $\hat{Q}_2$, which consist of the empirical distributions of $Q_1^{*r}, r = 1, \ldots, R$ and $Q_2^{*r}, r = 1, \ldots, R$.

The method above produces limits for *reproducibility*; that is, it takes into account the variability imparted upon the test by the different conditions in the different laboratories. In order to produce limits that reflect *repeatability*, simply set $\sigma_L^2 = 0$ in step 1, or equivalently fix $l_r^* = 0$, so that there is no variation in laboratory effects.

Example: Interlaboratory Study of Structural Wood Adhesives

A “round-robin” test was conducted to collect data toward establishing the reproducibility and repeatability limits for selected tests in the first edition of the CSA O112.9-10 standard [2]. Six laboratories participated by testing block shear specimens prepared centrally by FPInnovations in Vancouver, BC, Canada. Specimens consisted of lumber cut from Douglas fir meeting the wood quality characteristics as specified in CSA O112.9 [2]. Specimens were glued using one of four adhesives: one known to pass the new standard (labeled “W”), one near the border (“X”), and two failing the standard (“Y” and “Z”). Specimens were labeled and sent to laboratories, where each specimen was subjected to one of three different treatments: “dry” (or untreated), “vacuum pressure” (VP), or multiple cycles of “boil-dry-freeze” (BDF). Specimens were then shear-tested until failure. For each combination of adhesive and treatment, each laboratory tested 32 blocks, which were scored by a designated trained reader for that laboratory. The measurements from the first 30 blocks represented the primary test

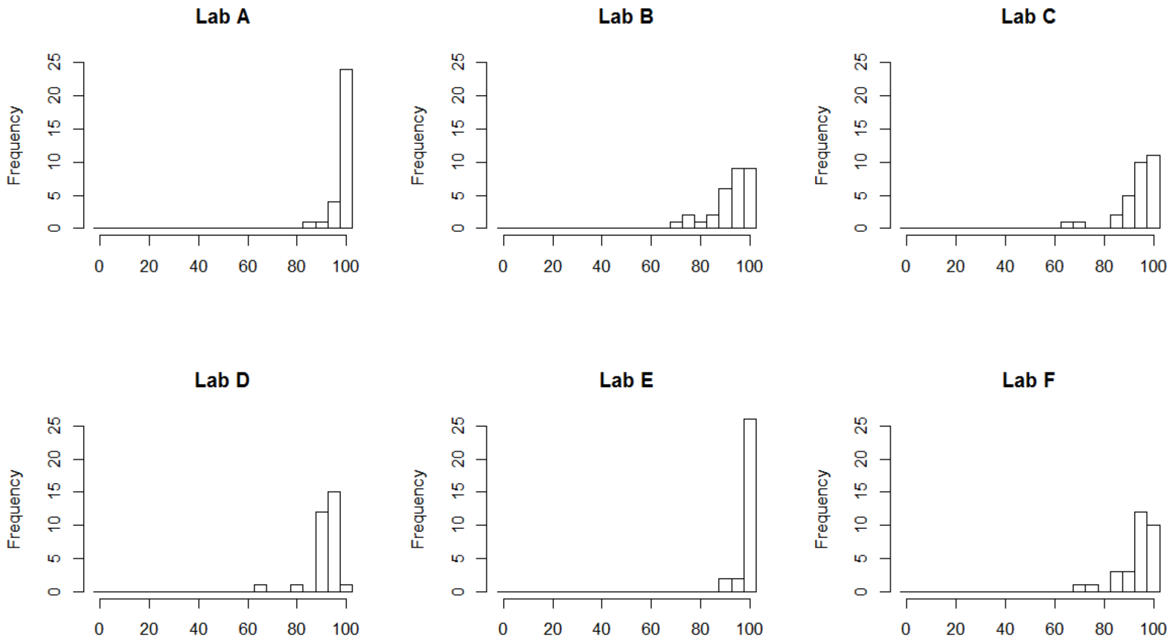


Figure 1: Histograms of wood failure percentage measurements in each laboratory for the best adhesive under the most severe treatment

results, while those on the last two were held in reserve in case a problem developed in testing one or two primary specimens.

The data for each combination of adhesive and treatment therefore consist of 30 WF% measurements from each laboratory: $Y_{ik}, i = 1, \dots, 6; k = 1, \dots, 30$. The data for the best adhesive under the most strenuous treatment (9 cycles of BDF) are shown in Figure 1 **Error! Reference source not found.**. From these data it is clear that models based on normal distributions within each laboratory are inappropriate. Furthermore, it appears that, while the adhesive performs well in all laboratories, there is some variation in the shapes, spreads, and central tendencies of the distributions across laboratories. In particular, the spread of the distribution in each laboratory tends to be narrower when the center lies closer to 100% than when it is more toward 50%. These are all expected properties of our binomial-based GLMM. A detailed description of the methodology used in the study is available in the original report [14].

Parameter Estimation

We applied model (1) to the transformed responses $W_{ik} = Y_{ik}/5$ separately for each combination of adhesive and treatment. Models were fit using the `glmer` function from the `lme4` package in R [15]. This function uses a Laplace approximation to the integrated log-likelihood from the binomial model when there is more than one random effect [12]. The parameter estimates are shown in Table 1. For reference, the values of $\hat{\pi}$, which are the inverted logits of $\hat{\alpha}$ for each adhesive under each treatment—$\hat{\pi} = e^{\hat{\alpha}}/(1 + e^{\hat{\alpha}})$—are also given in the table. To interpret a $\hat{\pi}$, recall that α is the unknown average logit of wood failure among all possible laboratories and blocks. Because the distribution of random effects is symmetric on the logit scale, α also represents the median log-odds of wood failure among all possible laboratories and blocks. Thus, each $\hat{\pi}$ represents the estimated median wood failure proportion for its adhesive and treatment across all possible laboratories and blocks.

Table 1: Parameter estimates from model (1) for each combination of treatment and adhesive.

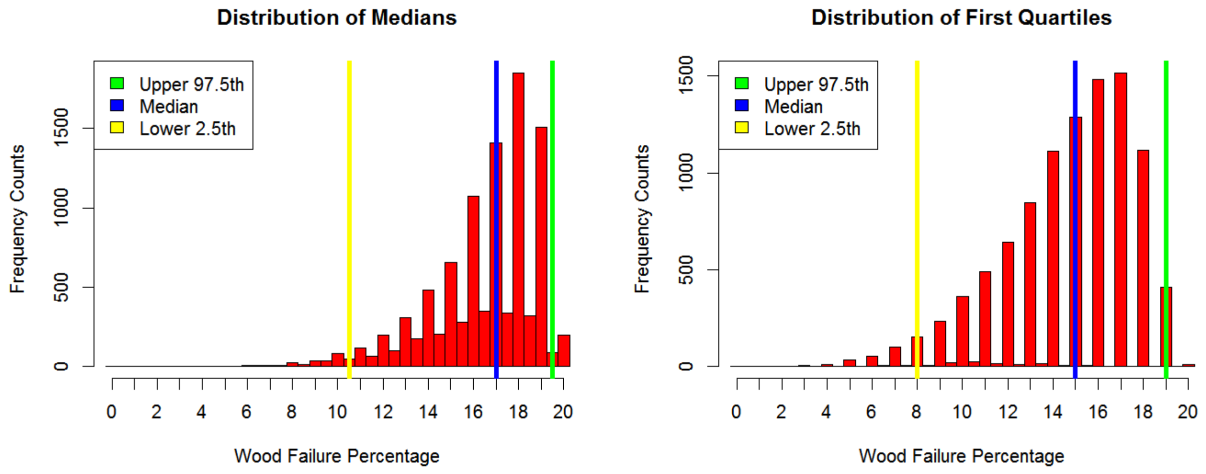
Adhesive	Treatment	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\sigma}_B^2$	$\hat{\sigma}_L^2$
W	BDF	3.35	0.97	0.51	0.65
W	Dry	3.29	0.96	1.17	0.06
W	VP	3.21	0.96	0.20	0.94
X	BDF	-2.34	0.09	3.17	0.57
X	Dry	2.09	0.89	1.17	0.08
X	VP	2.65	0.93	0.80	0.10
Y	BDF	-3.33	0.03	2.29	1.29
Y	Dry	4.49	0.99	1.73	0.24
Y	VP	4.85	0.99	2.38	0
Z	BDF	-2.24	0.10	3.44	0.07
Z	Dry	2.77	0.94	2.02	0.66
Z	VP	-0.82	0.40	0.68	0.34

The parameter estimates for α and π confirm expectations that adhesive W performs well (has high proportions of wood failure) after any of the three treatments. It is the only adhesive that performs well under the BDF treatment, while all adhesives do well under Dry. Results for VP are mixed, with the known worst adhesive, Z, performing poorly after this treatment.

Variance components for block effects and for laboratory effects take a variety of different values for different cases. Indeed, likelihood ratio tests for equality of the block and/or laboratory variance components across the 12 models reject the null hypothesis of equality strongly. Similarly, tests show that the variance components within a given treatment or within a given adhesive are not all equal. This is a disappointing result, but not unexpected due to the use of deliberately disparate adhesives. It suggests that it may not be possible to run a single interlaboratory test to set criteria that apply simultaneously to both performing and non-performing adhesives. That is, it may be that the higher-quality adhesives that manufacturers might actually consider testing for acceptance could have more similar variance components than what are shown here. Examining this issue further is beyond the scope of the present work.

To demonstrate the use of the model to derive acceptance limits for a future SWA, we consider further the results of the good adhesive, W, and the most difficult treatment, BDF. Suppose that we wish to set an acceptance criterion corresponding to a true median WF% of 85% under this treatment. We address the question, "What results might we expect from different laboratory tests of a particular SWA whose median lies at this boundary?" This allows us to assess the role that chance plays in

Figure 2: Estimated sampling distributions of median and first quartile wood failure percentages for adhesive W and treatment BDF based on 10,000 randomly selected laboratories with 30 blocks per laboratory. Also included on each plot are the median value of each distribution and the limits within which 95% of test results would be expected to fall.



determining whether such an adhesive is deemed acceptable. Note that this is a question about reproducibility of the test.

Using the estimated variance components for this case, we simulated 10,000 data sets, each consisting of 30 randomly selected blocks tested at a randomly selected laboratory, using the seven steps outlined in the previous section. The parameter values for the simulation were $\hat{\alpha} = \log(85/(100 - 85)) = 1.73$, $\hat{\sigma}_L^2 = 0.65$, and $\hat{\sigma}_B^2 = 0.51$. From each data set, we computed the sample median and first quartile of the 30 simulated WF% measurements. The estimated sampling distributions of these two quartiles are shown in Figure 2. Vertical lines indicate the 2.5th, 50th, and 97.5th percentiles of the respective distributions. Note that the alternating pattern of high-low bars is because of the discreteness of the WF% measurements. Quartiles calculations are more likely to fall on values that can be observed directly than on those in between..

The histograms show a considerable amount of variability in both statistics. If adhesive W had true median performance level right at the hypothetical boundary of

85%, then with probability approximately 0.95 it would provide a test median ranging between 52.5% and 97.5%. Similarly, the first quartile, which was not directly specified in the simulations, would range between 40% and 95% with a median value of 75%. Setting the laboratory variance component to zero for repeatability limits results in a much narrower spread of resulting medians and first quartiles, as one would expect (plots not shown). The median WF% ranges from 80% to 90% with probability 0.95 with a median value of 85%, while the first quartile's corresponding limits are 62.5% and 85% with a median value of 75%.

Assessing the Model Fit

It is important to examine whether the binomial mixed model provides a reasonable representation for the data on which it is fit. We address this by evaluating whether the models produce simulated data whose distribution is consistent with the actual data from the examples. We provide this assessment separately for each of the 12 combinations of adhesive and treatment using a parametric bootstrap goodness-of-fit test.

To begin, we are testing the null hypothesis that the observed data were generated by the statistical model represented in Equation (1). That is, WF% values generated by each model should follow a probability distribution that is "similar" to the relative frequencies observed in the actual data. Thus, for each combination of adhesive and treatment we need to perform three tasks: (a) establish the probability distribution of WF% for the estimated model, (b) compute an appropriate test statistic to compare the empirical distribution of the data to this distribution, and (c) compute a p-value for the test from the sampling distribution of the test statistic under the null hypothesis.

Unfortunately, there is no simple way to obtain the marginal probability distribution of responses from a binomial mixed model exactly. We therefore use massive simulation from the estimated model to provide a very close approximation to the true distribution. We simulated WF% values for 30 blocks from each of 1,000,000 laboratories, and used these 30,000,000 samples to estimate the probabilities of each possible response value. Refer to these estimated response probabilities as $P_0, \dots, P_{20}$, and let $C_0, \dots, C_{20}$ be the corresponding cumulative probabilities, $C_h = \sum_{g=0}^{h} P_g$, for $h = 0, \dots, 20$.

Next we require a test statistic that is appropriate for testing fit for a discrete probability distribution. Several such statistics are discussed in [16]. These statistics are based on comparing the cumulative distribution of the data to the true cumulative distribution, which in this case is estimated by $C_0, \dots, C_{20}$. Following the authors' recommendation, we use the Cramér-von Mises A^2 statistic, which for our data has the form:

$$A^2 = 180 \sum_{h=0}^{19} \frac{P_h (C_h - c_h)^2}{C_h (1 - C_h)},$$

where c_h is cumulative relative frequency of the observed data; i.e., the proportion of observed WF% values data at or below $5h$.

The sampling distribution of A^2 is not known when the true distribution follows a complex form such as our binomial mixed model. We therefore once again use a parametric bootstrap to estimate its distribution under the null hypothesis. The goal of the simulation is to estimate the p-value of the test, which is the probability that a value of A^2 would occur that is at least as large as the one observed with the original data, when data sets of the same size and structure as the original data are generated from the estimated binomial mixed model. This requires the following steps:

1. Compute A^2 on the original data. Call this A_0^2.
2. Simulate data for 6 randomly-drawn laboratories using steps 1-5 of the algorithm given in the "Analysis of the Model" subsection.
3. Use these parametrically resampled data to compute a new test statistic, A^{2*} by following *exactly* the same steps that led to A_0^2 from the original data:
    a. Fit model (1) to the parametrically resampled data to estimate model parameters.
    b. Use massive simulation with this newly estimated model to estimate its response probabilities and cumulative response probabilities, say $P_1^*, \dots, P_{21}^*$ and $C_1^*, \dots, C_{21}^*$, respectively.
    c. Compute A^2 on the parametrically resampled data and its estimated response probabilities. Call the result A^{2*}.
4. Repeat steps 2-3 a large number of times.
5. Compute the p-value for the test as the proportion of parametrically resampled data sets for which $A^{2*} \geq A_0^2$.

Note that, while step 3 of this algorithm is extremely computationally intensive, it is necessary to prevent the potential tautology that would arise from using the same data to estimate a model and assess its fit to the data. Although A_0^2 does, indeed, compare observed data to a model estimated from the same data, step 3 provides an estimate of the sampling distribution that a test statistic would have when it is calculated in exactly this way. Thus, any biases that would be inherent in such a statistic are explicitly incorporated into the estimated sampling distribution.

The results of this procedure as applied to each adhesive and treatment are given in Table 2. It is clear from this table that the model provides a reasonable fit for some of the data sets, but not all of them. Three cases have their null hypotheses rejected at the 0.05 level, and four others have p-values between 0.05 and 0.10. However, in only one case is the null hypothesis of an adequate model fit soundly rejected: adhesive Z under the dry treatment.

Table 2: P-values from goodness-of-fit tests for model (1) from each combination of treatment and adhesive.

Adhesive	Stress	p-value
W	BDF	0.072
W	Dry	0.024
W	VP	0.072
X	BDF	0.016
X	Dry	0.080
X	VP	0.852
Y	BDF	0.276
Y	Dry	0.612
Y	VP	0.976
Z	BDF	0.064
Z	Dry	0.000
Z	VP	0.204

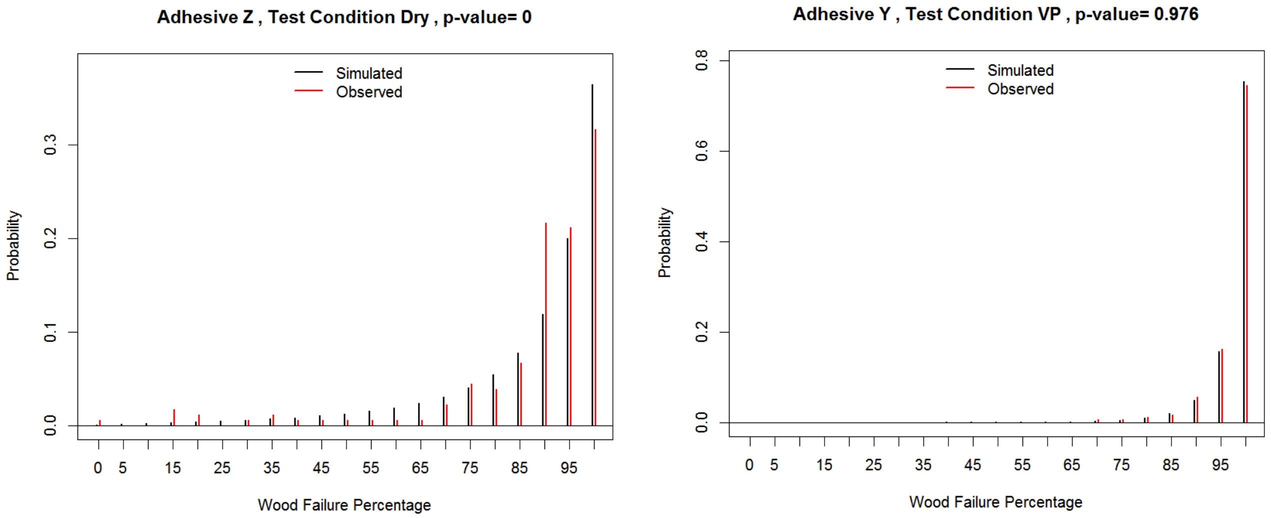


Figure 3: Comparison of observed and model-simulated marginal distributions for data from the cases with the smallest and largest goodness-of-fit test p-values.

In Figure 3 the histograms of observed responses and simulated model probabilities are shown for this worst case and for the case with the largest p-value, adhesive Y under VP treatment. In the latter case, the fit of the model is seen to be nearly perfect. The poor fit in the former case is caused mainly by the substantially higher numbers of blocks with WF% below 25%. Because the estimated model places very small probabilities on such values, any observed blocks there are unusual, and therefore the corresponding terms in the test statistic are very large. However, this adhesive is known to be a poor performer, as seen from its median probabilities for the VP and BDF treatments in Table 1, and would be unlikely to pass a test based on any reasonable standards. It is perhaps not a serious concern that the model does not fit this case.

Looking at the best adhesive, W, an interesting pattern is seen in the histograms (not shown). There is a tendency to have substantially more WF% values of 95 and fewer 100s than the model predicts . An example of this feature is shown in Figure 4.

We wonder whether this could indicate a reluctance on the part of the trained evaluators to assign a value of 100% to blocks where there appears to be a very small amount of flat surface remaining on the block, even though it may not be close to 5% of the total area. It is conceivable that evaluators may be uncomfortable giving “perfect” scores when there is slight evidence of imperfection. While this is speculation whose investigation is beyond the scope of this paper, it raises the possibility that the model could be used to identify possible unintended evaluator biases if it is found to be otherwise satisfactory in broader application.

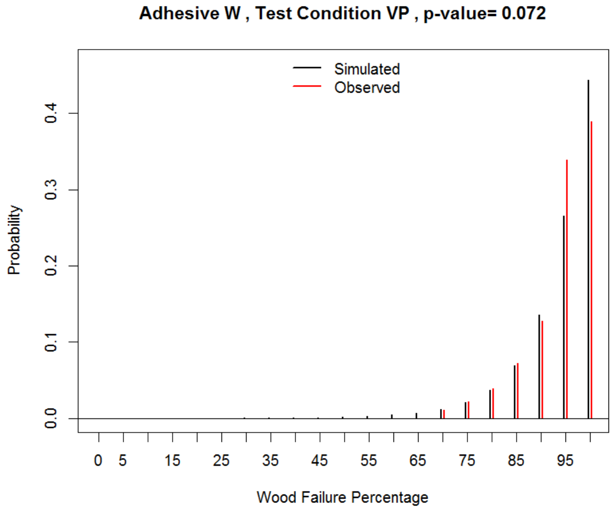


Figure 4: Comparison of observed and model-simulated marginal distributions for a case showing more recordings of 95% and fewer 100% than expected under the model.

Discussion and Conclusions

We have proposed a statistical model for wood failure measurements that are subject to multiple sources of random variability. We treat the 5-point interval percentage responses as having arisen from observing 20 independent binary responses on different parts of the block. The overdispersion that is likely to result from the practical inadequacy of this assumption is overcome empirically by the inclusion of a random effect for blocks. We have demonstrated parametric-bootstrap-based analysis of the model to provide sampling distributions for statistics that regulators might wish to use in setting standards for acceptance of wood adhesives. We provide procedures that address both repeatability and reproducibility of the test results. Similar computational methods are developed to assess the fit of the model. This model provides a reasonably good fit for actual data in many of the cases to which it was applied.

The model can be easily modified to account for other random effects besides block and laboratory effects. For example, one could run a study that considers different technicians within laboratories, different sources of wood, and so forth. Although we have not explored such extensions of the model, they are straightforward conceptually. The logic behind the parametric bootstraps also extends directly to more complex cases. All that is required is the ability to simulate data from the model for the logit, and subsequently use the simulated logits as the basis for generating random binomial responses.

In particular, other standards employ wood failure as an indicator of wood-adhesive bond quality, including: ASTM D3931 [17] for gap-filling adhesives; ASTM D7247 [5] for adhesive bonds in laminated wood products at elevated temperatures; ASTM D7469 [18] for end joints in structural wood products; and ASTM D906 [19] , PS 1-09 [20], CSA O121 [21] and CSA O151 [22] for adhesives in plywood type construction. A number of international wood adhesive standards also use wood failure as an indicator of bond quality. Most of these standards are based on average WF%, rather than median or first quartile, but this poses no problem for our analysis approach. It is straightforward to change the summary statistic in which distributions and intervals are based. What is needed are data sets of WF% for performing and non-performing adhesives, particularly when the test involves a different wood species, treatment (such as ASTM D7247), and/or test specimen (such as ASTM D7469 or D906).

More generally, this model can be used to represent any discretely-measured percentage responses that are subject to random effects from any identifiable sources. The crux of the model is the assumption that the discrete percentages can be transformed into consecutive integer values, which can then be viewed as having arisen from binomial pseudo-trials. Including a random effect for the subjects within which the pseudo-trials are measured should compensate for the overdispersion that results from this assumption.

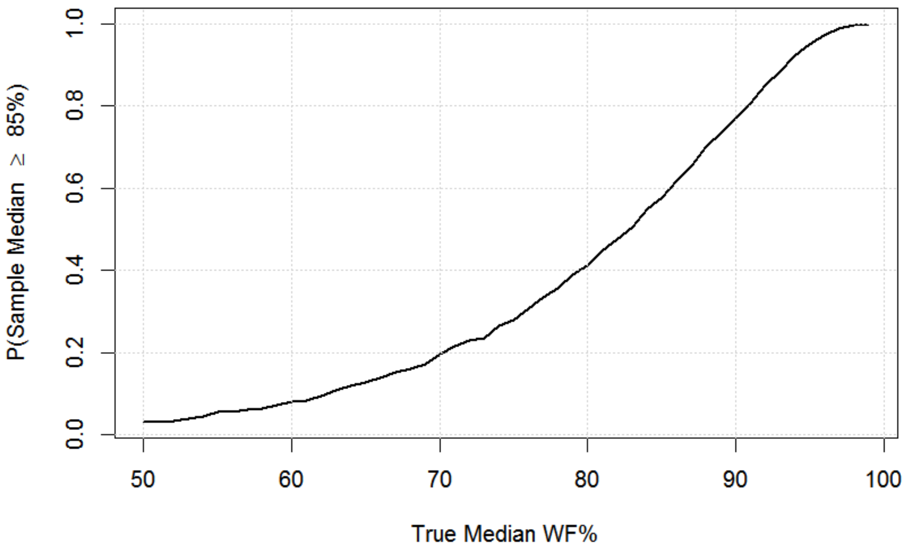


Figure 5: Estimated probability that a laboratory produces data with a median WF% of at least 85%, given different true median values

Finally, from a practical perspective the model allows us to demonstrate clearly the properties that we may expect among wood adhesive tests run in different laboratories. The example shows that the variability inherent in wood failure tests of SWAs, especially the variability between test laboratories, has a considerable effect on the potential outcome of an adhesive test, even though all facilities use the same testing protocol. The plots shown in Figure 1 suggest that setting acceptance limits for adhesives may therefore be a challenge, given the amount of variability that is induced upon the required quartiles. We reran the simulations using the same variance components as in the original example, but using adhesives whose true median performance was allowed to vary between 50% and 99%. The model-estimated probability that such an adhesive would pass a hypothetical standard on the median set at 85% is given in Figure 4. This figure suggests some concerning features about the testing process. For example, about one in five adhesives with true median WF% of

70% would pass a test based on an 85% limit, while one in five with a true median of 91% would fail in a test, merely because of the inter- and intra-laboratory variability inherent in the testing process. Of course, these results should be taken as tentative, and could change in a much larger study aimed at estimating the required variance components with more certainty. The indication is clear that differentiating between adhesives that are truly above the standard and those below it is a difficult task.

Acknowledgments

We gratefully acknowledge support from the National Science and Engineering Research Council of Canada, Grants CRDPJ 381211-09 and Discovery 342205.

References

[1] C. R. Frihart and C. G. Hunt, "Adhesives with Wood Materials: Bond Formation and Performance," in *Wood Handbook, Wood as an Engineering Material*, Madison, WI, U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, 2010, pp. 10-1 to 10-24.

[2] CSA O112.9-10 (R2014), Evaluation of adhesives for structural wood products (exterior exposure), Toronto, ON: CSA Group, 2010.

[3] CSA O112.10-08 (R2013), Evaluation of Adhesives for Structural Wood Products (Limited Moisture Exposure), Toronto, ON: CSA Group, 2008.

[4] ASTM D2559-10a, Standard Specification for Adhesives for Bonded Structural Wood Products for Use Under Exterior Exposure Conditions, West Conshohocken, PA: ASTM International, 2012.

[5] ASTM D7247-07a, Standard test method for evaluating the shear strength of adhesive bonds in laminated wood products at elevated temperatures, West Conshohocken, PA: ASTM International, 2011.

[6] ASTM D905-08, Standard Test Method for Strength Properties of Adhesive Bonds in Shear by Compression Loading, West Conshohocken, PA: ASTM International, 2013.

[7] ASTM D5266-99 (Reapproved 2005), Standard practice for estimating the percentage of wood failure in adhesive bonded joints, West Conshohocken, PA: ASTM International, 2005.

[8] ASTM E456-13ae1, Standard Terminology Relating to Quality and Statistics, West Conshohocken, PA: ASTM International, 2013.

[9] ASTM E691-09, Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method, West Conshohocken, PA: ASTM International, 2009.

[10] G. A. Milliken and D. E. Johnson, Analysis of Messy Data, Volume 1: Designed Experiments, Second edition, Boca Raton, FL: Chapman and Hall/CRC Press, 2004.

[11] G. Molenberghs and G. Verbeke, Models for Discrete Longitudinal Data, New York: Springer, 2005.

[12] C. R. Bilder and T. M. Loughin, Analysis of Categorical Data with R, Boca Raton, FL: CRC Press, 2014.

[13] A. C. Davison and D. V. Hinkley, Bootstrap Methods and their Applications, New York: Cambridge University Press, 1997.

[14] R. C. Casilla and C. Lum, "Round Robin Testing of Structural Wood Adhesives," FPInnovations, Forintek, Vancouver, BC, Canada, 2009.

[15] D. Bates, M. Maechler, B. Bolker and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4. R package Version 1.1-7," http://CRAN.R-project.org/package=lme4, 2014.

[16] V. Choulakian, R. A. Lockhart and M. A. Stephens, "Cramér-von Mises Statistics for Discrete Distributions," *The Canadian Journal of Statustics,* vol. 22, no. 1, pp. 125-137, 1994.

[17] ASTM D3931-08, Standard test method for determining strength of gap-filling adhesive bonds in shear by compression loading, West Conshohocken, PA: ASTM International, 2011.

[18] ASTM D7469-12, Standard test methods for end-joints in structural wood products, West Conshohocken, PA: ASTM International, 2013.

[19] ASTM D906-98 (Reapproved 2011), Standard test method for strength properties of adhesives in plywood type construction in shear by tension loading, West Conshohocken, PA: ASTM International, 2011.

[20] PS 1-09, US Dept. of Commerce Voluntary Product Standard: Structural Plywood, Gaithersburg: NIST, 2010.

[21] CSA O121-08 (R2013), Douglas Fir Plywood, Mississauga , ON: CSA Group, 2013.

[22] CSA O151-09 (R2014), Canadian softwood plywood, Mississauga, ON: CSA Group, 2014.